Thèse

Présentée pour l'obtention du grade de
DOCTEUR DE L'UNIVERSITÉ DE STRASBOURG
Discipline : Sciences du Vivant
Mention : Biophysique et Biologie Structurale

par

Morgan TORCHY

ETUDE STRUCTURE-FONCTION DU COMPLEXE DE REMODELAGE DE LA CHROMATINE NuRD

STRUCTURE-FUNCTION STUDY OF THE CHROMATIN REMODELLING COMPLEX NuRD

Soutenue le 16 Décembre 2014, devant la commission d'examen composée de :

Pr. Peter B. BECKER	Rapporteur externe
Dr. Sébastien FRIBOURG	Rapporteur externe
Pr. Jean CAVARELLI	Examinateur
Dr. Bruno KLAHOLZ	Directeur de thèse

Thèse préparée au sein du Département de Biologie Structurale Intégrative
de l'Institut de Génétique et de Biologie Moléculaire et Cellulaire (IGBMC)
Université de Strasbourg/ CNRS UMR7104 / INSERM U964
Illkirch, FRANCE

ACKNOWLEDGMENTS / REMERCIEMENTS

Pour commencer, je tiens à remercier Bruno pour m'avoir accueilli, il y a maintenant un peu plus de cinq ans. Merci de m'avoir permis de poursuivre mon doctorat au sein de ton équipe, d'avoir financé toutes ces années et de m'avoir fait confiance sur ce projet ambitieux. Je suis arrivé sans vraiment de connaissances en biologie, avec une appréhension certaine, mais j'en ressors en vrai biochimiste. Merci !

My gratitude also goes to the members of the jury, Pr. Peter Becker from Munich, Dr. Sébastien Fribourg from Bordeaux and Pr. Jean Cavarelli from Strasbourg. Thank you for reading these pages, and for accepting to judge my work.

My special words of appreciation go to Kareem. Thank you for being so demanding and for your advice. We've started a nice and complex project together and I can only hope it will be rewarding! I'm sure the hours we spent in the cold room at midnight weren't for nothing…

Évidemment, je veux remercier Isabelle. Tu as été mon premier contact quand je suis arrivé il y a cinq ans au labo, et j'aurais pu difficilement tomber sur une meilleure collègue… Ces quelques années sont finalement passées assez vite, et ta présence, au labo et en dehors, y a largement contribué. Je suis heureux de pouvoir te compter aujourd'hui parmi mes vrais amis.

Of course, I wanna thank all the members, past and present, of our team, for your support, advice and for constructive and fruitful discussions. Merci à Isabelle pour ton énergie et tes idées sans cesse bourgeonnantes; merci à Jean-François pour m'avoir initié aux rudiments de la reconstruction 3-D; *thanks to Sasha, for your patience and help in cryo-EM;* Sinthuja, pour avoir reconstitué des kilos de nucléosomes; *thanks to Heena, I'm wishing you all the best for the future (by the way, if a grammar or spelling mistake has crept into this manuscript, blame her!);* Karima, pour ta folie douce et tous les ragots… *And thanks to all of you!* J'aimerais également profiter de ces quelques lignes pour souhaiter bonne chance à Rachel qui a rejoint l'équipe cette année et qui prend la suite de ce projet.

Notre travail serait bien moins efficace sans le support et l'aide de la plateforme de génomique et biologie structurale. Un grand merci donc à Catherine et Florence, pour vos conseils précieux en biochimie et biophysique; Pierre pour ton support constant en cristallisation *and Alastair for your advice and support in crystallography (he's also to blame, together with Heena!).* Je me réjouis d'avance de rejoindre votre équipe !

Je souhaiterais remercier Ali Hamiche avec qui nous avons démarré une jolie collaboration. L'étude de ce complexe est un beau projet, avec tant de choses à découvrir, et j'espère qu'il se poursuivra aussi longtemps que possible. Merci également à Arnaud Depaux, qui a su se montrer

patient à mon arrivée à l'IGBMC et m'a appris l'art du clonage quand je savais à peine ce qu'était un plasmide…

Merci au service baculovirus, et plus particulièrement à Nathalie, pour les dizaines et dizaines de litres de culture que tu as pu me fournir !

Merci à Adeline, Virginie, Mathilde et Franck à l'IGBMC, ainsi que Philippe, Lauriane et Johana à l'IBMC, pour les analyses en spectro de masse.

Un grand merci à l'ensemble des membres du département de Biologie Structurale Intégrative, avec qui j'ai eu plaisir à travailler.

Durant ces quatre années de doctorat, j'ai eu la chance de participer à une superbe opération de promotion de la biologie auprès de lycéens. OpenLAB a vraiment été une expérience enrichissante et j'aimerais remercier Laurence Drouard, Michel Labouesse, Stéphane Vincent, Catherine Florentz et Serge Potier pour avoir permis de pérénniser cette opération et m'avoir donné l'opportunité d'y participer activement. Et évidemment, mes collègues et amis OpenLABistes durant ces deux jolies années: Mélanie, Morgane, Claire, Anaïs, Léa, Patrick et Benjamin.

Mais parce qu'il faut aussi savoir être bien entouré en dehors du labo, je veux remercier ici quelques personnes dont l'amitié, même si elle est parfois lointaine, fait ou a fait que les jours étaient plus beaux. Outre ceux que j'ai déjà cité plus haut et qui se reconnaitront, je pense particulièrement à Sophie, mon amie depuis 25 ans maintenant; Étienne, qui malgré mon acharnement, confond toujours ribosome et rhizome; Floriane, ma plus belle rencontre sur les bancs de la fac; Raphaël, mon ami biologiste qui ne comprend rien à ma biologie (et réciproquement); Pierre, mon ami nanto-lyonnais bientôt plus strasbourgeois que moi; Gautier, le plus beau soutien de mes premières années; et évidemment beaucoup d'autres qui n'ont pas besoin d'être cités pour savoir que je pense à eux.

J'aimerais enfin remercier ma famille, et plus particulièrement mes parents, qui n'ont jamais vraiment bien compris ce que je faisais mais qui m'ont tout de même soutenu depuis le début de mes études; qui ont dû s'arracher les cheveux en me voyant passer par 4 facs différentes en 4 ans; et qui s'inquiètent régulièrement de savoir si mes bêbêtes vont bien.

"Oui Maman ! mes bêbêtes vont bien !"

Les listes de remerciements sont toujours interminables et on finit forcément par oublier quelqu'un… Pour autant, vous pouvez être assuré que je vous remercie quand même ! Vous pouvez ajouter votre nom ci-dessous si jamais… ;)

TABLE OF CONTENTS

TABLE OF FIGURES AND TABLES

COMMON ABBREVIATIONS

AUC	Analytical ultracentrifugation
BME	β-mercaptoethanol
cDNA	complementary DNA
CHAPS	3-[(3-Cholamidopropyl)dimethylammonio]-1-propanesulfonate
CMC	Critical micelle concentration
Cryo-EM	Cryo-electron microscopy
DLS	Dynamic light scattering
DMSO	Dimethyl sulfoxide
DNA	Deoxyribonucleic acid
DTT	Dithiothreitol
EDTA	Ethylenediaminetetraacetic acid
EMSA	Electromobility shift assay
ESI-TOF	Electrospray ionization-Time of flight
HEPES	4-(2-hydroxyethyl)-1-piperazineethanesulfonic acid
ITC	Isothermal titration calorimetry
MALDI-TOF	Matrix-assisted laser desorption/ionization-Time of flight
MES	2-(N-morpholino)ethanesulfonic acid
MPD	2-methyl-2,4-pentanediol
mRNA	messenger RNA
NTA	Nitrilotriacetic acid
NP-40	Nonyl phenoxypolyethoxylethanol-40
OD	Optical density
PAGE	Polyacrylamide gel electrophoresis
PCR	Polymerase chain reaction
PCS	Photon correlation spectroscopy
PEG	Polyethylene glycol
PMSF	Phenylmethylsulfonyl fluoride
QUELS	Quasi elastic light scattering
RNA	Ribonucleic acid
rRNA	ribosomal RNA
SDS	Sodium dodecyl sulfate
SEC-MALLS	Size exclusion chromatography-Multi-angle laser light scattering
snRNA	small nuclear RNA
TBE	Tris-Borate-EDTA
TCEP	Tris(2-carboxyethyl)phosphine
TRIS	Tris(hydroxymethyl)aminomethane
tRNA	transfer RNA

Amino acids and nucleotides

Alanine, Ala, A	Isoleucine, Ile, I	Tyrosine, Tyr, Y
Arginine, Arg, R	Leucine, Leu, L	Valine, Val, V
Asparagine, Asn, N	Lysine, Lys, K	
Aspartic acid, Asp, D	Methionine, Met, M	Adenine, A
Cysteine, Cys, C	Phenylalanine, Phe, F	Cytosine, C
Glutamic acid, Glu, E	Proline, Pro, P	Guanine, G
Glutamine, Gln, Q	Serine, Ser, S	Thymine, T
Glycine, Gly, G	Threonine, Thr, T	Uracyl, U
Histidine, His, H	Tryptophan, Trp, W	

Units of measure

Å	Ånsgtrom (0.1 nm)	mS	millisiemens
bp	base-pair	nL	nanolitre
°C	degree centigrade or Celsius	nm	nanometre
cm	centimetre	pfu	plaque forming unit
Da	dalton	rpm	revolution per minute
g	gram	S	svedberg (10^{-13} seconds)
kbp	kilo base-pair	TB	terabyte
kDa	kilo Dalton	µg	microgram
kV	kilovolt	µL	microlitre
L	litre	µm	micrometre
M	molar (moles/litre)	µM	micromolar (micromoles/litre)
mA	milliampere	V	volt
mg	milligram	v/v	volume/volume
mL	millilitre	W	watt
mm	millimetre	w/v	mass/volume
mM	millimolar (millimoles/litre)	Xg	acceleration of gravity

SUMMARY OF THE THESIS

In an organism, every cell contains the same genetic material. Nevertheless, evolution has made possible the selective expression of specific genes, and the repression of some others, and thus allowed cell specialization. With the establishment of differential expression patterns, cells can differentiate and organisms can develop. For a given cell, various normal and pathological processes can occur, such as reactions to stress stimulation (nutrient deficiency, hypoxia, lack of growth factors, etc.), pathologies related to deregulations of gene expression (cancers, etc.) or simply the progress of the cell cycle. This modulated gene expression is made possible by chromatin remodelling, a process that is thought to be related with the accessibility of the DNA of target genes to transcription factors or RNA polymerase in particular.

In 1942, Conrad Waddington coined the term “epigenetic”, the branch of biology which studies “the causal interactions between genes and their products, which brings the phenotype into being”. Indeed, genes and more generally, chromatin, are targeted by covalent modifications, which can be recognised by protein effectors, allowing the recruitment of enzymes and other partners involved in chromatin remodelling. In 1998, several groups described a complex exhibiting an ATP-dependent remodelling activity, similar to that of ySWI/SNF from Saccharomyces cerevisiae, and coupled to a histone deacetylation activity. This complex, called NURD, NRD, Mi-2 complex, and finally, NuRD, standing for “Nucleosome Remodelling and histone Deacetylation”, is one of only two known complexes coupling two independent chromatin-remodelling activities. One possible reason for that could be that the ATP-dependent remodelling activity is necessary for the Histone Deacetylase (HDAC) subunits to access their target. This idea is supported by the observation that in absence of ATP, deacetylation is only possible on histone octamers, and not on nucleosomes. The binding site of HDACs could be somehow protected by the DNA, and thus inaccessible. Experiments carried out to determine whether ATP could stimulate deacetylase activity did not show any significant effect on free histone octamers. By contrast, when nucleosomes were tested, ATP was shown to stimulate deacetylase activity by two-fold: without ATP, 30-35% of acetylated H4 histones were deacetylated, while in the presence of ATP, 60-70% were[1].

The NuRD complex is highly conserved among superior eukaryotes, and is expressed in a large variety of tissues. It forms a large macromolecular assembly that consists of different protein subunits; however, different homologs and isoforms have been described for each of those subunits, leading to a horde of coexisting NuRD complexes, depending on the cellular, tissue, physiological or pathological context. Moreover, the stoichiometry of the different subunits remains an open question. Recently, the development of a new label-free quantitative mass spectrometry method, applied to the analysis of NuRD, suggested that it is composed of one CHD3 or CHD4 protein (Chromodomain, Helicase, DNA binding domain), one HDAC1 or HDAC2, three MTA1/2/3 (Metastasis Associated), one MBD3 (Methylated CpG-Binding), six RbAp46/48 (Retinoblastoma Associated protein), two p66α or p66β and two DOC-1 (Deleted in Oral Cancer)[2]. Those data are nevertheless in contradiction with the structural analysis of the HDAC1/MTA1 complex showing a dimerisation of MTA1, suggesting the presence of two MTA1/2/3 and two HDAC1 or HDAC2 in NuRD[3]. The

specificities of each isoform, together with the sharing of competences such as the opposite activities of deacetylation and remodelling, ensure that NuRD is a major actor in various biological processes, like embryonic development, cellular differentiation, haemato- and lymphopoeisis, tumour growth inhibition, or the general repression of transcription. Furthermore, it directly interacts with various partners, like the lysine specific demethylase 1 (LSD1/KDM1A)[4], Ikaros, Aiolos, Helios[5-7], B-cell lymphoma 6 (BCL6)[8,9], the oestrogen receptor α (ERα/NR3A1)[10-12] or Oct4/Sox2/Klf4/c-Myc (OSKM)[13,14]. This highlights the very broad and general role of NuRD, especially given that it is the most abundant form of deacetylases in mammals.

The work I carried out during my PhD is part of a global and long-term project to study the NuRD complex. Indeed, 16 years after its discovery, its role still remains poorly understood. Although this complex plays undoubfully a nearly ubiquitous role in our cells, we still lack biochemical, genetic and structural data to understand the precise role of a given NuRD complex, *in vitro*, but also in its cellular environment. Numerous studies are focused on isolated subunit, and results obtained are being extrapolated to the whole complex.

In the team of Dr. Bruno Klaholz, at the IGBMC, we study large complexes involved in gene expression regulation, by an integrative structural biology approach. Several techniques including X-ray crystallography, cryo-electron microscopy (cryo-EM) and a large panel of biophysical tools allow us to describe with great accuracy interactions between partners within a complex, like nuclear receptors on their DNA target, polyribosomes linked to a messenger RNA, or translation initiation complexes, all being involved in gene expression regulation. These same methods were used to study the structure-function relationship of the chromatin complex NuRD, of its subunits and their complexes with nucleosomes, in connection with functional studies carried out in the team of our collaborator, Dr. Ali Hamiche at the IGBMC.

This PhD work thus allowed setting up a new and ambitious project, with two main ideas in mind: the structural analysis of isolated subunits of the complex and stable subcomplexes within NuRD or with chromatin components like nucleosomes; and second, the study of the whole endogenous complex purified from human cells.

In order to reconstitute the NuRD complex *in vitro*, all its subunits were cloned in baculovirus expression vectors, for insect cells production. Eleven constructs with different affinity tags were thus designed and tested. Protein production from insect cells required a fine optimisation of the cultivation protocol: cell line, virus titration, cultivation time, etc. After cultivation optimisation, my efforts were mainly focused on three proteins of the NuRD complex: RbAp46, RbAp48 and MBD3. The latter remains very little studied and poorly understood. MBD3 belongs to the MBD family, binding methylated CpG islands. In mammals however, this protein has lost its ability to bind to methylated DNA and binds unmodified DNA instead. This is due to a point mutation (Y34F) that appeared with the emergence of the mammal class. The RbAp46 and RbAp48 proteins, for their part, are both histone chaperones, found within chromatin-associated complexes, like HAT-1, CAF1, Sin3A, Polycomb, EZH2/EED, NURF and NuRD. They have often been described as a stable

structural platform for these complexes, although their chaperoning role has been only poorly investigated. Their study in complex with the nucleosome will allow shedding light on new information about their primary function.

While this project was ongoing, crystal structures of RbAp46 and RbAp48, in complex with a short peptide of histone H4, were determined by the group of Ernest Laue[15]. These structures suggest that the RbAp chaperones can only bind free H4 histones, but not nucleosomes. Using reconstituted recombinant human nucleosomes, we decided to undertake first binding studies to verify the previous assumptions. Electromobility shift assays were carried out with both RbAp46 and RbAp48 and reconstituted nucleosomes, and surprisingly showed a positive result. The complex was reconstituted biochemically and crystallisation assays were carried out. First crystals of RbAp46-Nucleosome were obtained in a dozen of different conditions. Their diffraction was tested in the Swiss Light Source (SLS, Villigen, Switzerland). However, their high mosaicity and weak diffraction did not allow determining the structure. Crystallisation conditions are currently being optimised and will hopefully lead to the first structure of a nucleosome in complex with a histone chaperone.

In parallel, MBD3 was produced and purified using baculovirus expression system, with a good yield, but as soluble aggregates only. Optimisation of this purification process was, however, not pursued further because mass spectrometry revealed the presence of a shorter isoform, missing the DNA-binding domain. The cDNA of the long isoform of MBD3 being not available in cDNA banks, a synthetic gene was thus designed and synthesised, with optimised codon bias for both baculovirus and bacterial expression systems. The production of this protein was implemented and optimised in bacteria, but the non-negligible loss of insoluble material led to the development of a new purification protocol under denaturing conditions. With a yield of several tens of milligrams of pure protein per litre of culture, a large panel of *in vitro* refolding conditions could be tested. Unfortunately, regarding the instability of this protein in absence of denaturing agent, no quality sample could be obtained.

Production in a 20 or 100-litre bioreactor was thus implemented in order to get more material to achieve native purification. A pure protein could be obtained as soluble aggregates, and Thermofluor® experiments were carried out. In this way, new buffer conditions were defined, in which MBD3 seemed more stable. This new purification protocol indeed led to a final yield of 200 to 300 micrograms of pure protein per litre of culture. Size-exclusion chromatography showed the presence of a major species of 70 to 80 kilo Daltons, suggesting an MBD3 dimer, as mentioned in the literature[16]. However, analytical ultracentrifugation analysis revealed that the purified protein was in reality monomeric and partially unfolded. Binding assays with nucleosomes were nevertheless carried out and optimisation of salt concentration and MBD3-nucleosome ratio led to a positive result observed by electromobility shift assay. This result suggests that only the C-terminal part of MBD3, which is not involved in DNA-binding, is unfolded. It would thus explain, at least partly, the great instability of this protein towards minor temperature changes. Rigorous working conditions had to be implemented, including handling of the sample below 4°C exclusively. Crystallisation trials of this MBD3-nucleosome complex were carried out and crystals were obtained in various

conditions, mostly different from typically-observed conditions for nucleosome crystallisation. Their diffraction was tested in the SLS, and the good results obtained encourage pursuing crystal optimisation. In particular, the topology and space group of these crystals turned out to be different from crystals of the core nucleosome particle.

Considering that this project was built up from zero in our team, there is considerable scope of perspectives. First, the MBD3-nucleosome and RbAp-nucleosome crystals are currently being optimised and shall lead hopefully to high-resolution structures of these complexes. As part of the MBD3 project, several clonings have been carried out to produce the DNA-binding domain only, with different boundaries. First purification assays were very conclusive, and DNA oligos were used to carry out binding assays. Using the binding conditions already published for MBD2 and MBD4, I observed a positive binding of MBD3 on an 11-bp DNA oligo, using electromobility shift assay. Furthermore, binding assays of this DNA-binding domain have also been carried out on nucleosomes and in spite of the various optimisations made, no binding could be observed, suggesting a role of the C-terminal part of MBD3 in specifically binding nucleosomes.

With our expertise in cryo-EM, the full-length MBD3 protein in complex with nucleosomes was flash-frozen on EM-grids, and images collected. A first data collection led to a low-resolution 3-D reconstruction. Despite this limited resolution (25 Å), this first electron density shows a circular and flat shape in which the crystal structure of the nucleosome core particle could be fitted; an additional density was visible, in which the crystal structure of the MBD domain of MBD2 was fitted. This extra-density shows a clear interaction with the DNA on the side of the nucleosome but also spreads on the face of the nucleosome. However, the low resolution of this first reconstruction does not allow to conclude whether MBD3 is interacting with histones or not, although it would be consistent with the idea that the C-terminal part of MBD3 could be involved in nucleosome recognition, this part being highly acidic, which is fully compatible with basic histones binding. Recent functional studies support this hypothesis, as MBD3 has been shown to be involved in nucleosome organisation near promoters and in gene bodies[17]. New data collections are currently being processed and shall lead to a clear answer in the future.

Lastly, the baculovirus vectors that were designed for all the NuRD subunits can now be used for co-infections and produce *in vivo* subcomplexes of NuRD. With this in mind, the protein MTA2 seems to be the perfect starting point. Indeed, recent publications of crystal structures of MTA1 highlight its interactions with HDAC1 and RbAp48. These data suggests thus the existence of a stable subcomplex including HDAC1 or HDAC2, RbAp46 or RbAp48 and MTA2.

Finally, in the frame of our collaboration with Dr. Ali Hamiche on this project, his team designed a HeLa cell line for endogenous NuRD complex production and purification. This will open new roads in the study of the entire NuRD complex and hopefully address fundamental questions about the function of this complex, such as the localisation of the individual subunits inside the complex and their interaction surfaces, the subunit stoichiometry and, most importantly, its overall structure.

STRUCTURE-FUNCTION STUDY OF THE CHROMATIN REMODELLING COMPLEX NuRD

INTRODUCTION AND STATE OF THE ART

1. Postulates, hypotheses and discoveries timeline

Year	Author	Postulate/Hypothesis/Discovery
1801	Lamarck	**Biology: theory of living organisms.**
1859	Darwin	« On the origin of species by means of natural selection, or the preservation of favoured races in the struggle for life».
1866	Mendel	« Experiments in plant hybridization ».
1871	Miescher	Erythrocyte nuclear extracts contain nuclein, a phosphate-rich substance.
1880	Flemming	The chromatin, dense coloured fibres constituting the nucleus.
1881	Zacharias & Flemming	Chromatin is made of nuclein.
1884	Kossel	Erythrocyte nuclear extracts contain proteins, called histones.
1885	Hertwig	The nucleus contains the support of heredity.
1888	Waldeyer	The chromosome, or coloured body.
1889	Altmann	Nuclein is renamed « nucleic acid ».
1896	Wilson	Nucleic acid is the support of heredity.
1902	Sutton & Boveri	Chromosomes bear Mendelian heredity factors.
1905	Bateson	**Genetic: the study of inheritance and the science of variation.**
1911	Morgan	Demonstration of Sutton & Boveri's theory in *Drosophila melanogaster*.
1924	Feulgen & Voit	Chromosomes are made of thymonucleic acids.
1928	Griffith	The bacterial transforming principle.
1929	Levene	The four bases of DNA: adenine, cytosine, guanine and thymine.
1929	Levene	Histones are the support of heredity.
1941	Beadle & Tatum	One gene = one enzyme.
1942	Waddington	**Epigenetic: the branch of biology which studies the causal interactions between genes and their products, which bring the phenotype into being.**
1944	Avery	DNA is the support of genetic information.
1949	Chargaff	Chargaff's rule: in DNA, A/T and G/C ratios are quasi-equals and close to 1.
1953	Watson, Crick & Franklin	The structure of DNA.
1957	Waddington	Cellular differentiation isn't due to modifications of the genome, but to epigenetic phenomena.
1958	Meselson & Stahl	Semiconservative replication.
1959	Jacob & Monod	Structural genes are first transcribed into an mRNA prior to being translated into proteins.
1962	Allfrey et al.	Histone methylation.

1964	Allfrey et al.	Histone acetylation.
1964	Allfrey et al.	There is a link between histone chemical modifications and transcription regulation.
1965	Nirenberg, Khorana & Ochoa	The genetic code.
1965	Arber	Restriction enzymes.
1965	Doskocil et al.	DNA methylation.
1972	Fiers et al.	First sequenced gene (*MS2 phage* capsid).
1974	Olins & Olins	The beads on a string, observed in TEM.
1975	Oudet, Bellard & Chambon	The nucleosome.
1975	Jackson et al.	Histone phosphorylation.
1976	Woodcock et al.	The 30 nm fibre observed in TEM.
1976	Finch & Klug	The 30 nm fibre: the solenoid model.
1977	Sanger et al.	DNA sequencing.
1977	Goldknopf	Histone ubiquitination.
1978	Laskey et al.	Histone chaperones.
1983	Mullis	The polymerase chain reaction technique (PCR).
1983	Feinberg & Vogelstein	Gene methylation is altered in cancer cells.
1984	Woodcock et al.	The 30 nm fibre: the zigzag model.
1984	Richmond et al.	X-ray structure of the nucleosome at 7 Å resolution.
1985	Keshet et al.	DNA methylation plays a key role in gene regulation in animal cells.
1987		The chromatin immunoprecipitation technique (ChIP).
1989	Greger et al.	CpG islands in promoter regions of tumour-suppressor genes are hypermethylated.
1991	Arents et al.	X-ray structure of the histone octamer at 3.1 Å resolution.
1993	Turner et al.	The epigenetic information is stored in chemical modifications of histones N-terminal domains.
1993	Stoger et al.	Imprinted genes carry their epigenetic marks throughout the entire life of an organism.
1994	Cairns et al.; Peterson et al.	The Swi/Snf complex purified from yeast.
1994	Cote et al.; Imbalzano et al.; Kwon et al.	The yeast and mammalian Swi/Snf complex alters nucleosome structure in an ATP-dependent fashion.
1994	Elfring et al.	The ISWI complex in *Drosophila melanogaster*.
1995	Fleischmann et al.	First prokaryotic genome to be sequenced: *Haemophilus influenzae*
1995	Games et al.	First mouse model for cancer epigenetics studies.

1996	Cairns et al.	The RSC remodelling complex in yeast.
1996	Goffeau et al.	First eukaryotic genome to be sequenced: *Saccharomyces cerevisiae*
1996	Campbell & Wilmut	Dolly, first mammal to be cloned.
1997	Luger et al.	X-ray structure of the nucleosome at 2.8 Å resolution.
1998	Kuo & Allis	Histone acetyltransferases and deacetylases.
1998	Tong et al.; Wade et al.; Xue et al.; Zhang et al.	The NuRD complex.
2001	Consortium public international & Celera Genomics	The human genome is fully sequenced.
2001	Jenuwein & Allis	The histone code hypothesis.
2002	Davey et al.	X-ray structure of the nucleosome at 1.9 Å resolution.
2006	Rhodes et al.	The 30 nm fibre: the interdigitated solenoid model.
2006	Yamanaka et al.	Induced pluripotent stem cells (iPS cells).
2010	Neandertal Genome Consortium	The Neanderthal man genome is sequenced.

2. Chromatin, the compacted state of genetic information

In Eukaryotes, genetic information is stored in a long deoxyribonucleic acid molecule, called DNA, that can measure up to 2 m long, but has to fit in a nucleus that is a about 200'000 times smaller. This long molecule must thus be compacted, with the help of proteins, to form chromatin, the condensed state of DNA. The basic unit of this compaction is the nucleosome.

2.1 The nucleosome: some structural and functional generalities

2.1.1 Discovery

The whole story about nucleosome starts in the early 1970's. At that time, Robert Williamson carried out biochemical studies on isolated chromatin from rat hepatic cells. He observed the migration of different DNA fragments on a polyacrylamide gel, whose size was a multiple of 135 kilo Daltons. This observation led him to assume that nuclear DNA was degraded[18]. In 1973, Dean Hewish and Leigh Burgoyne discovered that chromatin is accessible to Ca^{2+}/Mg^{2+}-endonucleases,

resulting in a series of fragments similar to those observed by Williamson[19]. These fragments are composed of 180 to 200 base-pairs.

In 1974, Olins & Olins confirmed this observation using electron microscopy, highlighting the beads on a string, a structure in which they could observe regularly spaced particles on DNA[20] (*figure 1*). The same year, Kornberg carried out crosslinking studies, allowing him to precisely determine a 1 to 1 stoichiometry between the DNA and the proteins of these particles[21,22]. From these observations resulted the suggestion that chromatin was composed of a fundamental unit, that Pierre Oudet, Maria Gross-Bellard and Pierre Chambon named in 1975, the nucleosome[23].

In 1984, the first X-ray structure at 7 Å resolution allowed to deduce the shape of this particle: a flat disc[24] (*figure 2*). It took then until 1997 to get a X-ray structure at 2.8 Å resolution, required to highlight a major part of the interactions between the histones and the DNA of the complex[25]; and finally 2002, for a finer resolution at 1.8 Å[26] (*figure 3*).

Since then, the PDB (Protein Data Bank, *http://www.rcsb.org/pdb*) has been enriched with numerous structures, each of which contributes in unique ways. In mid-2014, over 80 crystallographic structures could be counted. Among these, 24 structures of nucleosomes carrying a point mutation on histone H3 or H4; 7 structures of nucleosome in complex with a protein partner like Sir3 or RCC1; 8 structures of nucleosomes with a histone variant (in particular, H3.2, H3.3, CENP-A or Macro-H2A); 8 structures of nucleosomes with different DNA sequences (poly dAdT, 601, 145 base-pairs, etc.); 2 structures of nucleosomes with post-translational modifications (H3K79me2 and H4K20me3); 5 structures of nucleosomes chelating ions like cobalt, nickel, rubidium or caesium); 10 structures of nucleosomes in complex with an antitumoral agent like cisplatine, oxaliplatine or osmium and ruthenium complexes; but also several structures of nucleosomes from different species, among which the yeast, D*rosophila melanogaster* and human nucleosome.

2.1.2 The protein component of the nucleosome: the histone

Each nucleosome is composed of a set of two times four proteins: histones H2A, H2B, H3 and H4. These 8 proteins are arranged in a central tetramer $(H3\text{-}H4)_2$, flanked on both side by a dimer (H2A-H2B), to form an octamer. With this octamer is associated a 147 base-pairs DNA, called nucleosomal DNA. The global structure of this octamer is well-known since it was solved in 1991 by *Arents et al*[27]. In the following sections of this manuscript, I will be using the generic term “histone” to refer to these four nucleosomal histones only. I will precisely point out in the case of histones H1 or H5, also called linker histones.

Histones are nuclear proteins, extremely conserved, and found in all Eukaryotes (except in dinoflagellates[28]), as well as some Archaea.

2.1.2.1 General structure

Each histone has a central structured domain called histone core, and an amino-terminal tail.

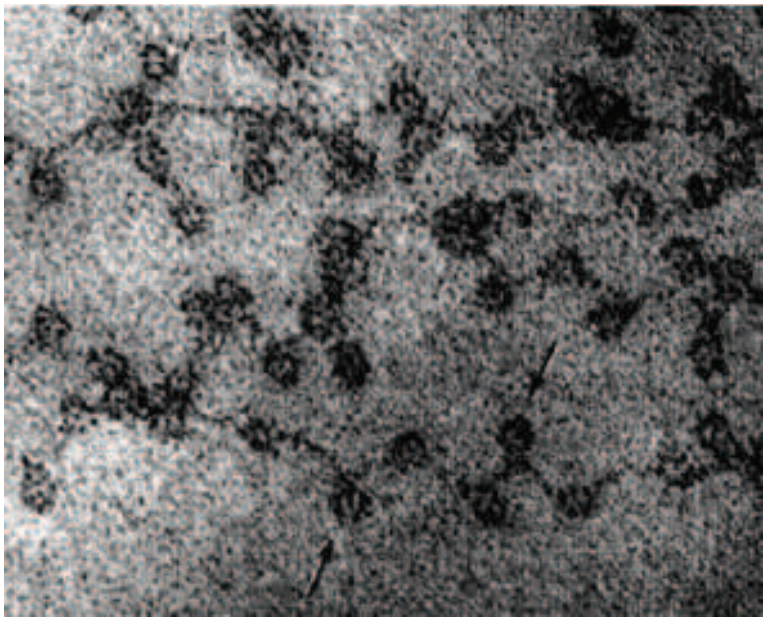


FIGURE 1

The beads on a string

Ultrastructure of a repeated spheroid chromatin unit, observed in electron microscopy.

Olins & Olins, 1974

FIGURE 2

The nucleosome at 7 Å resolution

Richmond et al., 1984

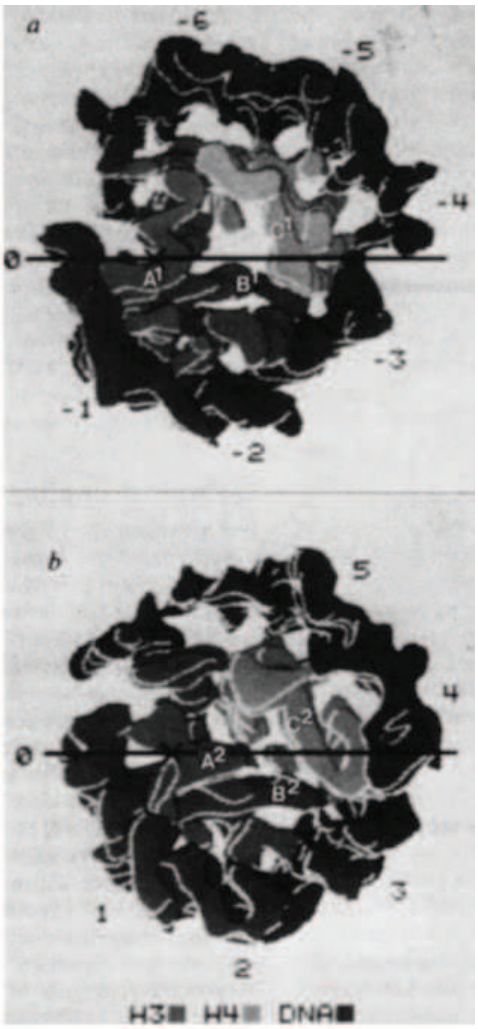


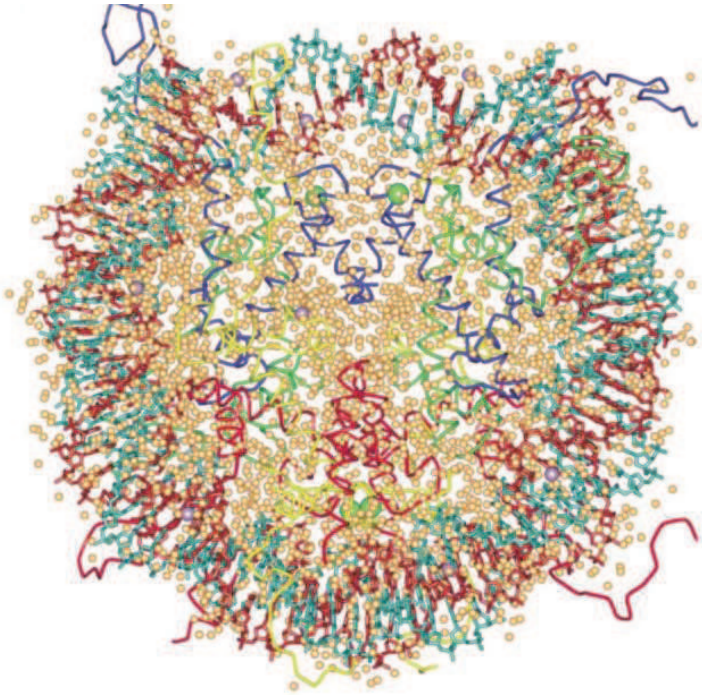
FIGURE 3

The nucleosome at 1.8 Å resolution

The high-resolution structure shows in particular the important solvatation state of this nucleoprotein complex.

Davey et al., 2002

(pdb : 1kx4)



2.1.2.1.1 The histone core

The histone core consists in three α helices, namely α1, α2 and α3, and two loops L1 and L2. These structural elements associate to form a specific motif, conserved in all four histones, and called histone-fold. It is noted α1-L1-α2-L2-α3, the two loops linking the three α helices together. Two histone folds can interact to form the specific pattern called handshake motif (*figure 4*). Histones can thus form heterodimers (H3-H4 and H2A-H2B), strongly linked by hydrogen bonds and hydrophobic interactions. Each histone fold is associated to its homolog in an antiparallel manner, leading to a pseudo 2-fold symmetry, of which the axis travels between the two long α2 helices, at the level of their point of intersection. This antiparallel organization allows the loop L1 of one partner to be juxtaposed with the loop L2 of the other partner, shaping a DNA binding site called L1L2, described later on.

Within a histone, helices α1 and α3 fold back on roughly the same side of the central α2 helix. The two α1 helices of a pair join to form a DNA binding site, called α1α1, on the convex side of the dimer, whereas α3 helices, on the concave side, are not in contact.

The C-terminal end of α2 helices along with α3 helices are the primary determinants of histone assembly to form the octamer, since they participate in the formation of three four-helix bundles, one between H3 and its homolog H3' to form the $(H3\text{-}H4)_2$ tetramer, and one between H2B and H4 to form the octamer. Needless to say, other interactions can be observed during octamer formation, notably between H3–H2A', H4–H2A', H4–H2B', H2A–H2A' and H2A–H2B'. Finally, the α3 helix of H2A might be involved in internucleosomal interaction to form higher order compaction state.

The four-helix bundle between H3 and H3' comprises buried charged groups, responsible for intermolecular hydrogen bonds. These interactions involve aspartic acid D123 of histone H3 and histidine H113 of its homolog H3' (*figure 5a*). The four-helix bundle between H2B and H4 is somewhat similar in its conformation since the Root-Mean Square Deviation (RMSD) of both structures centred on α carbons is only 1.85 Å. Histidine H75 of histone H4 hydrogen-bonds glutamic acid E90 of histone H2B (*figure 5b*). There is nevertheless one difference between these two types of four-helix bundles: while the H3-H3' bundle displays a cysteine in its middle (H3-C110), the H2B-H4 bundle exhibits a tyrosine (H2A-Y72), lying flat on tyrosine Y80 of histone H2B, in the very place where the histidine H113 is found in histone H3'. Consequently, the prevailing hydrophobicity in the H2B-H4 bundle, higher to that in the H3-H3' bundle, explains, at least partly, the instability of the octamer compared with that of the tetramer at low salt concentration.

2.1.2.1.2 The amino-terminal tail

Each histone comprises, in addition to the histone core, an unstructured N-terminal end, more or less long. It represents 28 % of histones, and is extremely basic owing to a significant

FIGURE 4

The histone-fold and the hand-shake motif

In green, H3; in blue, H4;

In red, H2A; in yellow, H2B

(pdb : 1aoi)

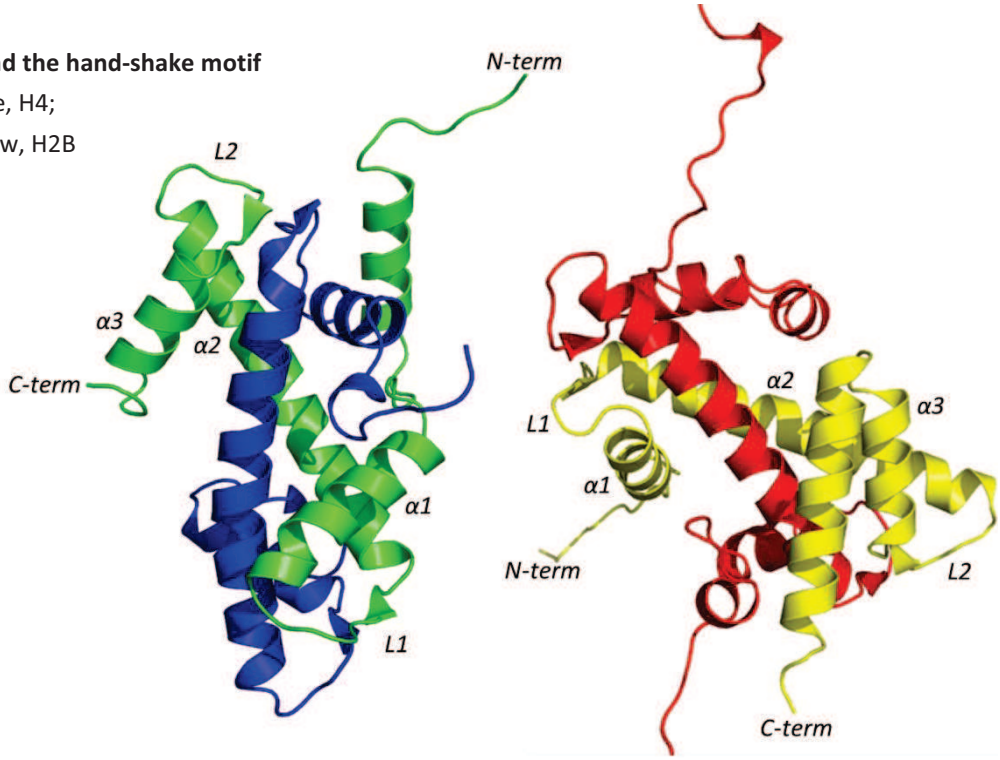


FIGURE 5a

The 4-helix bundle of H3-H3'

In green: H3

In dark green: H3'

(pdb : 1aoi)

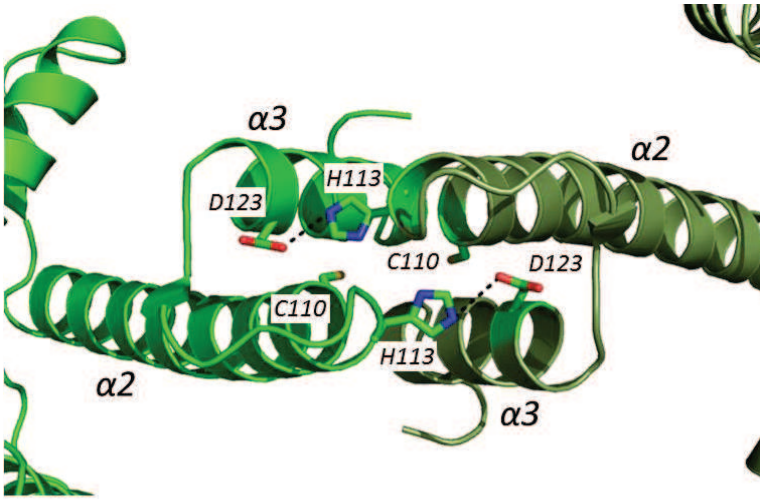


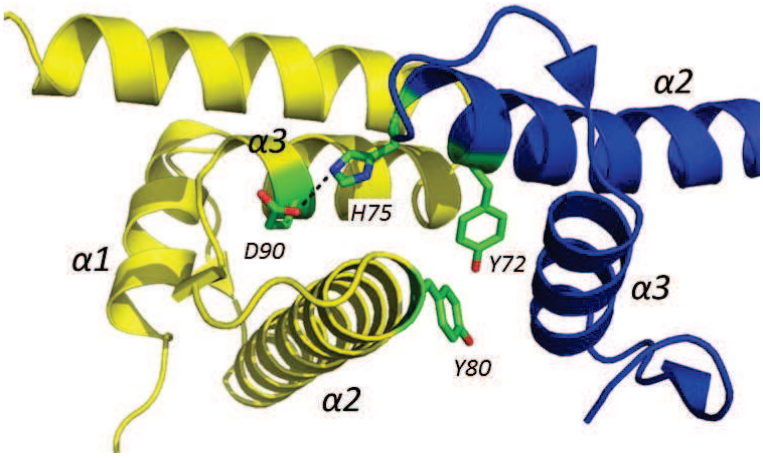
FIGURE 5b

The 4-helix bundle of H4-H2B

In blue: H4

In yellow: H2B

(pdb : 1aoi)



proportion of lysine and arginine residues[29]. This extremity is composed for H2A, H2B, H3 and H4 of 17, 35, 45 and 31 amino acids, respectively.

As we shall see later on, these tails are targeted by covalent modifications resulting in a large panel of physiological responses.

2.1.2.2 Histone H2A and its variants

Histone H2A is composed of 129 amino acids after removal of the initiator methionine, for a mass of around 14 kilo Daltons.

Histone H2A is the one that comprises the most variants, by far, totalling 19. These are coded by 26 genes, majorly organised in gene clusters *(table 1)*.

One H2A variant differs enormously in size. It is variant macro-H2A, a large protein of 327 amino acids, for a mass of around 40 kilo Daltons. It plays a specific role in X-chromosome inactivation[30] as well as in transcription regulation. H2A.Bbd (*Barr-body deficient*), on the contrary, is only present on active X-chromosome[31]. Other variants, like H2A.X or H2A.Z, are constitutively expressed throughout the genome and play an important role in DNA repair or thermal stress response.

2.1.2.3 Histone H2B and its variants

Histone H2B is composed of 126 amino acids after removal of the initiator methionine, for a mass of around 14 kilo Daltons.

Histone H2B is, together with H2A, the one that comprises the most variants, by far, totalling 19. These are coded by 23 genes *(table 2)*. However, out of the four histones, H2B remains the least studied so far.

2.1.2.4 Histone H3 and its variants

Histone H3 is composed of 135 amino acids after removal of the initiator methionine, for a mass of around 15 kilo Daltons.

Histone H3 comprises only 6 variants. These are coded by a set of 18 genes, divided into 2 gene clusters *(table 3)*. Among these variants, CENP-A is slightly larger in size, with a mass of around 16 kilo Daltons. CENP-A variant is specifically found in centromeric regions, and its N-terminal tail is highly divergent compared to the other histones. Another variant, the euchromatinian histone H3.3, only diverges by 4 residues compared to the canonical H3.1 histone. It is expressed constitutively and is mainly found in transcriptionally active regions[32].

2.1.2.5 Histone H4

Histone H4 is composed of 102 amino acids after removal of the initiator methionine, for a mass of around 11 kilo Daltons, being thus the smallest of all four histones.

Histone H4 doesn't have any variant. The unique canonical histone is coded by a set of 14 genes, divided into three gene clusters *(table 4)*. It is the most conserved histone throughout evolution.

TABLE 1
Histone H2A and its variants
From "HIstome: The Histone Infobase" (http://www.actrec.gov.in/histome)

Variant	Uniprot	Number of coding genes
Histone macro-H2A.1	O75367	1 (H2AFY)
Histone macro-H2A.2	Q9P0M6	1 (H2AFY2)
Histone H2A type 1	P0C0S8	5 (HIST1H2AG, HIST1H2AI, HIST1H2AK, HIST1H2AL, HIST1H2AM)
Histone H2A type 1-A	Q96QV6	1 (HIST1H2AA)
Histone H2A type 1-B/E	P04908	2 (HIST1H2AB, HIST1H2AE)
Histone H2A type 1-C	Q93077	1 (HIST1H2AC)
Histone H2A type 1-D	P20671	1 (HIST1H2AD)
Histone H2A type 1-H	Q96KK5	1 (HIST1H2AH)
Histone H2A type 1-J	Q99878	1 (HIST1H2AJ)
Histone H2A type 2-A	Q6FI13	2 (HIST2H2AA3, HIST2H2AA4)
Histone H2A type 2-B	Q8IUE6	1 (HIST2H2AB)
Histone H2A type 2-C	Q16777	1 (HIST2H2AC)
Histone H2A type 3	Q7L7L0	1 (HIST3H2A)
Histone H2A-Bbd type 1	P0C5Y9	1 (H2AFB1)
Histone H2A-Bbd type 2/3	P0C5Z0	2 (H2AFB2, H2AFB3)
Histone H2A.J	Q9BTM1	1 (H2AFJ)
Histone H2A.V	Q71UI9	1 (H2AFV)
Histone H2A.X	P16104	1 (H2AFX)
Histone H2A.Z	P0C0S5	1 (H2AFZ)

TABLE 2
Histone H2B and its variants
From "HIstome: The Histone Infobase" (http://www.actrec.gov.in/histome)

Variant	Uniprot	Number of coding genes
Histone H2B type 1-A	Q96A08	1 (HIST1H2BA)
Histone H2B type 1-B	P33778	1 (HIST1H2BB)
Histone H2B type 1-C/E/F/G/I	P62807	5 (HIST1H2BC, HIST1H2BE, HIST1H2BF, HIST1H2BG, HIST1H2BI)
Histone H2B type 1-D	P58876	1 (HIST1H2BD)
Histone H2B type 1-H	Q93079	1 (HIST1H2BH)

Histone H2B type 1-J	P06899	1 (HIST1H2BJ)
Histone H2B type 1-K	O60814	1 (HIST1H2BK)
Histone H2B type 1-L	Q99880	1 (HIST1H2BL)
Histone H2B type 1-M	Q99879	1 (HIST1H2BM)
Histone H2B type 1-N	Q99877	1 (HIST1H2BN)
Histone H2B type 1-O	P23527	1 (HIST1H2BO)
Histone H2B type 2-E	Q16778	1 (HIST2H2BE)
Histone H2B type 2-F	Q5QNW6	1 (HIST2H2BF)
Histone H2B type 3-B	Q8N257	1 (HIST3H2BB)
Histone H2B type F-M	P0C1H6	1 (H2BFM)
Histone H2B type F-S	P57053	1 (H2BFS)
Histone H2B type W-T	Q7Z2G1	1 (H2BFWT)
Putative histone H2B type 2-C	Q6DN03	1 (HIST2H2BC)
Putative histone H2B type 2-D	Q6DRA6	1 (HIST2H2BD)

TABLE 3
Histone H3 and its variants
From "HIstome: The Histone Infobase" (http://www.actrec.gov.in/histome)

Variant	Uniprot	Number of coding genes
Histone H3-like centromeric protein A	P49450	1 (CENPA)
Histone H3.1	P68431	10 (HIST1H3A, HIST1H3B, HIST1H3C, HIST1H3D, HIST1H3E, HIST1H3F, HIST1H3G, HIST1H3H, HIST1H3I, HIST1H3J)
Histone H3.1t	Q16695	1 (HIST3H3)
Histone H3.2	Q71DI3	3 (HIST2H3A, HIST2H3C, HIST2H3D)
Histone H3.3	P84243	2 (H3F3A, H3F3B)
Histone H3.3C	Q6NXT2	1 (H3F3C)

TABLE 4
Histone H4
From "HIstome: The Histone Infobase" (http://www.actrec.gov.in/histome)

Variant	Uniprot	Number of coding genes
Histone H4	P62805	14 (HIST4H4, HIST2H4A, HIST2H4B, HIST1H4A, HIST1H4B, HIST1H4C, HIST1H4D, HIST1H4E, HIST1H4F, HIST1H4H, HIST1H4I, HIST1H4J, HIST1H4K, HIST1H4L)

2.1.2.6 Regulating histones

2.1.2.6.1 At the gene level[33]

Histone genes are organized in one major cluster namely HIST1 and three minor clusters, namely HIST2, HIST3 and HIST4.

- HIST1 is located on chromosome 6 (6p21-p22) in human and chromosome 13 (13A2-3) in mouse. This cluster extends over more than 2.1 megabases. In human, it comprises 49 histone genes: 10 for H3, 12 for H4, 15 for H2B and 12 for H2A.
- HIST2 is located on chromosome 1 (1q21) in human and chromosome 3 (3F1-2) in mouse. It extends over 100 kilobases. In human, this cluster comprises 6 histone genes: one for H3, one for H4, one for H2B and three for H2A.
- HIST3 is also located on chromosome 1 (1q42) in human and chromosome 11 (11B2) in mouse. It extends over 35 kilobases. In human, it comprises 3 genes: one for H2B, one for H2A and one for a modified canonical H3 histone (Histone H3.1t), only expressed in primary spermatocytes.
- finally, HIST4 is located on chromosome 12 (12p13.1) in human. It is the smallest of all four clusters since it only comprises one gene coding for histone H4.

2.1.2.6.2 At the RNA level

Canonical histones are expressed in a replication-dependent manner, and shall, at each cell cycle, reach a very high concentration during the S phase, corresponding to the DNA replication phase. This contrasts with histone variants, whose genes are expressed in a non-replication-dependent manner and throughout the whole cell cycle.

Canonical histone mRNAs exhibit, like all eukaryotic mRNAs, a 7-methylguanosine (or cap) at their 5'-end. However, they don't possess any polyadenylated 3'-end, but instead, a stem-loop structure of 6 base-pairs for the stem and 4 bases for the loop[34]. These stem-loop are highly conserved among all metazoans. They are located at a hundreds of nucleotides downstream the stop codon. One unique protein is able to bind to this structure, SLBP (Stem-Loop Binding Protein), contributing thus to the metabolism of these particular mRNAs. The basal expression if this protein during G1 phase is then increased by a factor 10 when entering the S phase, and falls down quickly at the end of the S phase, after phosphorylation-dependent degradation.

Finally, these mRNA do not have any introns. Their maturation only consists in a 3'-end endonucleolytic cleavage, between the stem-loop and a purine-rich region called HDE (Histone Downstream Element)[35]. To achieve this reaction, the spliceosomal RNA U7 (snRNA U7) binds by base-complementarity to the HDE, further stabilised by SLBP bound to the stem-loop. After cleavage, SLBP remains on the mature mRNA, escorts it in the cytoplasm, where translation can then occur (*figure 6*).

This stem-loop is the critical regulatory element of these histone mRNAs. During S phase, the half-life of a histone mRNA is around 45 to 60 minutes; this half-life falls to 10 minutes at the end of the S phase. It has been suggested that SLBP would protect the mRNA during S phase, and its degradation at the end of the S phase leads to a quick degradation of the mRNA itself.

It has moreover been suggested that, although mRNAs and SLBP accumulate at the same moment in the cell cycle, regulatory signals controlling their expression are not the same. SLBP is regulated by cell cycle signals, whereas mRNA are directly regulated by the histone demand, thus by the DNA synthesis[36].

2.1.2.6.3 At the protein level

In 2003, Gujan carried out studies on Rad53, a kinase required of DNA repair and replication[37]. He was then able to show that Rad53 was sensitive to histone overexpression. This protein could indeed detect an excess of histones and address them to the degradation pathway.

2.1.3 Properties of the nucleosomal DNA

Nucleosomal DNA is a B-form DNA. It ideally comprises 145 to 147 base-pairs, wrapped around the histone octamer in a 1.67-turn left-handed superhelix[25].

Despite all appearances, the nucleosome symmetry isn't perfect: indeed, one could think that a 146-base-pairs DNA would be organized around the octamer in such a way that the dyad axis would pass between the 73rd and 74th base-pairs, dividing the DNA into two equal halves. Though, it is preferentially organized with a base-pair on the dyad axis. The nucleosomal DNA is thus divided into a large 73-base-pairs half, and a small 72-base-pairs half. The rotational orientation of the DNA is described according to this central base-pair. It is called SHL0 (SuperHelix Location zero), and for each helix turn, this position number increases until SHL+7 on the 73-base-pair DNA side, and decreases until SHL-7 on the 72-base-pair DNA side (*figure 7*).

Among the 146 base-pairs, 129 are directly organized around the histone octamer, in a 1.59-turn left-handed superhelix. The remaining base-pairs at each extremity have thus a negligible participation in the curvature of the DNA, which explains their higher affinity for protein factors compared to the rest of the nucleosomal DNA. Nevertheless, some counterexamples can be mentioned, like the HIV-1 integrase[38], which binds preferentially to highly curved DNA on the nucleosome rather than uncurved nucleosomal DNA or naked DNA.

Each of the four pairs of histone fold in the octamer is associated with 27 to 28 DNA base-pairs, leaving 4 free base-pairs between each segment. The region of the $(H3\text{-}H4)_2$ tetramer binds the central part of the nucleosomal DNA, from SHL-3 to SHL+3, while the (H2A-H2B) dimers binds the extremities of the nucleosomal DNA, from SHL-6 to SHL-3 and SHL+3 to SHL+6.

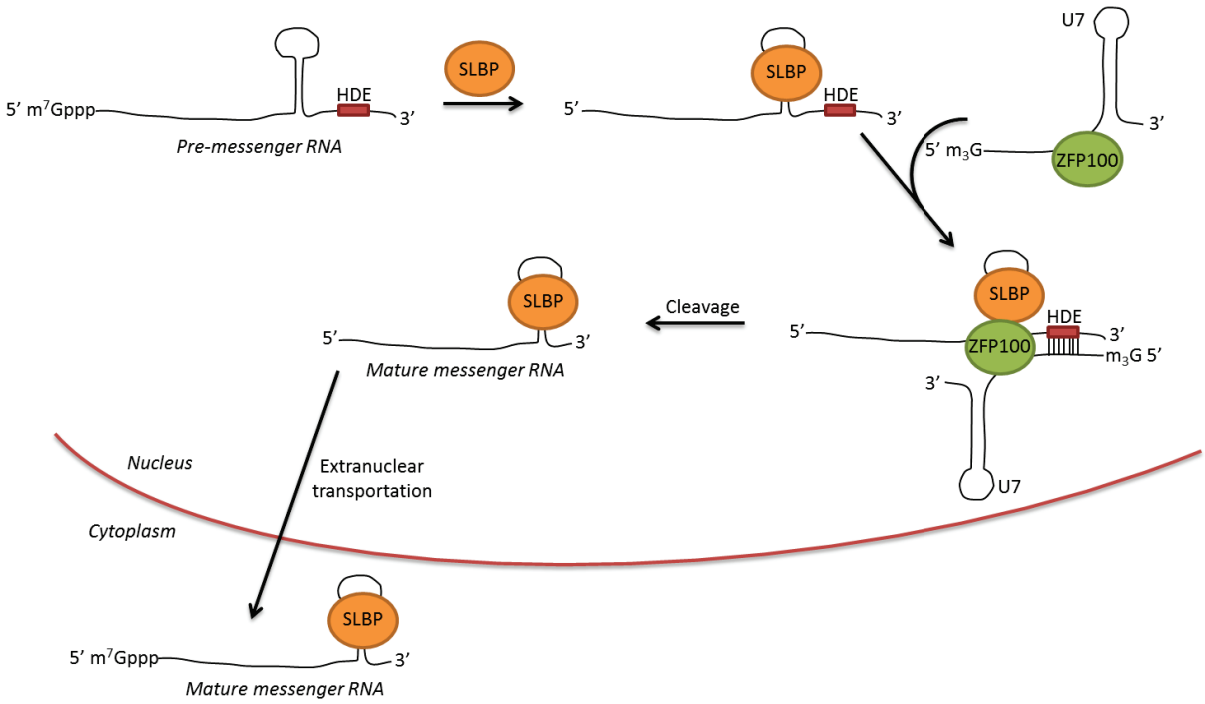


FIGURE 6

Histone mRNA maturation

From Marzluff W.F. & R.J. Dunorio. Histone mRNA expression: multiple levels of cell cycle regulation and important developmental consequences. Curr Opin Cell Biol. 2002, 14:692–699

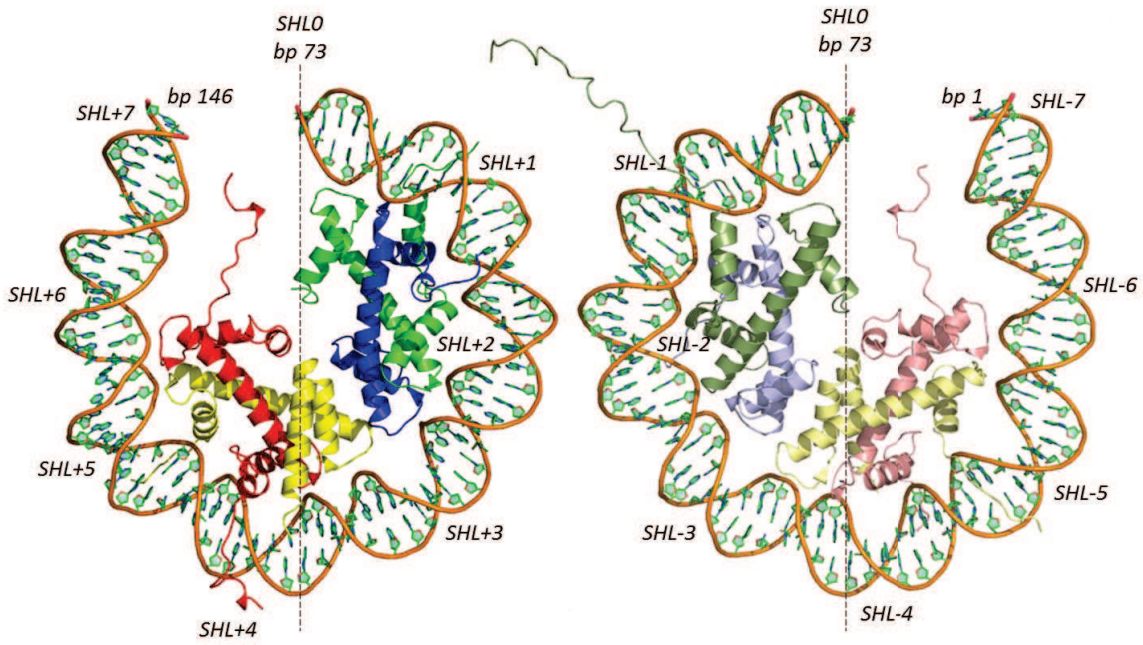


FIGURE 7

Nucleosomal pseudo-symmetry

The pseudo 2-fold axis is aligned vertically with the DNA, passing through a central DNA base-pair (bp 73), and dividing the DNA into two halves. Both are represented separately: on the left, the long 73 bp half; on the right, the short 72 bp half. *(pbd : 1aoi)*

2.1.4 Formation and dynamic of the nucleosome

Within the nucleosome, the compacted DNA is quasi-inaccessible to protein factors involved in the various cellular processes. Imbalzano and then later Godde notably demonstrated that the recognition by TBPs (TATA-box binding proteins) of a TATA-box located in a nucleosome is strongly diminished, leading to a decreased transcription activity[39,40]. In this regard, Schieferstein & Thomas showed, back in 1998, that DNA repair was inhibited *in vitro* if the damage was located in a nucleosome[41]. The question raised was thus to understand how nucleosomal DNA can be rendered accessible to protein effectors *in vivo*. To answer this question, Kimura and Jamai carried out experiments, on fluorescence recovery after photobleaching and chromatin immunoprecipitation, respectively, and showed that the nucleosome is a dynamic structure, that assembles and disassembles *in vivo*, according to the needs[42,43]. Furthermore, the kinetics for this disassembly/reassembly process has to be fast, in the second range[44].

2.1.4.1 Assembly dynamic and chaperoning

Karolin Luger published in 1999 a fully detailed protocol in which she described production and purification of recombinant histones, as well as the reconstitution of nucleosomes *in vitro*[45]. This reference protocol is based upon a spontaneous assembly of the nucleosome in presence of the four histones and a fragment of double-stranded DNA of adequate size (between 145 and 147 base-pairs), through a simple dialysis from 2 M to 250 mM potassium chloride. The reverse operation on the other hand can lead to a progressive disassembly of the reconstituted nucleosome. However, these observations do not give any information on the dynamic, strictly speaking, of assembly and disassembly of the nucleosome.

To date, two different schools stand out: the first one pleads for a sequential mechanism, in which the two (H2A-H2B) dimers move away from the central $(H3\text{-}H4)_2$ tetramer, allowing the latter to dissociate and finally release the DNA[46-50]; the second one supports a one-step-mechanism, in which the DNA frees the octamer which dissociate then entirely[51-54].

The numerous developments in single-pair Förster resonance energy transfer (spFRET) over the past few years allowed to highlight conformational states of the nucleosome[55,56] and to answer in a somewhat objective manner, though not definitive, to these two theories. Figure 8 therefore shows different conformational states observed by Buning & Van Noort and by Böhm *et al.*:

- State 1 corresponds to the nucleosome in its maximal folding state, as observed in X-ray crystallography. It comes as a stable state, observed *in vivo* at low ionic strength.
- State 2 was suggested for the first time in 1995 by Polach & Widom. It corresponds to a decompaction over the 30 first base-pairs of the DNA on each side of the nucleosome. DNA is "breathing" and is thus partly accessible to protein effectors.
- State 3 represents the histone octamer without DNA, as it has been crystallised by Arents *et al*. in 1992. However, this state has never been observed *in vivo*.

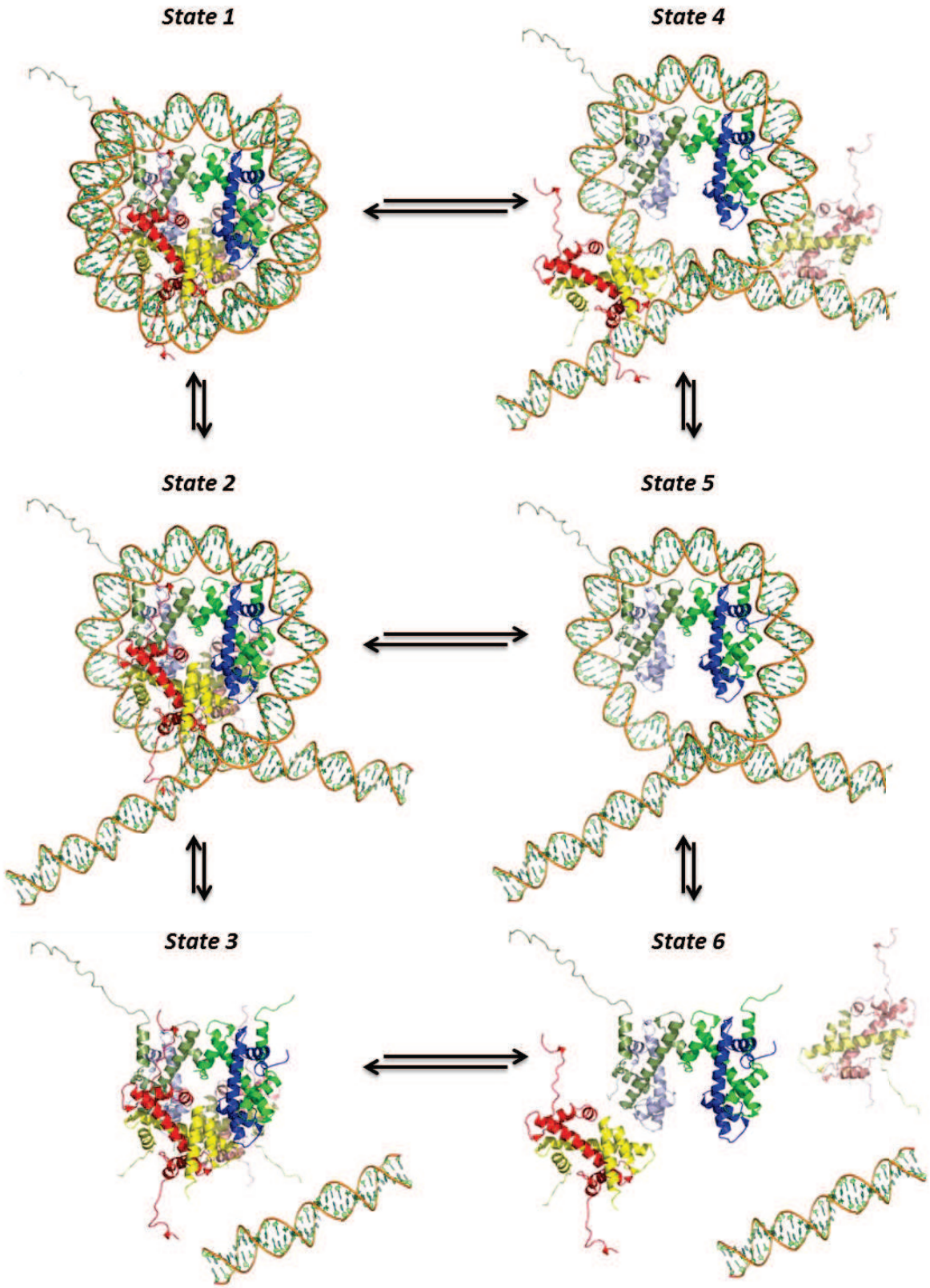


FIGURE 8

Dynamic of nucleosome formation

Modified from Bohm, V. et al. Nucleosome accessibility governed by the dimer/tetramer interface. Nucleic Acids Res 39, 3093-3102 (2011) (pbd : 1aoi).

- State 4 shows the nucleosome in a conformation in which DNA is fully binding the histone proteins, but where the interactions between the central $(H3\text{-}H4)_2$ tetramer and the (H2A-H2B) dimers are partially lost.
- State 5 corresponds to the $(H3\text{-}H4)_2$ tetramer, interacting with nucleosomal DNA, (H2A-H2B) dimers being absent. This state is probably the one that gave rise to the more controversies. This structure, called tetrasome, had indeed been proposed from 1991 by Dong & Van Holde, who demonstrated that the $(H3\text{-}H4)_2$ tetramer could bind DNA *in vitro*[57]. This observation was of course in agreement with a sequential assembly dynamic of the nucleosome, although nearly twenty years were needed to successfully demonstrate the existence of this state *in vivo*.
- State 6 illustrates finally a complete dissociation in one single step of the histones and the DNA.

In a more visual manner, the theory of a sequential mechanism could be summarized by the transition 1 → (4 or 2) → 5 → 6, while the one-step-mechanism would correspond to the sequence 1 → 2 → 3 → 6.

Without addressing the pros and cons of each theory in this manuscript, since it's not the aim of this thesis, the theory of a sequential mechanism has seemed to be confirmed these last few years by an increasing number of results. First of all, the salt effect has been one of the first subject matter in the history of nucleosome dynamics. Monomeric histones don't exist in solution, being too unstable and labile. It is however possible to observe (H2A-H2B) dimers and $(H3\text{-}H4)_2$ tetramers in physiological conditions. Nevertheless, they don't interact together unless in presence of DNA. The histone octamer is thus not a stable structure in physiological conditions, and is only observable at 2 M salt. Finally, the sequential model is also confirmed by the differences in residency time of histones on DNA *in vivo*: when the (H2A-H2B) dimer only stays for a few minutes on DNA, the $(H3\text{-}H4)_2$ tetramer can, on the contrary, stay bound to the DNA in a stable state for several hours. This observation vouches thus for a probably independent dynamism of the different components of the nucleosome[43,58].

Assembly and disassembly of the nucleosome *in vivo* implies, in addition to histones and DNA, a variety of other temporary partners: the histone chaperones. It was in 1978 that Laskey described them for the first time[59]. Since then, their role has been studied and their function evolved. Like so, these chaperones make it possible to evict nucleosomes in promoter regions to regulate transcription[60], participate to the regulation of histone post-translational modifications[61,62] and prevent mismatches between histones and DNA[63].

Classically, one can distinguish chaperones according to their affinity for either the (H2A-H2B) dimer or the $(H3\text{-}H4)_2$ tetramer. In this way, FACT, NAP1, Chz1, nucleophosmin (NPM), nucleoplasmin or nucleolin are related to the (H2A-H2B) dimer, whereas Asf1, Spt6, HIRA and CAF1 are rather related to the $(H3\text{-}H4)_2$ tetramer, wrongly since we know now that Chz1 for example is

TABLE 5
Histones chaperones
Modified from De Koning, L. et al. Histone chaperones: an escort network regulating histone traffic. Nat. Struct. Mol. Biol. 14, 997-1007 (2007).

Chaperone classification		Chaperones	Target	Main functions
Class 1 Single chaperones	H3 H4	Asf1	H3.1-H4 H3.3-H4	Histone recruiter for CAF-1 and HIRA
		Fkbp39p	H3-H4	rDNA repression
		N1/N2/Nasp	H3-H4	H3-H4 storage in oocytes
		Spt6	H3-H4	Transcription (Initiation/Elongation)
		Rtt106	H3-H4	Heterochromatin silencing
		HJURP/Scm3	CENP-A-H4	Centromeric chromatin
	H2A H2B	Nucleoplasmin Nucleophosmin	H2A-H2B	H2A-H2B storage in oocytes; Cytosol/nucleus transportation
		Chz1	H2A.Z-H2B	H2A.Z incorporation by SWR1
		Nap1 Nap1L2 SET/TAF1b/CINAP/Vsp75	H2A-H2B	Cytosol/nucleus transportation; Transcription; Replication
		Nucleolin	H2A-H2B	Elongation; Remodelling
Class 2 Multi-chaperones complexes	H3 H4	CAF-1 complex	H3.1-H4	Synthesis-dependent incorporation (repair, replication)
		HIRA/Hir complex	H3.3-H4	Synthesis-independent incorporation (transcription, decondensation)
	H2A H2B	FACT complex	H2A-H2B H3-H4	Transcription elongation
Class 3 Chaperones within enzymatic complexes	H3 H4	Hif1	H3-H4	Assists Hat1/Hat2
		Rsf-1	H3-H4	Assists RSF
	N.D. Multiple	Arp4	?	Assists HAT
		Arp7/Arp9	?	Assist SWI/SNF, RSC, BAP, BAF
		Arp8	H3-H4	Assists INO80
		Acf1	H2A-H2B H3-H4	Assists ACF/CHRAC
Multiclass chaperones	H3 H4	RbAp46 RbAp48	H3.1-H4 H3.3-H4 CENP-A-H4	Synthesis-dependent incorporation (replication, enzymatic activities assistance, centromeric chromatin maintenance)

specific for the H2A.Z variant[64] and that NAP1 has similar affinity towards both (H2A-H2B) dimer and $(H3\text{-}H4)_2$ tetramer[65] (*table 5*).

The usefulness of chaperones in general, and in our case in particular of histone chaperones, would appear to be beyond doubt. However, as explained previously, assembly and disassembly mechanisms remain poorly understood. In 2006, English published the structure of a (H3-H4) dimer in complex with the Asf1 chaperone[66]. Without providing all the answers, this structure revealed that Asf1 could interact with a (H3-H4) dimer only, and not a tetramer. English suggested then that Asf1, while interacting with H4 through the H4-H2A interaction site, could destabilise the interactions between the (H2A-H2B) dimer and the central tetramer. After eviction of both (H2A-H2B) dimers, Asf1 could fully interact with H3 through the H3-H3' interaction site required for tetramer stability. This ultimate destabilisation would finally cause the final disassembly of the nucleosome. Donham, in 2011, yet raised a reserve concerning this mechanism, by showing that Asf1 could not interact with a tetrasome *in vitro*[67]. It is probable that other factors, protein or biochemical, shall be required to allow Asf1 to interact with the (H3-H4) dimer.

2.1.4.2 DNA-histone binding sites

Nucleosome is a highly charged structure, globally anionic. The DNA, with its 146 base-pairs, provides two negative charges for each phosphate group, giving a total of 292 negative charges. The histone octamer on its side is globally basic, since it comprises 220 cationic amino acids (arginines and lysines) and only 74 anionic amino acids (aspartate and glutamate), giving a net total of 146 positive charges. There is thus an imbalance between the charges brought by the DNA and those brought by the histones. This difference seems, consequently, to be a crucial factor for chromatin compaction.

DNA binding sites are divided into two categories: first of all, α1α1 sites, composed by two α1 helices from a pair of histones (*figure 9*); then, L1L2 sites, composed by L1 and L2 loops, and by the N-terminal part of one α1 helix and the C-terminal part of the other α1 helix from a pair of histones[25] (*figure 10*).

Generally, five characteristics are observed when the deoxyribophosphate backbone of the DNA faces the histone octamer:

- the N-terminal part of α1 helices of H3, H4 and H2B, as well as α2 helices of the four histones are implicated in a hydrogen bond with a phosphate group of the DNA.
- the hydrogen bonds with phosphate groups always involve the nitrogen atom of the peptide chain from amino acids located in the last turn (or close to the last turn) of α1 and α2 helices.
- the side chains of arginine residues travel through the minor groove of the DNA 14 times. Ten times out of 14, these arginines belong to the histone fold. The four remaining times, they belong to the N-terminal tails.
- long-range contacts are possible with deoxyribose groups, on the DNA backbone.

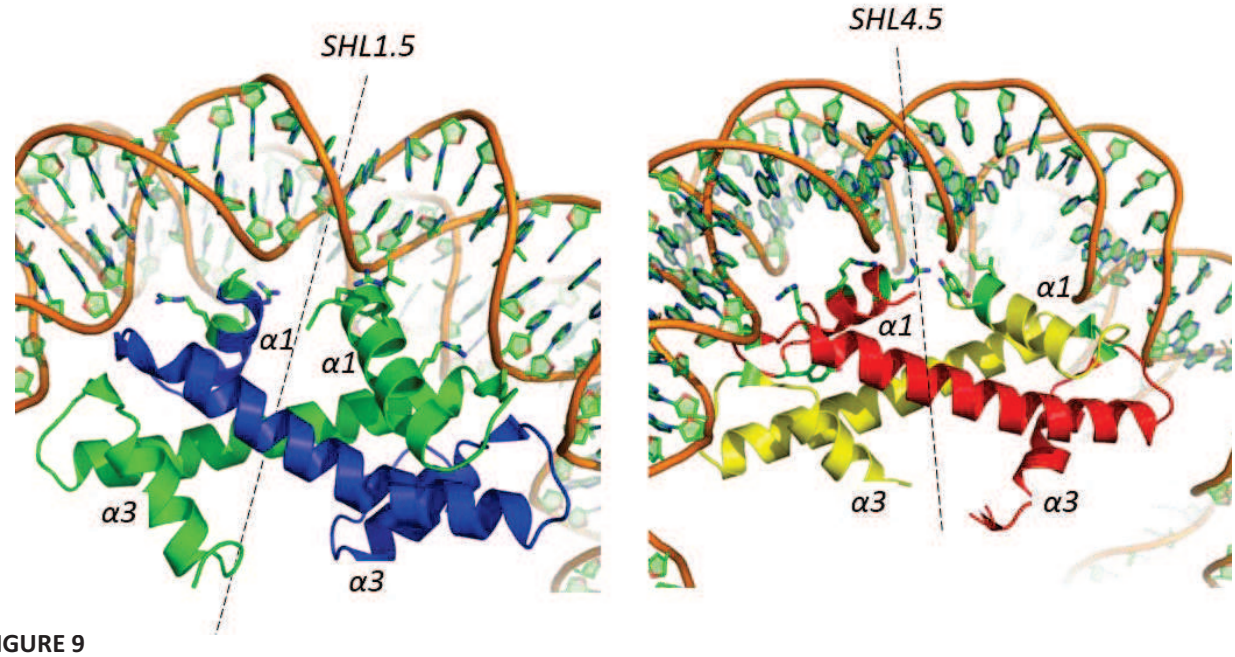


FIGURE 9

α1α1 interaction sites

On the left: H3-H4 dimer (only α1-L1-α2-L2-α3 are shown). A pseudo 2-fold axis runs vertically, through SHL1.5. In green, H3; in blue, H4.

On the right: H2A-H2B dimer. Same as H3-H4, but the pseudo 2-fold axis crosses SHL4.5. In red, H2A; in yellow, H2B. Note H2A α1 helix orientation, which is different from the other α1 helices.

(pdb : 1aoi)

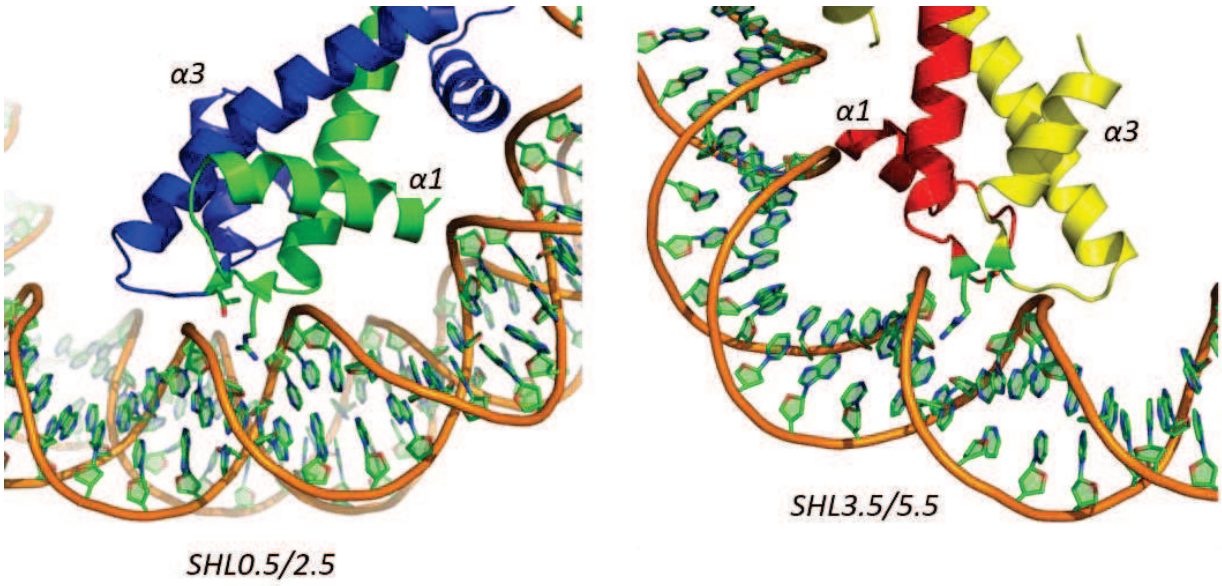
FIGURE 10

L1L2 interaction sites

On the left: H3-H4 L1L2 site, at SHL+2.5. In green, H3; in blue, H4.

On the right: H2A-H2B L1L2 site, at SHL+3.5. In red, H2A; in yellow, H2B.

(pdb : 1aoi)



- hydrogen bonds and salt bridges are often observed between oxygen atoms from the phosphate groups of the DNA and the hydroxyl groups as well as the basic groups of amino acid side chains of histones. These side chains bind oxygen atoms of the phosphate groups when the DNA backbone is flipped inside the nucleosome.

In α1α1 sites, H3, H4 and H2B have their α1 helix N-terminal part pointing towards a phosphate group. For H3 and H4, in position SHL+1.5, the side chain of the proline located in the first turn of α1 helix interacts with the deoxyribose group of the DNA backbone (H3-P66 and H4-P32), while the nitrogen atom of the peptide chain of residues K64 and L65 of histone H3 hydrogen-bonds the oxygen atom of the DNA phosphate group. In the case of H2B, in position SHL+4.5, the residue I36 makes contact with the deoxyribose group while the nitrogen atom of the peptide chain of residue S33 hydrogen-bonds the oxygen atom of the DNA phosphate group.

On the side of the α1 helix that points towards the major groove of DNA, the side chains of arginine, lysine, histidine and tyrosine generally make hydrogen bonds with the phosphate groups of DNA (H3-R69, H3-R72, H4-R35, H4-R36, H2A-R29, H2A-R32, H2A-R35 and H2B-Y37) or electrostatic interactions. In one single case, the residue L65 manages to reach the major groove of DNA and makes a hydrophobic interaction with the 5-methyl group of a thymidine.

Concerning the α1 helix of H2A, the presence of a tyrosine 39 in loop L1 causes a different orientation compared with the three others. Thus, only arginines 29 and 32 bind to the DNA backbone. No contacts are made with the deoxyribose groups.

In a L1L2 site, the N-terminal end of a α2 helix points toward a phosphate group, and, in the case of H3, H4 and H2A, the side chain of an arginine residue in the L1 loop penetrates the minor groove of the DNA. The nitrogen atoms of the peptide chain in the L1-α2 junction make hydrogen bonds with the phosphate groups of the DNA. Arginine residues in loops L1 of H3, H4 and H2A (H3-R83, H4-R45 and H2A-R42) form hydrogen bonds with the hydroxyl group of threonines in the neighbouring loop L2 (H4-T80, H3-T118 and H2B-T85, respectively). These interactions probably aim to restrict the conformation of the arginine side chain, avoiding hydrogen-bonding with the O2 atoms of cytosine or thymidine, or the N3 atoms of adenine or guanine. In one case however, H4-R45 makes a hydrogen bond with the O2 atom of a thymidine in position SHL+0.5, while maintaining its hydrogen bond with the threonine residue.

In the sequence of H2B, the arginine residue in the loop L1 is replaced by a glycine (H2B-G50) and the threonine in the neighbouring loop L2 is replaced by an arginine residue (H2A-R77), which penetrates inside the minor groove of the DNA in position SHL+5.5.

In 2002, the X-ray structure of the nucleosome at 1.9 Å resolution allowed to highlight new interactions stabilising the DNA molecule around the histone octamer, involving water-mediated hydrogen bonds[26]. In total, 121 water molecules interact both with the DNA and the histone octamer, facilitating thus the interactions between the two partners. No less than 116 direct hydrogen bonds and 358 indirect water-mediated hydrogen bonds were thus observed in this new

structure. The distribution of these new indirect bonds has a 1.4 to 1 ratio between the DNA backbone and the DNA bases; and a 1 to 2 ratio between the peptide chain and the side chains of the amino acids in the octamer. Comparatively, the direct hydrogen bonds between the DNA and the octamer have a ratio of 6.7:1 between the backbone and the bases of the DNA; and a ratio of 1:1.3 between the peptide chain and the side chains of the amino acids in the octamer.

Even though the majority of the interactions between the histones and the DNA involves the phosphate groups of the DNA, crucial interactions take place via the side chains of the histones penetrating the minor groove of the DNA. These side chains belong to arginine residues, extending in twelve out of the fourteen minor grooves facing the octamer. These residues (H4-R45, H3-R63, H2A-R42, H2B-R30, H2A-R77) are strictly conserved in canonical histones of all species, and, with the exception of H2A-Bbd, also are in histone variants. Eight times out of twelve, a threonine residue of the neighbouring histone intervenes to form a hydrogen bond with the guanidine group of the arginine residue, in a direct-manner (H3-T118 with H4-R45) or indirect-manner through a water molecule (H4-T80 with H3-R83, H2B-T85 with H2A-R42 and H2A-T76 with H2A-R77). It is assumed that these arginine side chains act as ratchets to restrain the sliding of nucleosomes along the DNA molecule.

Together, the twenty side chains concerned by these interactions lead to a loss of 2962 Å² of accessible surface area (ASA), that is to say, about ¼ of the total ASA loss in the nucleosome. Indeed, the total ASA of the nucleosome being 74049 Å², and that of the DNA alone and octamer alone being respectively 52410 Å² and 34310 Å², the total loss due to DNA/protein interactions can be estimated up to 12671 Å².

To finish, N-terminal ends of histones H3 and H2B have unstructured tails traveling through the gyrus of the DNA superhelix. For H3, five amino acids (H39, R40, Y41, R42 and P43) go through a tunnel in the superhelix formed by the superimposition of the minor grooves in position SHL+6.7/SHL-0.7 and SHL-6.7/SHL+0.7. For H2B, an extremely basic tail of 8 amino acids (K24, K25, R26, R27, K28, T29, R30 and K31) crosses over the superhelix in location SHL+4.7/SHL-2.7 and SHL-4.7/SHL+2.7. A periodicity of these histone tails can thus be observed in the minor grooves of the DNA, around every twenty base-pairs.

2.2 Higher chromatin compaction orders

2.2.1 The beads on a string

Also known as the 10 nm fibre, this string of beads was observed for the first time in 1974 by Olins & Olins[20]. The picture recorded by electron microscopy showed a repetition of nucleosomes, spaced by a few tens of base-pairs. This structure was observed only at low salt concentration (less than 5 mM NaCl) and it lacked an essential partner for chromatin compaction: histone H1, playing a role of internucleosomal linker. When increasing the ionic strength, the DNA condenses into

unstructured corpuscles. Thus, knowing that the physiological NaCl concentration is around 154 mM, the beads on a string most probably doesn't exist *in vivo* (*figure 11*).

2.2.2 The 30 nm fibre

In 1976, Finch and Klug were the first to suggest that the nucleosome, in presence of linker histone H1 or magnesium ions, is compacted to form a "30 nm chromatin fibre"[68] (*figure 11*). This suggestion follows on from their observations in TEM. The model they proposed, known as the solenoid model, shows nucleosomes constitutively lined up in the fibre, forming a simple helix.

Later, another model, called the zigzag model, was proposed[69]. In this conformation, the nucleosomes aren't constitutively aligned. Instead, the nucleosome "n" is linked to the nucleosome "n+2", but not to the "n+1". In 2004, Richmond showed that his results support this second model[70]. Furthermore, he published the first X-ray structure of a tetrasome at 9 Å resolution[71], confirming the zigzag model.

Then, recently, Rhodes achieved *in vitro* reconstitution of long chromatin fibres, using a long DNA, histone octamers and the linker histone H5, a H1 isoform found in avian erythrocytes[72,73]. Electron microscopy then showed that this fibre was indeed a solenoid, but different from the one described in 1976. Instead, it would rather be an interdigitated solenoid, in which the nucleosome "n" interacts with nucleosomes "n+5" and "n+6".

To date, no absolute certainty is established but Rhodes suggested that the two conformations – interdigitated solenoid and zigzag – only depend on the length of the internucleosomal DNA[74]. More recently, Grigoryev showed that both conformations could coexist within the same fibre under certain conditions[75].

2.2.3 The metaphase chromosome

The metaphase chromosome is the chromatin maximal compaction state that one can observe. It appears during mitosis and reaches its highest level of compaction during metaphase, when the genetic material is moving towards the equatorial plate (*figure 11*).

In 1880, Walther Flemming discovered that stainable basophilic fibres where the major constituent of the Eukaryotic nucleus, and that they play a role, then unknown, in cell division. Eight years later, Heinrich Waldeyer named these fibres "chromosomes", or coloured bodies.

This ultimate stage of chromatin compaction is to date still poorly understood. This is mainly due to the fact that chromosome studies are laborious because of their instability and ephemeral nature.

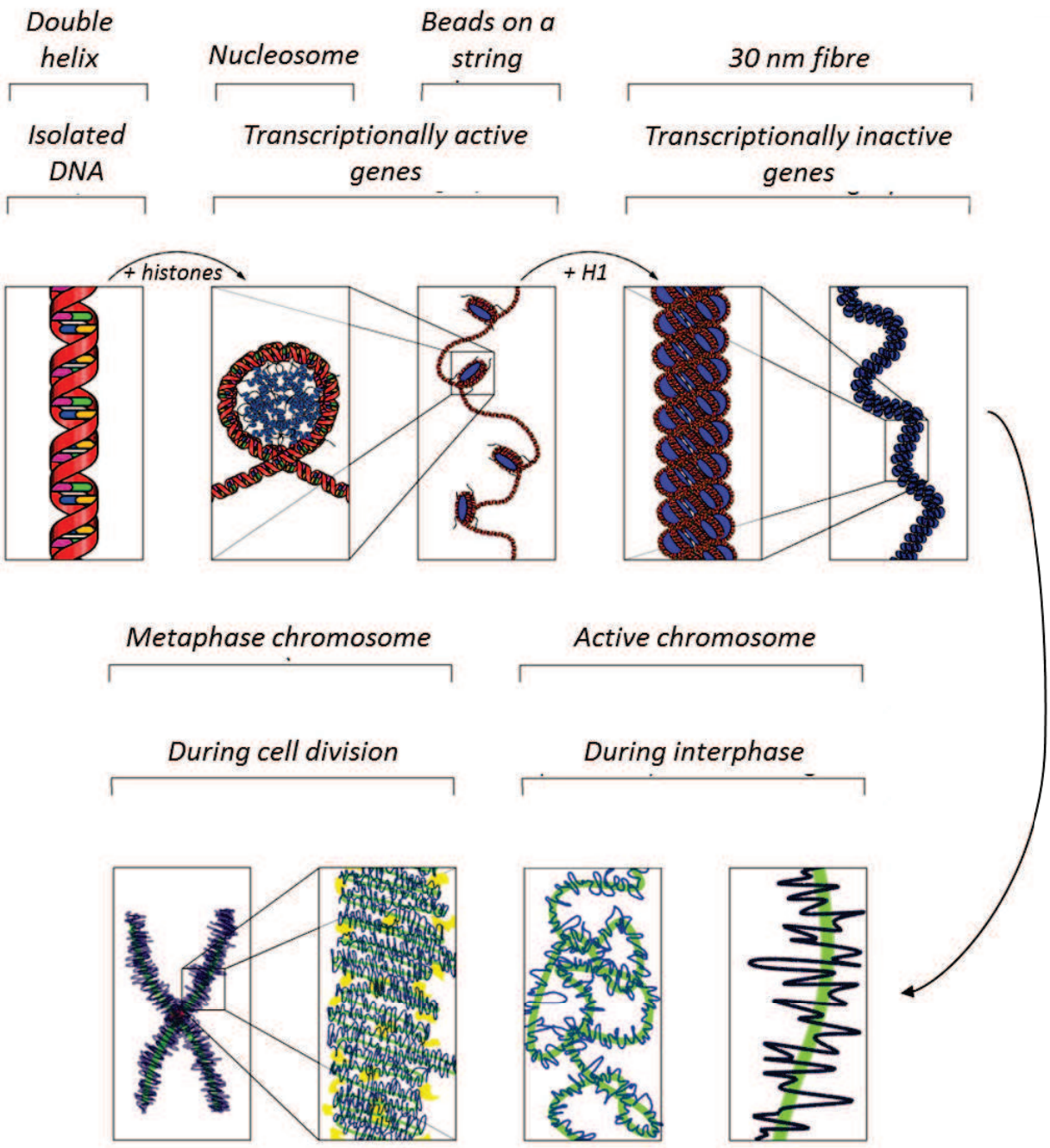


FIGURE 11

The different compaction states of chromatin

Modified from Richard Wheeler. The major chromatin structures. Wikimedia commons (2005).

3. The chromatin remodelling

During the various stages of development of an organism, the cell fate is played on activation and/or repression of specific genes, coding for proteins that are more or less important and determining. The chromatin must thus be remodelled in order to ensure accessibility of the DNA to transcription factors, or, on the contrary, to block the access.

Within the chromatin, the DNA and the histones can be targeted by covalent modifications. These modifications have two roles: a direct role, in the case for example of histone acetylation or deacetylation, which influences the compaction state of the chromatin. Or an indirect role, by recruiting enzymes or other protein partners ensuring the remodelling.

3.1 DNA methylation

Methylation of the DNA is due to the covalent addition of a methyl group on a cytosine to form a 5-methylcytosine (5mC) (*figure 12*). This modification is found in most of the living beings, from bacteria to human. In bacteria for example, this modification helps differentiate between endogenous and exogenous DNA, non-methylated, e.g., resulting from a phage infection. In Eukaryotes, some organisms nonetheless have lost this modification, like the yeast *Saccharomyces cerevisiae* or the worm *Caenorhabditis elegans*[76]. In plants, cytosine methylation only occurs in CpG, CpXpG and CpXpX islands (with X = A, C or T). Finally, in mammals, only CpG islands are subject to methylation by methyltransferases.

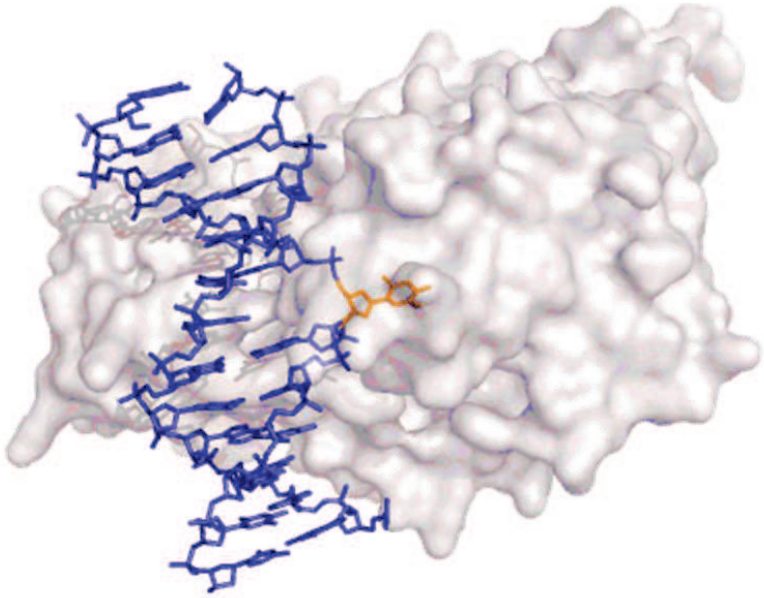
3.1.1 The functions of DNA methylation

In mammals, DNA methylation and demethylation are involved in various processes, like embryogenesis[77,78], stem cells differentiation[79,80], X-chromosome inactivation[81,82], genomic imprinting[83,84] or extinction of transposable elements[85]. Finally, DNA methylation is heavily involved in cancer development[86].

3.1.1.1 In healthy cells

Complex changes of methylation levels occur during embryonic development. Right after fertilization of a female gamete by a male gamete, the methylation level of both genomes, maternal and paternal, differs enormously, the first one being way less methylated than the second one[77]. Within 3 to 6 hours, the maternal genome is rapidly methylated, while the paternal genome is passively demethylated; finally, the maternal genome is in turn passively demethylated in such a way that after a few divisions, the methylation levels should not differ anymore. *De novo* methylation occurs only after embedding of the blastocyst, and this methylation is targeted: its level increases quickly in the ectoderm but remains low in the trophoblast or primitive endoderm[87,88] (*figure 13*).

a)



b)

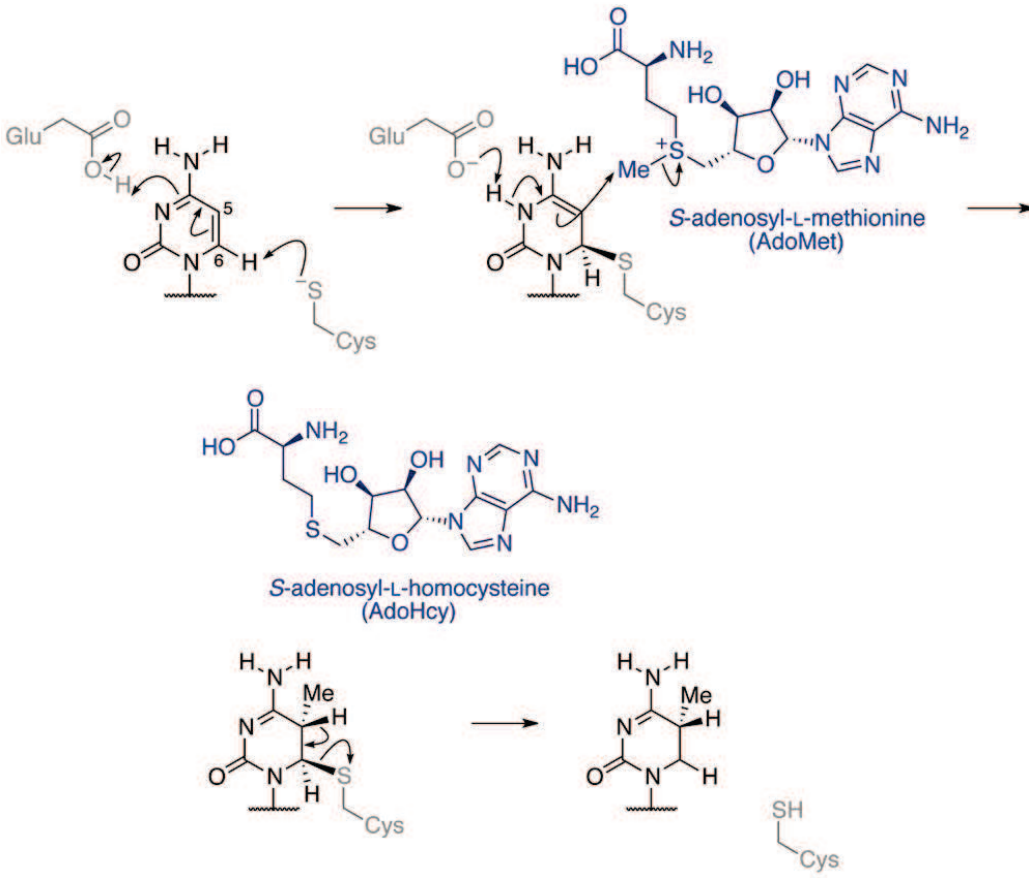


FIGURE 12

Cytosine methylation

a) The structure of DNMTs in complex with DNA shows how the cytosine is flipped out, allowing its binding to the active site.

b) DNMTs have a conserved cysteine, whose thiolate acts as a strong nucleophile and attacks the cytosine. The resulting negative charge of the cytosine is stabilized by a glutamate residue from the DNMT. The nucleophilic attack of the cytosine on the SAM cofactor allows then transferring a methyl moiety, before the final β-elimination, releasing the 5-methylcytosine.

Modified from Tom Brown & Tom Brown Jr. Nucleic Acids Book

DNA methylation is also essential for cell differentiation and the proper functioning of differentiated cells. Pluripotency genes are usually hypomethylated in stem cells and methylated during cell differentiation[89]. In this respect, it has been shown in 2009 that demethylation of the Oct4 and Nanog pluripotency genes promoters by AID (Activation-induced cytidine deaminase), could be useful to reprogram *in vitro* neurones into pluripotent cells[90,91].

Somatic tissues each express different genes, to ensure their specific functions. It was thus shown that DNA methylation was critical for regulating tissue-specific gene expression[92-94].

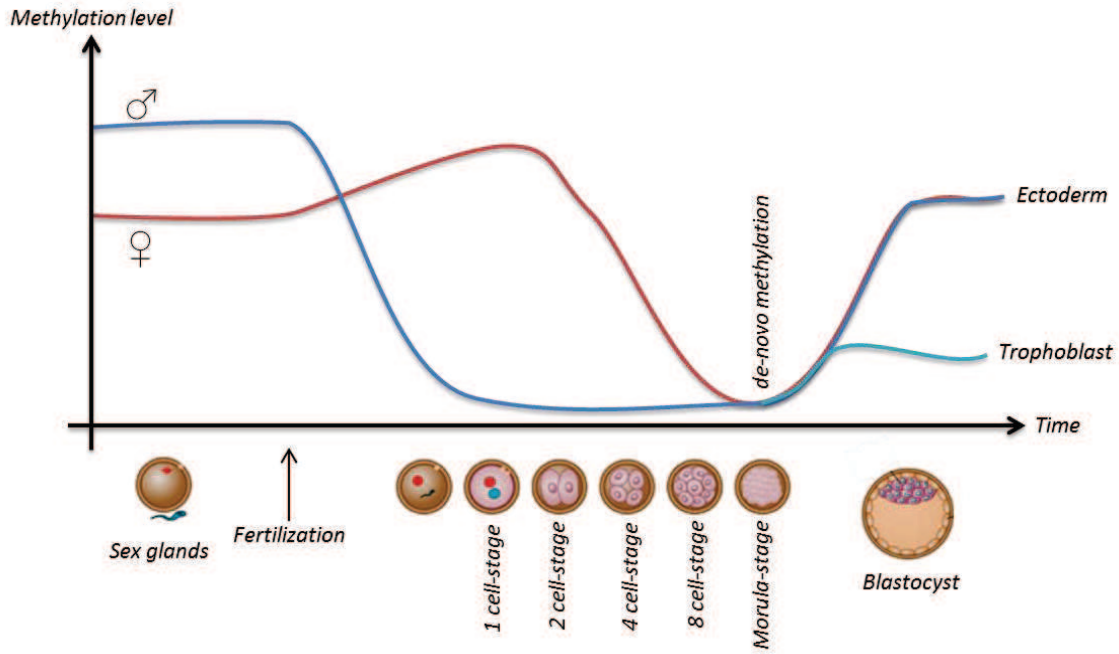


FIGURE 13

Methylation state during development

After fertilization, passive demethylation of the paternal genome (in blue), later followed by passive demethylation of the maternal genome (in red), are maintained until the morula stage. Following *de novo* methylation marks the start of cell differentiation, trophoblast becoming then placenta and ectoderm giving rise to the embryo.

Modified from Xiangzhong Yang et al. Nuclear reprogramming of cloned embryos and its implications for therapeutic cloning. Nat Genet. 2007. 39(3):295-302.

3.1.1.2 In cancer cells

One characteristic of many cancers is the global hypomethylation state and localised hypermethylation of CpG islands[95].

Global hypomethylation affects 10 to 30 % of the genome and is essentially localised on repeat elements like satellite-2[96] or alu[97]. It is associated with a recovery of the ectopic expression of

proto-oncogenes[98], retrotransposons, or imprinted genes. Furthermore, it is responsible for chromosomal instability leading to tumorigenesis[99,100].

CpG islands hypermethylation seems, for its part, to be responsible for tumour suppressor genes repression[101,102]. Thus, in hepatic cancers, the promoter of the gene CDH1, coding for E-cadherin, as well as that of HIC1, undergo a progressive hypermethylation[103]. This is linked to a loss of cell adhesion, characteristic of cancer cells[104]. In 2004, the concept of CIMP for "CpG Island Methylator Phenotype" was suggested[105], following assumptions that CpG islands hypermethylation could be cancer-specific. For example, the promoter of p16 is hypermethylated in various cancers[106], but that of p14 only is in colon cancer[107]. Also, the promoter of TIMP-3 is hypermethylated in colorectal and kidney cancers[108,109] but surprisingly, it is hypomethylated in acute myeloid leukaemia[110]. This observation could thus call into question the idea of a systematic hypermethylation of tumour suppressor genes promoters.

3.1.2 The methyltransferases

In mammals, three methyltransferases are found. These enzymes are able to catalyse the transfer of a methyl group from the SAM (S-adenosyl-L-methionine) to the position 5 of a cytosine, forming thus a 5mC[111].

The first methyltransferase to be described was DNMT1, in mouse, in 1988[112-114], and then in human in 1992[115]. Functional studies showed the high affinity of this enzyme towards hemimethylated sites. Like so, in human, DNMT1 has been shown to be 7 to 20-fold more active *in vitro* on hemimethylated sites than on unmethylated DNA[116]. For this reason, this enzyme is considered as responsible for the maintenance of methylation at hemimethylated sites, like for example, after replication by methylating the neo-synthesised strand (*figure 14*).

Gene knockout experiments carried out in mouse could highlight a global hypomethylation of the genome, characterised by a deregulation of imprinted genes, as well as an activation of the lyonised X-chromosome, leading to early embryonic death[117,118]. In contrast, cultured embryonic stem cells do not show any abnormal phenotype upon dnmt1 gene knockout, with the exception of their inability to differentiate. These observations corroborate the idea that DNA methylation would be involved in stem cell differentiation[117].

In 1998, two new enzymes were identified: DNMT3A and DNMT3B[119,120]. Unlike DNMT1, these two enzymes are able to methylate hemimethylated sites but also non-methylated DNA (*figure 14*). They are thus responsible for *de novo* methylation of the DNA, and are targeted to their sites by recruitment factors. It was indeed shown that DNMT3 enzymes were highly expressed during early stages of embryonic development, followed by a near-disappearance in differentiated cells[121,122]. Hence, these enzymes set up a personal methylation map in the embryo, after complete loss of the maternal and paternal methylomes.

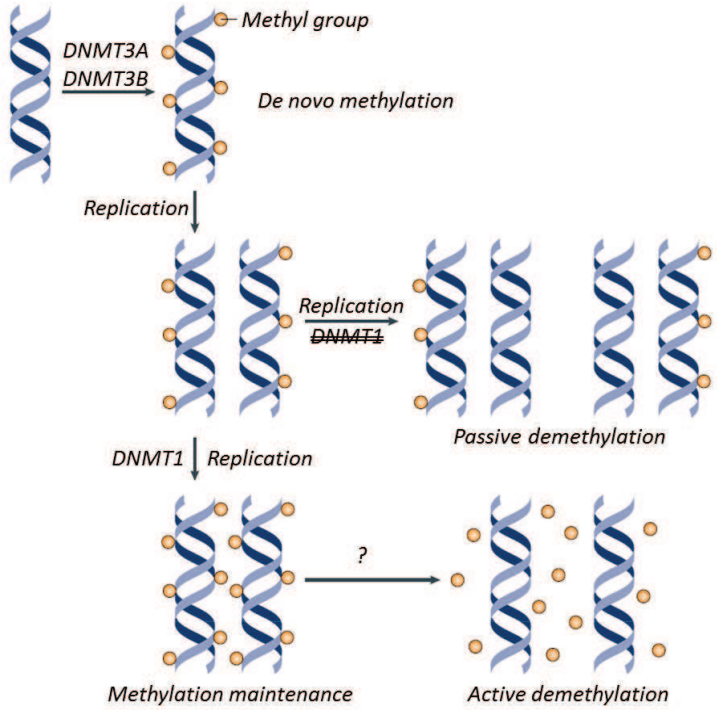


FIGURE 14

DNA methylation and demethylation mechanisms

In embryo, *de novo* methylation is ensured by DNMT3A/B, whereas maintenance of these marks is achieved by DNMT1 during replication.

If the latter is absent or inhibited, neosynthetized DNA can not be methylated: this results in passive demethylation.

Modified from Susan C. Wu & Yi Zhang. Active DNA demethylation: many roads lead to Rome. Nat Rev Mol Cell Biol. 2010, 11(9):607-20.

3.1.3 The demethylation

To date, no DNA demethylase has been discovered. The previously described mechanism, involving AID, to demethylated Oct4 and Nanog promoters, can not be considered as a global way to demethylated DNA since it involves deamination of 5mC, hence replacing it by an uracyl residue.

Demethylation can thus only occur in two different manners: either in a passive way, in the course of replications, by not methylating the neo-synthesised strand; or in an active way, in a replication-independent process. But concerning this latter point, very little is known. Only one mechanism highlighted in plants was studied to date. It involves a family of four 5mC glycosylases (ROS1, DME, DML2 and DML3), which act by base excision repair[123] (*figure 15*). In mammals though, no 5mC glycosylase was identified thus far. A similar mechanism, although indirect, was nevertheless suggested. It is based on the observation that a methylated cytosine has 10 to 50-fold more chances than any other nucleotide to spontaneously deaminate into a thymine residue, resulting then in 5mCpG/TpG mismatches[124]. Precisely, it so happens that two proteins are able to recognise and repair this kind of mismatch: TDG and MBD4. But this suggestion, although to be considered, implies a way too important random variable – spontaneous deamination of 5mC – to contemplate as an active mechanism.

For several years, 5mC was the only modified nucleotide described in mammals. But recently, hydroxymethylated cytosines (5hmC) as well as enzymes responsible for this modification – TET1 to 3 –, were discovered and gave rise to new expectations in understanding the demethylation mechanism[125,126] (*figure 16*). TET1 was initially discovered in acute myeloid leukaemia, as a partner of the KMT2A methyltransferase[127,128]. Later, studies highlighted the oxidative activity of TET1

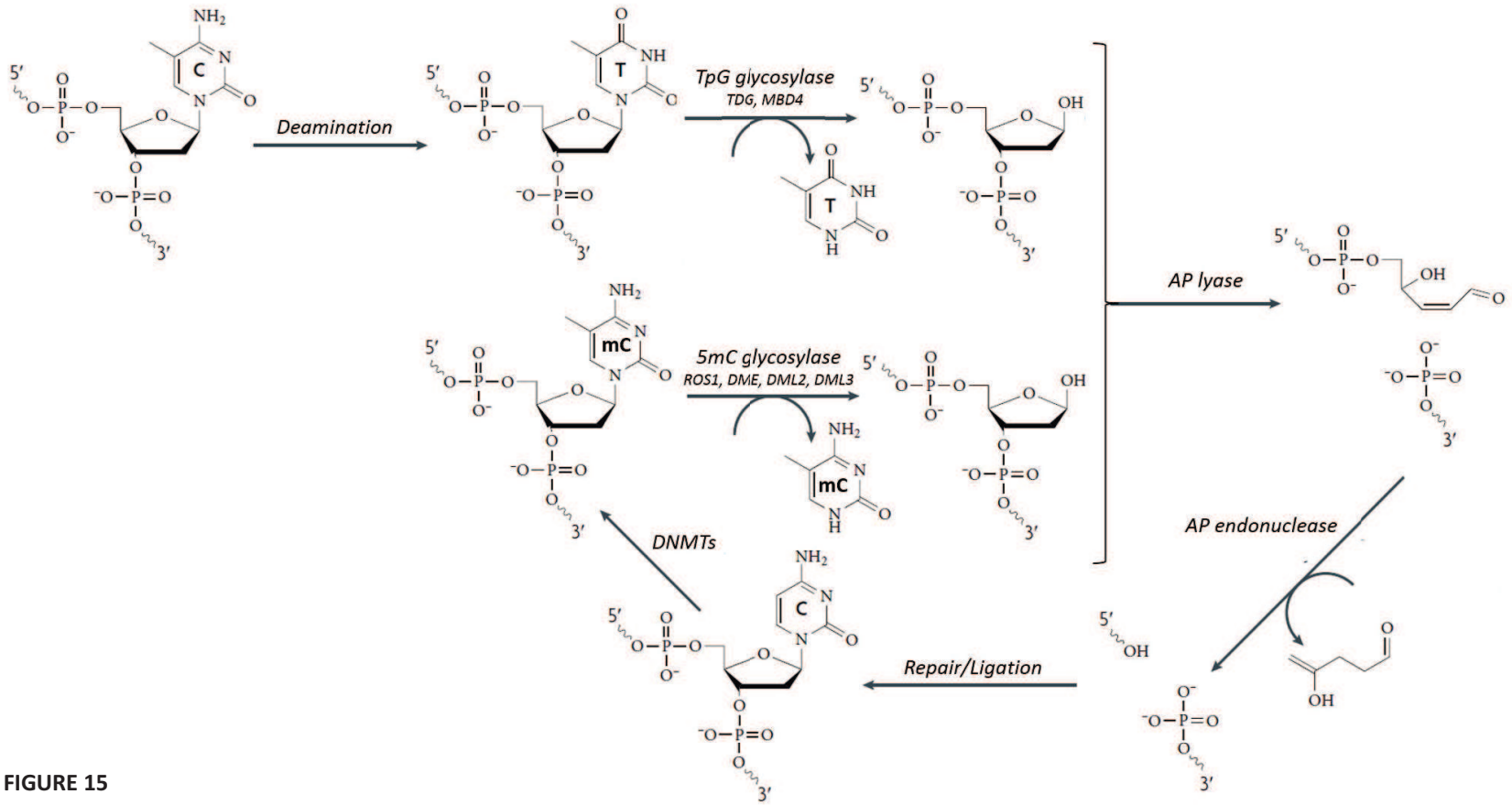


FIGURE 15

DNA demethylation

In mammals, 5mC deamination to thymine allows TDG and MBD4 mismatch-specific glycosylases to initiate a base excision/repair mechanism. In plants however, this base excision/repair process is directly possible via the glycosylase activity of the ROS1 enzymes. After cleavage of the DNA backbone on the 3' side of the abasic site by AP lyase activity, and excision by an AP endonuclease, the DNA polymerase together with a ligase can re-incorporate the missing nucleotide, by complementarity with the other DNA strand.

Modified from Susan C. Wu & Yi Zhang. Active DNA demethylation: many roads lead to Rome. Nat Rev Mol Cell Biol. 2010, 11(9):607-20.

towards 5mC, to produce 5hmC, but also higher oxidative level products like 5-formylcytosine (5fC) and 5-carboxylcytosine (5caC)[126]. The frequency of 5hmC has been described in two cell lines: Purkinje cells, in which this modification is only twice less abundant than 5mC[125]; and embryonic stem cells, in which one can estimate an occurrence of one 5hmC every 3000 nucleotides[126]. This thus clearly shows the pertinence at the physiological level of this modification.

In this way, in embryonic stem cells, TET1 has shown its ability to maintain hypomethylation on Nanog promoter. 5hmC seems thus to play an important role in the demethylation mechanism, maybe acting as a reactive intermediate. Indeed, this modification not being recognised by DNMT1[129], neo-synthesised DNA could this way be passively demethylated. In parallel, 5hmC could also be an active demethylation intermediate, leading to the exchange of 5mC into cytosine. Several pathways are proposed to lead to this result. First of all, a glycosylase activity towards 5hmC of the T2 bacteriophage was detected in calf thymus extract, back in 1988[130], although no enzyme responsible for this could be identified to date. Also, 5hmC spontaneous deimination into 5hmU allows base excision repair by SMUG1 (single-strand-selective monofunctional U DNA glycosylase 1)[131]. Plus, 5hmC could be spontaneously reduced in response to a high pH or exposure to UV-rays[132-134]. And finally, 5fC and 5caC could, for their part, undergo glycosylation followed by excision repair by TDG, mentioned earlier[135-137], but also enzymatic decarboxylation, as observed in the thymine rescue pathway, by an enzyme similar to iso-orotate decarboxylase. In this manner, a decarboxylation activity of 5caC was observed in mouse embryonic stem cells, suggesting that such an enzyme might exist[138]. And indeed, researchers suggested this year, based on their observations *in vitro*, that DNMTs could ensure the role of highly oxidized-cytosine demethylase[139].

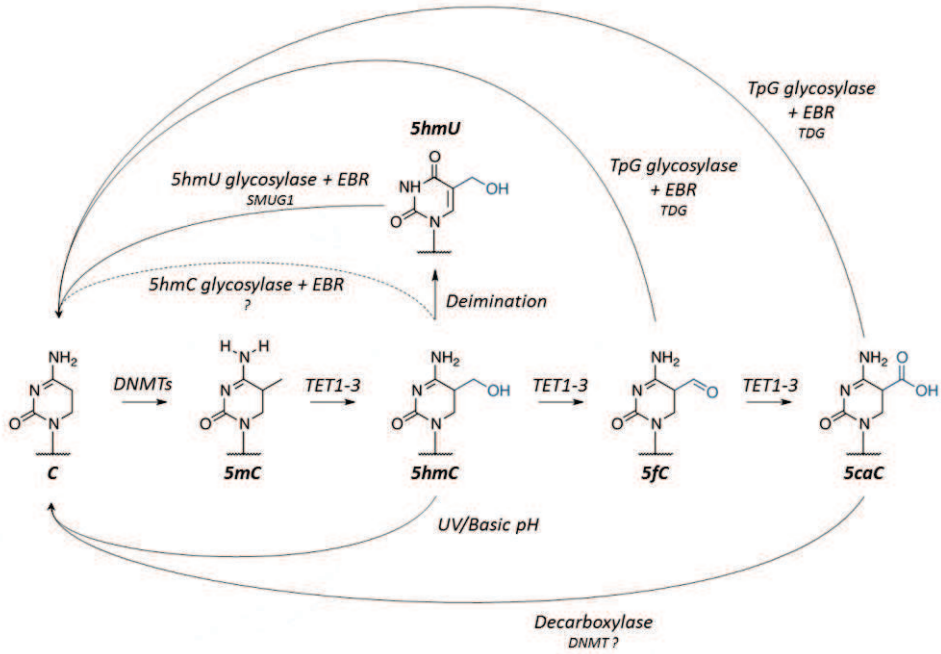


FIGURE 16

Oxidation pathways and active DNA demethylation

Modified from Susan C. Wu & Yi Zhang. Active DNA demethylation: many roads lead to Rome. Nat Rev Mol Cell Biol. 2010, 11(9):607-20.

3.2 Histone post-translational modifications

One distinctive feature of histones, and especially of their N-terminal tail, is the high number of post-translational modifications they can undergo. Some ten different types can be listed: lysine acetylation, lysine and arginine methylation, serine and threonine phosphorylation, lysine ubiquitination, lysine sumoylation, glutamic acid ADP-ribosylation, arginine deimination into citrulline, isomerisation of prolines *cis* to *trans* and vice-versa, and, discovered more recently, lysine crotonylation. The close link between these chemical modifications and transcription regulation has been highlightened since 1964 by *Allfrey et al*[140,141].

To date, over 170 modification sites were inventoried, mostly discovered by mass spectrometry and antibody-based detection methods (*figure 17*). These modifications offer the cell a vast panel of physiological responses, like transcription activation or repression, DNA repair, chromatin compaction or decompaction, etc.

3.2.1 The lysine acetylation

Acetylation was mentioned for the first time in 1964 as a regulatory mechanism for RNA synthesis, histone hyperacetylation leading to an increase of transcriptional activity, whereas hypoacetylation inhibits transcription[140,141]. But it was not until the end of the 1990s that histone acetylation was shown to be finely regulated by enzymes. In 1996, Kuo & Allis discovered histones acetyltransferases (HATs), their first representative being Gcn5p[142], while Taunton discovered, the same year, histones deacetylases (HDACs), with their first representative Rpd3p[143] ; and two years later, Kuo & Allis described the role of these enzymes, in close relation with gene expression[144].

Unlike amino-terminal acetylation undergone by plenty of proteins during their maturation in the Golgi apparatus, lysine ε-amine acetylation in histones runs in a post-translational manner, and is fully reversible. Addition of an acetyl group, brought by coenzyme A, on lysines, neutralize their positive charge and increases their hydrophobicity (*figure 18*). This modification is thus inextricably associated with a loss of interactions between the histone octamer, initially basic, and the DNA, acidic, resulting in a local decompaction of the chromatin and transcription activation. Conversely, histone deacetylation restores their positive charge and ensures thus the repression of gene expression, genes becoming inaccessible to the transcription machinery with the recompaction of the chromatin. It is therefore a very finely controlled regulation mechanism.

In yeast, there are at least nine complexes carrying an acetyltransferase or deacetylase activity known to date. These complexes generally consist in a protein with catalytic activity and auxiliary proteins which potentialise, regulate and/or guide the activity towards precise genomic locations.

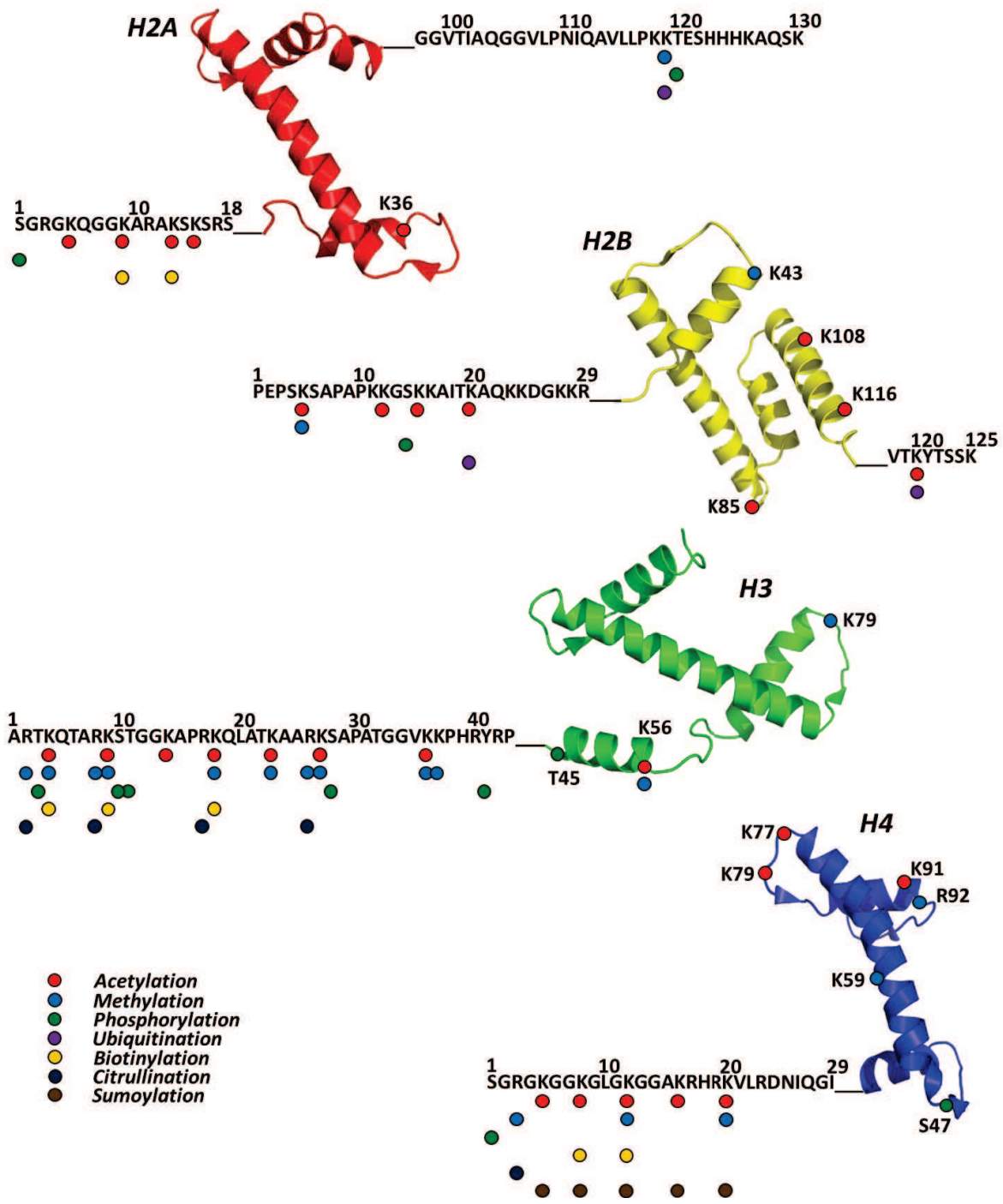


FIGURE 17

Post-translational histone modifications

This figure gives a non-extensive overview of several post-translational modifications that are observed on histones. One can note that the majority of these modifications occurs on N-terminal tails. The main acetylation sites (red), methylation sites (light blue), phosphorylation sites (green), ubiquitination sites (purple), biotinylation site (yellow), citrullination sites (dark blue) and sumoylation sites (brown) are shown.

(Uniprot - Histone hH2A.1A: Q96QV6; Histone hH2B.1B: P33778; Histone hH3.1: P68431; Histone hH4: P62805; pdb: 1aoi)

a)

b)

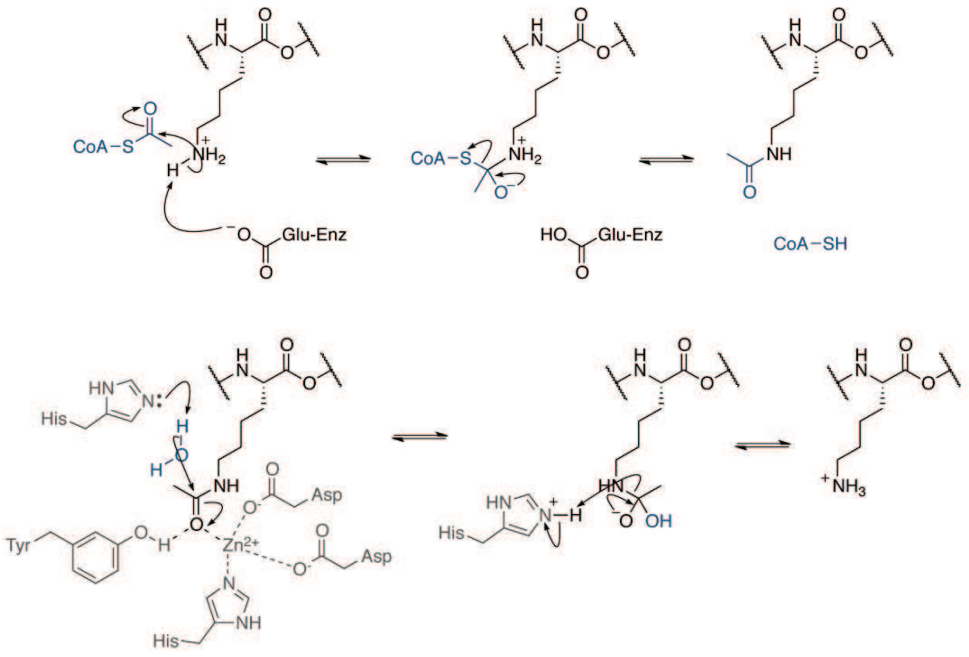


FIGURE 18

Histone acetylation and deacetylation mechanism

a) Histone acetylation

b) Histone deacetylation

Modified from Tom Brown & Tom Brown Jr. Nucleic Acids Book

3.2.1.1 Histones acetyltransferases (HATs)

So far, 19 histone acetyltransferases were identified in mammals, classified into two types: type A, which is found in the cell nucleus, and acetylates nucleosomal histones; whereas type B is mostly found in the cytoplasm and acetylates neo-synthesised histones. However, this classification was made obsolete by the discovery of HATs that can be both nuclear and cytoplasmic.

Histone acetyltransferases can also be divided into several families described hereafter (*table 6*). These families were established on sequence analysis, linking some HATs upon their great sequence similarity while others shared only little if any similarity. Furthermore, each of these families showed preferences for different substrates, their role being thus indicated in precise functional contexts.

TABLE 6
Histones acetyltransferases

HATs			Associated complex	Substrat	Role/Function
GNATs	Gcn5	KAT2A	SAGA, SLIK (SALSA), ADA, HAT-A2, ATAC, TFTC	H3K9, H3K14, H3K56, H4K5, H4K8, H4K12, H4K16, H4K91	Transcriptional coactivator
	PCAF	KAT2B	PCAF	H3K9, H3K14	
	Hat1	KAT1	HAT-B, NuB4, HAT-A3	H2AK5, H4K5, H4K12	Histones deposition
	Elp3	KAT9		H3K9, H3K18	Transcription elongation
	Hpa2	KAT10	HAT-B		Unknown
	Hpa3				
	Nut1				
MYST	MOZ	KAT6A	MSL	H3K9, H3K14	Malignant diseases
	Ybf2/Sas3				Repression
	Sas2			H4K16	
	Tip60	KAT5		H2AK5, H4K5, H4K8, H4K12, H4K16	Interacts with HIV-TAT
	Esa1		NuA4, piccolo NuA4		Cell cycle regulation
	MOF	KAT8	MSL	H4K16	Dosage compensation
	MORF	KAT6B	MSL		
	HBO1	KAT7	ORC	H3K14, H4K5, H4K8, H4K12,	Interacts with replication origin recognition complex
CBP/p300	CBP	KAT3A		H3K14, H3K18, H3K27, H3K56, H4K5, H4K8, H4K12, H4K16	Transcriptional coactivator
	p300	KAT3B			
GTFs	TFIIIC	KAT12	TFIIIC	H3K14	ARN-Pol III-dependent transcription
SRC	SRC-1	KAT13A	ACTR/SRC-1	H3K14	Transcriptional coactivator
	SRC-2	KAT13C			
	SRC-3	KAT13B	ACTR/SRC-1		
	CLOCK	KAT13D			
Others	Rtt109	KAT11			
	AFT-2				Transcriptional coactivator

GNATs (Gcn5-related N-acetyltransferases) constitute the first family, comprising Gcn5p, as well as all other related HATs[142,144,145]. One can mention PCAF, the cytoplasmic acetyltransferase Hat1 or the elongation factor Elp3. These HATs interact with a panel of transcription activators, and acetylate mainly the lysine 14 of histone H3, and to a lesser extent, lysines 8 and 16 of histone H4. They possess a HAT domain of around 160 residues, as well as a bromodomain in C-terminal position. Gcn5p is part of several multiprotein complexes, like Ada, SAGA or HAT-A2; PCAF is also part of a protein complex, similar to SAGA.

The MYST family was named after its four original representatives: **M**OZ, **Y**bf2/Sas3, **S**as2 and **T**ip60. These HATs show diverse biological functions, including their involvement in gene repression (Sas2 and Sas3)[146], in cell cycle regulation[147,148], in dose compensation (MOF)[149] or in leukemogenesis (MOZ and TIF2)[150,151]. Apart from Sas3, these HATs show a preference for histone H4 as a substrate. They have a HAT domain of around 250 residues, as well as a chromodomain in N-terminal position, and a zinc finger. Esa1, another member of this family, is found in the Piccolo-NuA4 complex, MOF is found in the MSL complex and Sas3, in the NuA3 complex.

CBP/p300 are global transcriptional regulators. They possess, in addition to a 500-residue HAT domain, a bromodomain, a KIX domain for CREB-binding and three cysteine-histidine-rich domains (TAZ, PHD and ZZ), probably regulating protein-protein interactions[152].

The family of general transcription factors (GTFs) includes the subunits of TFIIIC carrying a TATA-box binding domain[153], like TFIIIC220, TFIIIC110 or TFIIIC90. These proteins also contain two kinase domains, one at each extremity, as well as a double bromodomain in C-terminal position[154]. They show a particular affinity for histone H4 and are probably responsible for nucleosome eviction in the context of RNA Pol II and III-dependent transcription.

AFT-2 is a protein that resembles no other HAT known to date. It is the unique transcription activator factor that binds to DNA in a sequence-specific manner, and that carries an acetyltransferase activity[155]. Its DNA binding domain, a leucine-zipper, is located at its C-terminal end[156].

Some nuclear receptors, responding to steroid hormones, also possess a histone acetyltransferase activity. This family, named NCOA, includes SRC-1 (or NCOA1), SRC-2 (or NCOA2/TIF2/GRIP1) and SRC-3 (or ACTR/AIB-1/pCIP/RAC3/TRAM-1). These proteins share a basic helix-loop-helix motif, which increases the transcriptional efficiency.

Despite their sequence divergences, all these proteins share a great structural homology within their HAT domain[157] (*figure 19*). The latter is composed of three antiparallel β-strands, followed by an α-helix extending parallely, above the plane of the sheet. This motif plays a particular role in the binding of the cofactor acetyl coenzyme A, required of the acetyltransferase activity. It is largely stabilised inside the cavity formed by the sheet and the helix, each functional group of its pantethine-pyrophosphate arm being in close contact with the enzyme, through hydrogen bonds or van der Waals interactions. The adenosine cycle however is not stabilised inside the cavity and can thus adopt different orientations. Some studies lead to believe that this adenosine cycle could be implicated in a dimerisation phenomenon, stabilised by base-pair interactions[158].

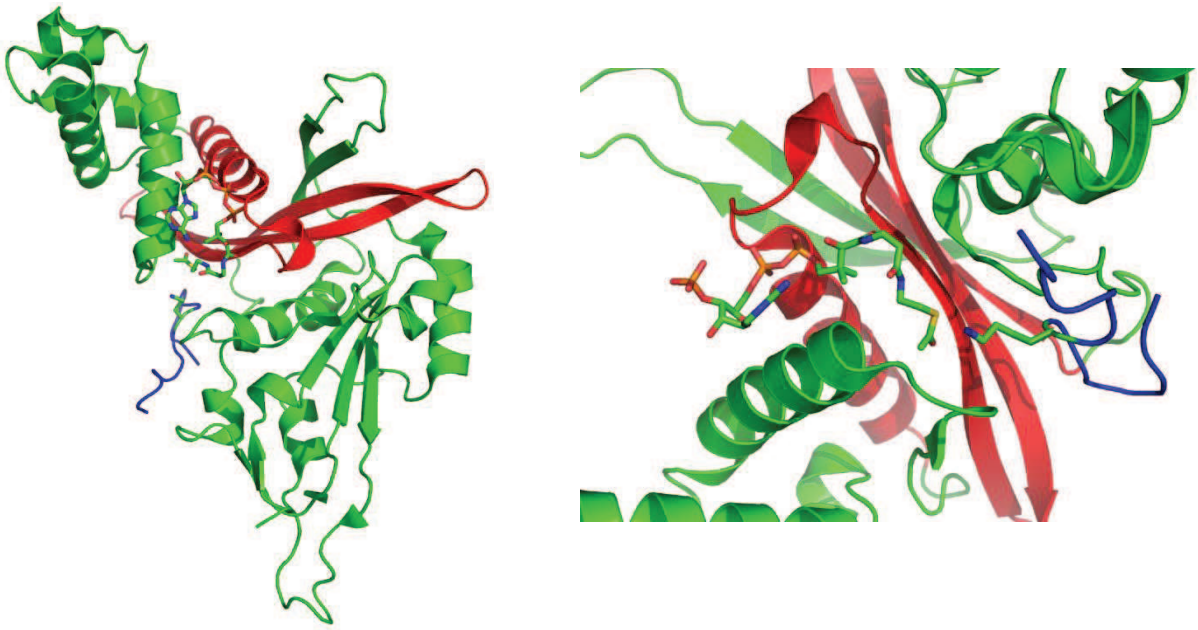


FIGURE 19

Structure of the HAT domain of histones acetyltransferases

The Hat1 enzyme is shown here. In red, its HAT domain (residues 198 to 267) composed of three β-strands and one α-helix. In the center of this domain, coenzyme A is ready to give its acetyl moiety up to the H4 lysine (in blue). *(pdb : 2pow)*

3.2.1.2 Histones deacetylases (HDACs)

In 1990, Itazaki tried to highlight new molecules able to counter the tumorigenic effect of the *v-sis* gene on 3T3 fibroblasts. He then identified a new molecule, trapoxin, able to revert the oncogenic action, but without knowing its target[159]. Three years later, Kijima showed that trapoxin-treated cells were hyperacetylated, and that their histone deacetylation function was inhibited[160]. Finally, in 1996, Taunton used trapoxin as a bet to purify its target by affinity chromatography. Mass spectrometry studies revealed that this target was an homolog of Rpd3p, a yeast transcription regulator[143].

Since then, 18 histone deacetylases were identified and classified into three groups: HDACs I, II and III (*table 7*). The first class, including HDACs 1-3 and HDAC 8, is based upon a very high homology with the yeast HDAC Rpd3p. They possess between 400 and 500 residues, and are exclusively located in the nucleus. They are moreover ubiquitous in the whole organism. These four enzymes are sensitive to trichostatin A as well as other common HDAC inhibitors (HDIs), but it has been shown that HDAC8 gene could be upregulated in presence of an inhibitor, to counteract its effect[161]. These class I HDACs are part of large protein complexes, allowing them to play a gene-specific transcription regulation role. Notably, we can mention the SMRT, CoREST, Sin3 and NuRD complexes, among others.

TABLE 7
Histone deacetylases

	HDACs	Role/Function	Location	Involvement in cancer
Class I	HDAC 1	Proliferation Apoptosis	Nucleus	Stomach Breast Colon Prostate
	HDAC 2			Stomach Prostate Colon-Rectum Cervix
	HDAC 3			Breast Colon
	HDAC 8		Nucleus Cytoplasm	
Class II	HDAC 4	Differentiation	Nucleus Cytoplasm	
	HDAC 5			
	HDAC 6	Protein degradation Tubulin regulation	Cytoplasm	Breast Colon
	HDAC 7	Differentiation	Nucleus Cytoplasm Mitochondria	
	HDAC 9		Nucleus Cytoplasm	
	HDAC 10	Protein degradation	Cytoplasm	
	HDAC 11	Tubulin regulation	Nucleus Cytoplasm	
Class III	SIRT 1		Nucleus Cytoplasm	Colon Prostate Lung
	SIRT 2	Tubulin regulation Protein degradation	Cytoplasm	
	SIRT 3	Stress resistance Metabolism	Nucleus Mitochondria	
	SIRT 4		Mitochondria	
	SIRT 5			
	SIRT 6	Genomic stability	Nucleus	
	SIRT 7		Nucleolus	

Once the yeast protein Hda1 was characterised, several teams were able to isolate human homologs based on databases analysis[162-165]. Thus, the second class of HDACs includes HDACs 4-7 and HDACs 9-11. These proteins are twice bigger than those of class I, with around 1000 residues. Most of them exhibit their catalytic domain at the C-terminal end. HDAC6 has two catalytic domains, at each end. Like class I HDACs, those of class II are also sensitive to trichostatin A. However, they are cytosolic proteins, and can be recruited to the nucleus if necessary. To finish, they are not ubiquitous, but largely expressed in the brain, heart and skeletal muscle.

Last, the third class of HDACs, also called sirtuins, includes SIRT 1 to 7, which are homologs of the yeast deacetylase ySir2. They possess between 300 and 400 residues, with the exception of SIRT-1 which possesses 747 amino acids. Unlike the two first classes of HDACs, class III isn't sensitive to trichostatin A, and its catalytic activity depends on the cofactor NAD^+. These enzymes still remain little known compared to the other HDACs, but it has been shown that SIRT-1 could associate with and deacetylate p53, repressing thus transcription of p53-regulated genes and preventing DNA-damage induced apoptosis[166,167].

3.2.2 Arginine and lysine methylation

Histone methylation is a way more complex phenomenon than acetylation, in the sense that it can be associated either with activation or repression of gene expression, according to the residue that is methylated and its level of methylation (monomethylated, dimethylated or trimethylated). Thus, for example, di- and trimethylation of lysine 9 of histone H3 (H3K9me2/3) as well as trimethylation H3K27me3, within a gene promoter, repress the transcriptional activity of the said gene. By contrast, H3K4 trimethylation within the transcription start site as well as H3K79me1/2/3 and H3K36me3 modifications within the coding region of a gene enable transcription activation. This combination of modifications forms therefore the necessary epigenetic information for transcription factors to establish their own role.

3.2.2.1 Arginines methyltransferases (PRMTs)

The methylation of histone arginine residues is achieved by enzymes called arginine methyltransferases. These enzymes are not specific for histones and showed their activity on other proteins like nucleolin, fibrillarin or certain helicases[168,169]. Targeted arginines can be monomethylated or dimethylated on their η-amine group, by transfer of a methyl moiety from the SAM. Furthermore, dimethylation can be symmetrical (one methyl moiety on each η-amine group) or asymmetrical (two methyl moieties on the same η-amine group). Arginine trimethylation, through methylation of the ε-amine of an already dimethylated arginine, has, to date, been observed only in yeast.

Two classes of PRMT were described. Type 1 PRMTs generate monomethylated arginines (Rme1) and asymmetrically dimethylated arginines (Rme2as) (PRMT1, PRMT3-4, PRMT6, PRMT8); while type 2 PRMTs can generate monomethylated arginines and symmetrically dimethylated

TABLE 8
Histones methyltransferases

	HMTs	Role/Function	Known substrates	Involvement in cancer
Type-I PRMT	PRMT1	Transcription activation Transduction RNA splicing DNA repair	Histone H4 Fibrillarin Nucleolin SAM68 FGF2 STAT1/3/6 EWS…	Breast Prostate Lung Colon Bladder Leukaemia
	PRMT3	Ribosome homeostasis	rpS2 Fibrillarin SAM68 STAT1…	Breast
	PRMT4 CARM1	Transcription activation RNA splicing DNA repair Cell cycle progression	Histone H3 PABPC1 CBP/p300 RAC3/p/CIP…	Breast Prostate Colon-Rectum
	PRMT6	Transcription regulation	Histone H4 Histone H2A Fibrillarin VIH-Tat/Rev/NC RNA Pol β MBP…	Lung Bladder
	PRMT8	Brain-specific functions	SmD1/3 hnRNP A1 PABPC1…	Ovaries Skin Large intestine
Type-II PRMT	PRMT2	Transcription regulation	/	Breast
	PRMT5	Transcription repression Transduction piRNA pathway	Histone H3 Histone H4 Histone H2A Fibrillarin MBP SmD1/3 MBD2…	Stomach Colon-Rectum Lung Lymphoma Leukaemia
	PRMT9	/	Histone H4 Histone H2A MBP…	/
	PRMT10	/	/	/
	PRMT11	/	/	/
Type-III PRMT	PRMT7	Genomic imprinting (in male germ lines)	Histone H4 Fibrillarin MBP SmD1/3…	Breast

arginines (Rme2s) (PRMT5, PRMT7, PRMT9)[170,171]. In total, 11 PRMTs were described in human to date, including two putative enzymes (PRMT10 and PRMT11) (*table 8*).

PRMT1 was the first arginine methyltransferase to be discovered in mammals in 1996, on account of its marked predominance compared with all other PRMTs described hereafter[172-174]. This predominance is most notably exemplified by an affinity for over 40 different substrates[169,175]. Besides histone H4 arginine methylation, PRMT1 is also involved in methylation of the elongation factor SPT5, regulating thus its interaction with RNA Pol II, if to cite only one example among dozens[176,177].

PRMT2 was identified in 1997, by sequence homology with PRMT1[174] and its yeast homolog, yRMT2[178]. Two-hybrid screenings showed an affinity of PRMT2 for some nuclear receptors, like ERα[179] or AR[180]. However, no enzymatic activity was revealed to date, despite the great sequence homology with PRMT1, which leads to believe that these two enzymes share a similar activity[174].

PRMT3 was discovered as a protein partner of PRMT1[181]. The substrate specificity of this enzyme is conferred by its N-terminal zinc finger. PRMT3 is indeed involved in RNA-associated protein methylation, like ribosomal proteins[182].

PRMT4 was identified as an arginine methyltransferase interacting with GRIP1, a coactivator. This gave it its nickname “CARM1” (Coactivator associated arginine methyltransferase 1)[183]. GRIP1 is a member of the p160 family, which includes SRC-1 (or NCOA1), SRC-2 (or NCOA2/TIF2/GRIP1) and SRC-3 (or ACTR/AIB-1/pCIP/RAC3/TRAM-1). These proteins are able to directly bind a panel of nuclear receptors, playing then a primary coactivator role; while they recruit other proteins acting as secondary transcriptional mediators[184]. CARM1 is one of those mediators, which role is to amplify the transactivation of nuclear receptors through methylation of H3 N-terminal arginines[183-185].

PRMT5 plays an important role in transcription control and modulation. Indeed, it is able to methylate major regulatory proteins like E1 cyclin[186] or IL-2[187]. Furthermore, PRMT5 interacts with COPR5 (Cooperator of PRMT5), allowing its specific recruitment, to methylate histones and some transcription elongation factors[177,188,189].

PRMT6 was the first arginine methyltransferase to be identified through structural homology of the catalytic motif[190]. Besides histone H3 arginine methylation[191], PRMT6 is also able to methylate some proteins of HIV-1, decreasing then gene expression and thus, viral replication[192-194].

PRMT7 was discovered by sequence homology and motif comparison[195,196]. It is the only PRMT to possess two methyltransferase domains, when all the other PRMTs only have one[195]. Recently, PRMT7 was downgraded from the type 2 PRMTs, to integrate a new type of methyltransferases: type 3 PRMTs. Indeed, Clarke’s group showed in 2012 that PRMT7 was not able to produce symmetrically dimethylated arginines but only monomethylated arginines[197].

PRMT8 was identified by sequence homology, since it shares over 80 % of identity with PRMT1, making it the closest homolog[198,199]. Whereas all the other PRMTs are expressed in a more or less ubiquitous manner in all cell types – with some specificities somehow –, PRMT8 is largely expressed in brain[198]. This enzyme distinguishes itself in membrane protein methylation, to which it is recruited through its myristoylation domain[198,199]. However, its great homology with PRMT1 confers them some common substrates, like histones or RNA-binding proteins[200,201].

PRMT9 was found by motif homology, although it shares only a small sequence homology with the other PRMTs[202]. This methyltransferase is able to methylate histones, but also MBP (Maltose-binding protein).

PRMT10 was predicted by sequence homology with PRMT7, but no biochemical or enzymological studies were carried out to date[203-205].

PRMT11 was also predicted by sequence homology, this time with PRMT9, and again, no biochemical or enzymological studies were carried out to date[203,204].

3.2.2.2 Lysines methyltransferases (HKMTs)

In 2000, SUV(39)H1 was described as the first lysine methyltransferase, targeting lysine 9 of histone H3[206]. Since then, numerous HKMTs were identified, all of them using SAM as a methyl donor. It is interesting to mention that of all HKMTs studied to date, only one, Dot1, is able to methylate a lysine located within the histone core (H3K79)[207], the others methylating only N-terminal lysines. Furthermore, Dot1 is deprived of a structural domain found in the other HKMTs, called SET domain, and responsible for the enzymatic activity. This domain contains 130 residues and is highly conserved.

Among the SET-containing HKMTs, four subgroups can be distinguished: SET1, SET2, SUV39 and RIZ.

3.2.2.3 Demethylases (HDMs)

In 2004, Yang Shi discovered the existence of proteins carrying a lysine demethylase activity[208]. They were classified into two families, according to their catalytic mechanism[209,210]. The first family, LSD1, includes two representatives: LSD1/KDM1A and LSD2/KDM1B[208,211]. These enzymes are flavin monoamine oxidases, able to demethylate mono- and dimethylated lysine residues (*figure 20*). LSD1 exhibits a double specificity for H3K9me1/2 and H3K4me1/2, depending on its associated factor: in the presence of the androgen receptor, LSD1 acts as a transcriptional activator by demethylating H3K9[212,213]; whereas when associated to CoREST, it represses transcription by demethylating H3K4[214-216]. Recent work highlights coordination of these actions with H3 tail phosphorylation state[217].

Unlike LSD1, LSD2 has a zinc finger and acts rather as an RNA Pol II-dependent elongation effector, by demethylation of H3K36 within the targeted gene[218].

The second family includes all the other histone demethylases known to date. Their common feature is a JMJC domain (Jumonji C-terminal domain) responsible for demethylase activity. These enzymes are Fe^{2+} and α-oxoglutarate-dependent, unlike LSD1 and 2[210]. They are divided into several subgroups. In the first one, KDM2A and KDM2B have proven their high specificity for H3K36me2, and mainly act in the nucleolus, confirming thus a very precise physiological role. They indeed repress RNA Pol I and II-transcription[219-222].

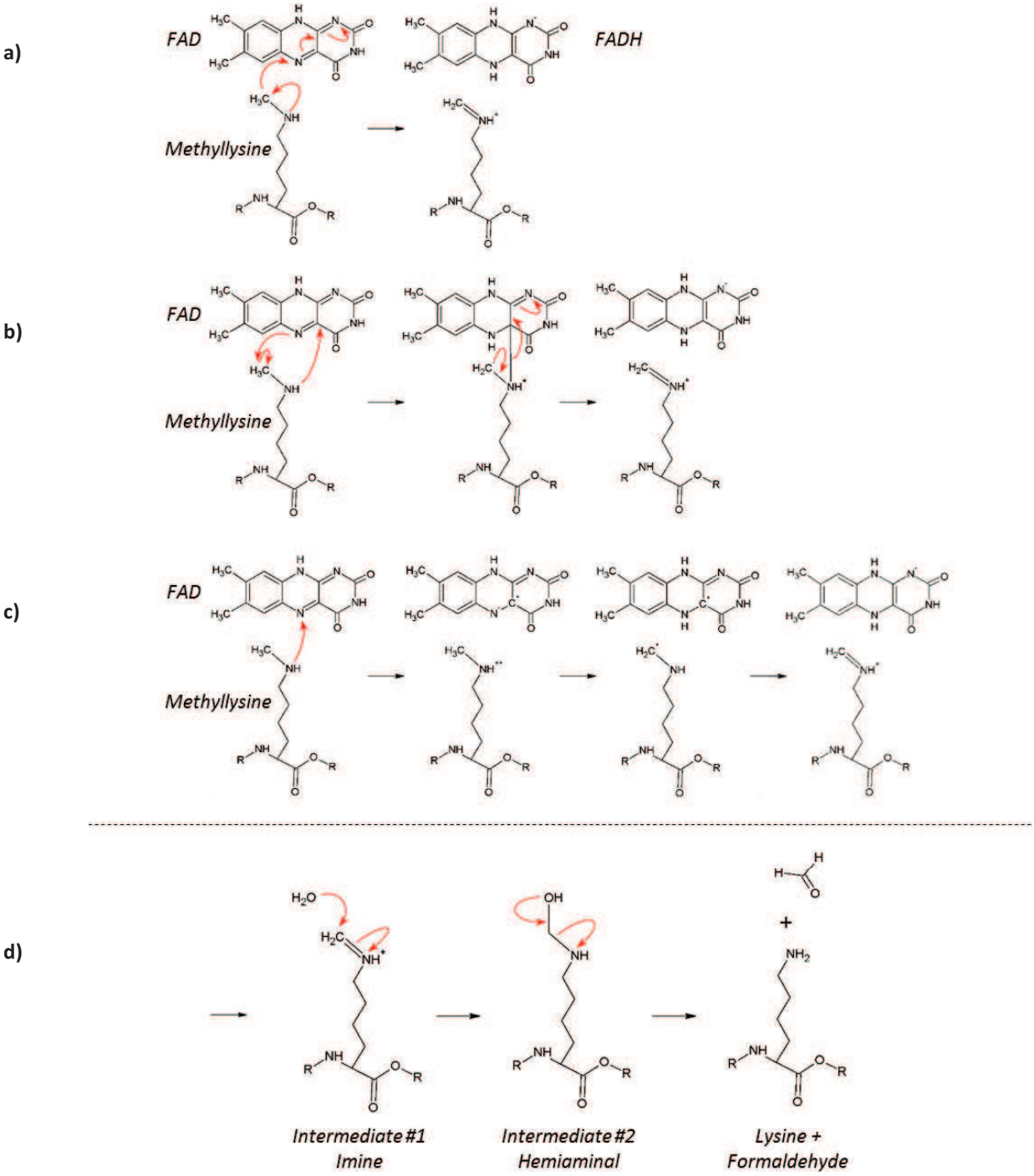


FIGURE 20

Demethylation mechanism by LSD1/KDM1A

LSD1 demethylates lysines via an oxidative FAD-dependent process. The molecular mechanism leading to the first reaction intermediate, an imine, is subject to discussion and three suggestions have been made:

a) Hybrid transfer mechanism
b) Nucleophilic substitution mechanism
c) Radical mechanism
d) Final common steps

KDM3A, KDM3B and KDM3C form a second subgroup in this second family of demethylases. To date, only KDM3A has shown a specific lysine demethylase activity, directed towards H3K9me1/2. It is involved in a variety of processes, like transcriptional activation of metabolic genes, of genes regulated by the androgen receptor, or genes involved in spermatogenesis[223,224]. It also controls reprogramming of stem cells into neurones[225].

KDM4A, KDM4B, KDM4C and KDM4D were the first enzymes to be described as being able of trimethylated lysine demethylation. They exhibit a double specificity for H3K9 and H3K36[226-228]. They are also able to demethylate lysine 26 of isotype 4 of histone H1, as well as other non-histone proteins like G9a, a SET-containing lysine methyltransferase[229,230]. To date, KDM4C has only shown a role in gene transcription activation[213,231]. KDM4A on the contrary proved to be a corepressor[232,233], by interacting with histone deacetylases and proteins of the Rb family (Retinoblastoma protein) to repress genes of the E2F group[233].

KDM5A, KDM5B, KDM5C and KDM5D are specific to H3K4me2/3[234-236]. These enzymes, through their ARID domain (AT-rich interaction domain), are able to bind DNA in a sequence-specific manner[237]. KDM5A participates in E2F genes repression in differentiated cells, in cooperation with Sin3[238]. KDM5B is able to block cell differentiation[239] and is involved in neural ridge development in embryo[240]. KDM5C is a repressor associated within the REST complex (repressor element 1-silencing transcription factor) with HDAC1, HDAC2 and G9a, inhibiting neuronal gene expression in non-neuronal tissues[241]. KDM5D finally is a repressor of developmental genes in embryonic kidney cells[242], and also a regulator of chromatin compaction during spermatogenesis[243].

The next subgroup includes KDM6A, KDM6B and UTY. The latter hasn't shown any demethylase activity to date. The first two on the other hand are specific to H3K27me3[244-247]. They are transcriptional coactivators involved in development, differentiation[246-248], inflammatory response[248,249] or cell cycle[250]. It is interesting to mention that these demethylases are also involved in bivalent promoters gene expression[251,252]. These promoters exhibits at the same time an activation mark (H3K4me) as well as a repressive mark (H3K27me). RNA Pol II is already recruited to this promoter, and is waiting for H3K27me demethylation to start transcription. This process allows thus an efficient and fast gene transcription at a given moment, i.e., at certain stages of the development.

KDM7A, KDM7B and KDM7C were recently described doubly-specific to H3K9me1/2 and H3K27me1/2[253,254]. KDM7B was, in addition, described as the very first demethylase able to demethylate residue H4K20me1[255,256]. KDM7A plays an important role in brain development[253,257] ; KDM7B interacts with RNA Pol I and II, in the nucleolus, where it activates ribosomal DNA transcription[258-260].

To finish, KDM8 recently unveiled its demethylase activity towards H3K36me2. It activates the A1 cyclin gene[261], participating thus in the cell cycle progression.

Arginine demethylases are still largely unknown, and with good reason, since only one was discovered to date. Named JMJD6, it belongs to the second family of lysine demethylases[262]. In addition to demethylating lysine residues, it is also able to demethylate H3R2 and H4R3. Another

way to demethylate arginines is through deimination, ensured by PAD 1 to 4 (Peptidyl arginine deiminase), but no reverse-enzyme, to convert citrulline back into arginine, is known to date.

3.2.3 Other post-translational modifications (*figure 21*)

Like acetylation, histone phosphorylation is highly dynamic. It occurs on serine, threonine and tyrosine residues. Phosphorylation levels are regulated by kinases on one hand, which phosphorylate the residues; and by phosphatases on the other hand, which dephosphorylate them[263]. ATP is used as a γ-phosphate donor, a negatively charged moiety which induces thus a change of charge on histones, probably leading to a direct modification of the chromatin structure.

Little is known about enzymes responsible for this modification. Aurora B kinase has long been characterised as responsible for overall-genome phosphorylation of residues H3S10 and H3S18 during mitosis[264,265]. MAPK1 is a kinase which possesses a DNA-binding domain, via which it can be recruited, but site specificity probably requires other associated factors. JAK2, another kinase, catalyses the phosphorylation of a residue inside the histone core: H3Y41[266].

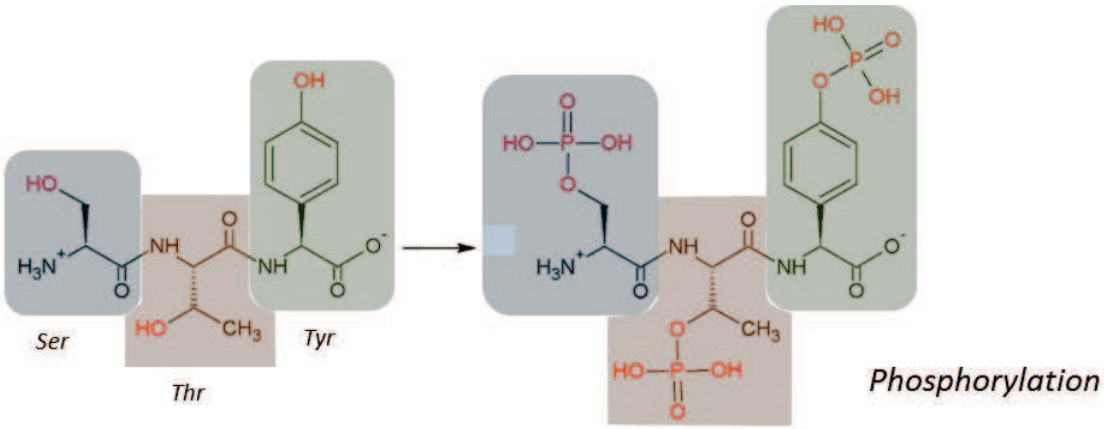
Even less is known about phosphatases, except that they have a strong activity inside the nucleus, as illustrated by their rapid action. For example, PP1 phosphatase works in an antagonist-manner to Aurora B.

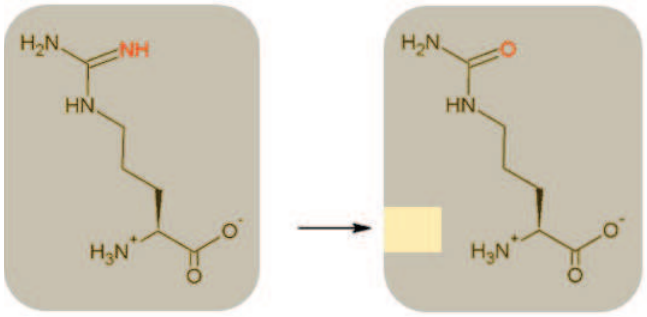
Deimination is the conversion of an arginine residue into a citrulline. PAD 1 to 4 proteins (Peptidyl arginine deiminase) ensure this function on histones H3 and H4. This reaction antagonises the activator effect of arginine monomethylation[267]. The reverse reaction has not been described yet, although its existence is assumed, since the promoter of the oestrogen-dependent pS2 gene exhibits cyclically either citrullines or arginines[268].

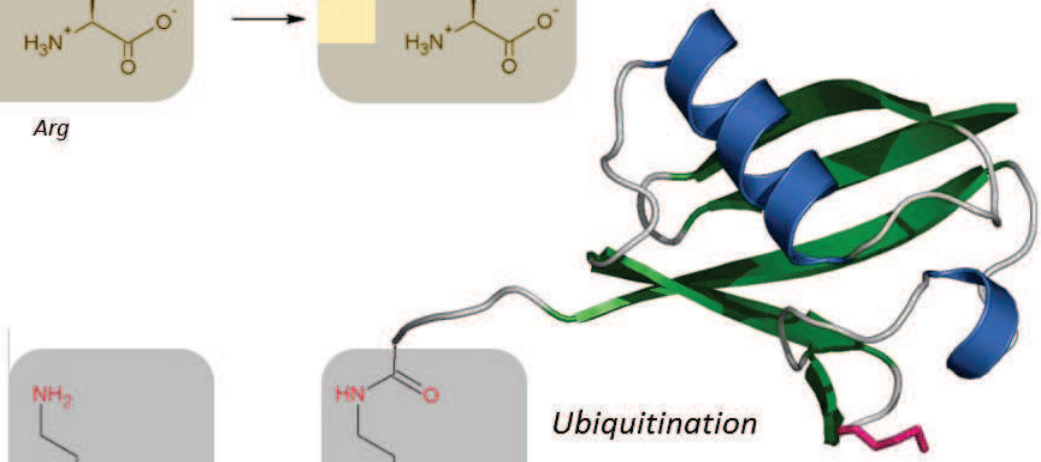
Lysine ubiquitination has been discovered in 1975[269], but it was not until 30 years later that this modification started to be studied[270]. It results in the covalent binding of an ubiquitin protein to the lysine residues of histones H2A and H2B. This ubiquitin protein is a large peptide of 76 amino acids, for a mass of 8.5 kilo Daltons. In humans, the Polycomb complex, and in particular the protein Ring1B[271,272], was the first protein identified as being able to catalyse H2AK119 ubiquitination[273]. This epigenetic mark is associated with a repressed transcriptional activity of polycomb genes. Later, other proteins with E3-ubiquitine-ligase were highlighted like Ring1A, BMI1[272,274] or BRCA1[275,276].

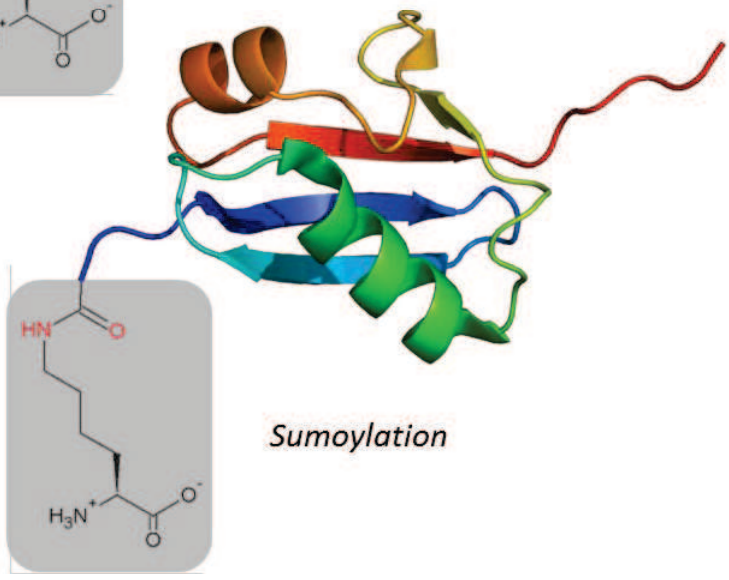
On the contrary, H2BK120 ubiquitination[277] is catalysed by RNF20/RNF40 and has an activator effect[278,279]. The role of this modification remains still little-known but it could have a physical implication in chromatin decompaction by virtue of its size, by blocking the open state of the DNA.

Sumoylation, just like ubiquitination, consists in adding a large peptide onto lysine residues of the four histones. The SUMO protein has a mass of 12 kilo Daltons, with a hundreds of amino acids, and antagonise both lysine acetylation and ubiquitination. This modification has thus a repressive role in transcription[280,281].


Ser
Thr
Tyr
Phosphorylation


Arg
Deimination


Lys
Ubiquitination


Sumoylation

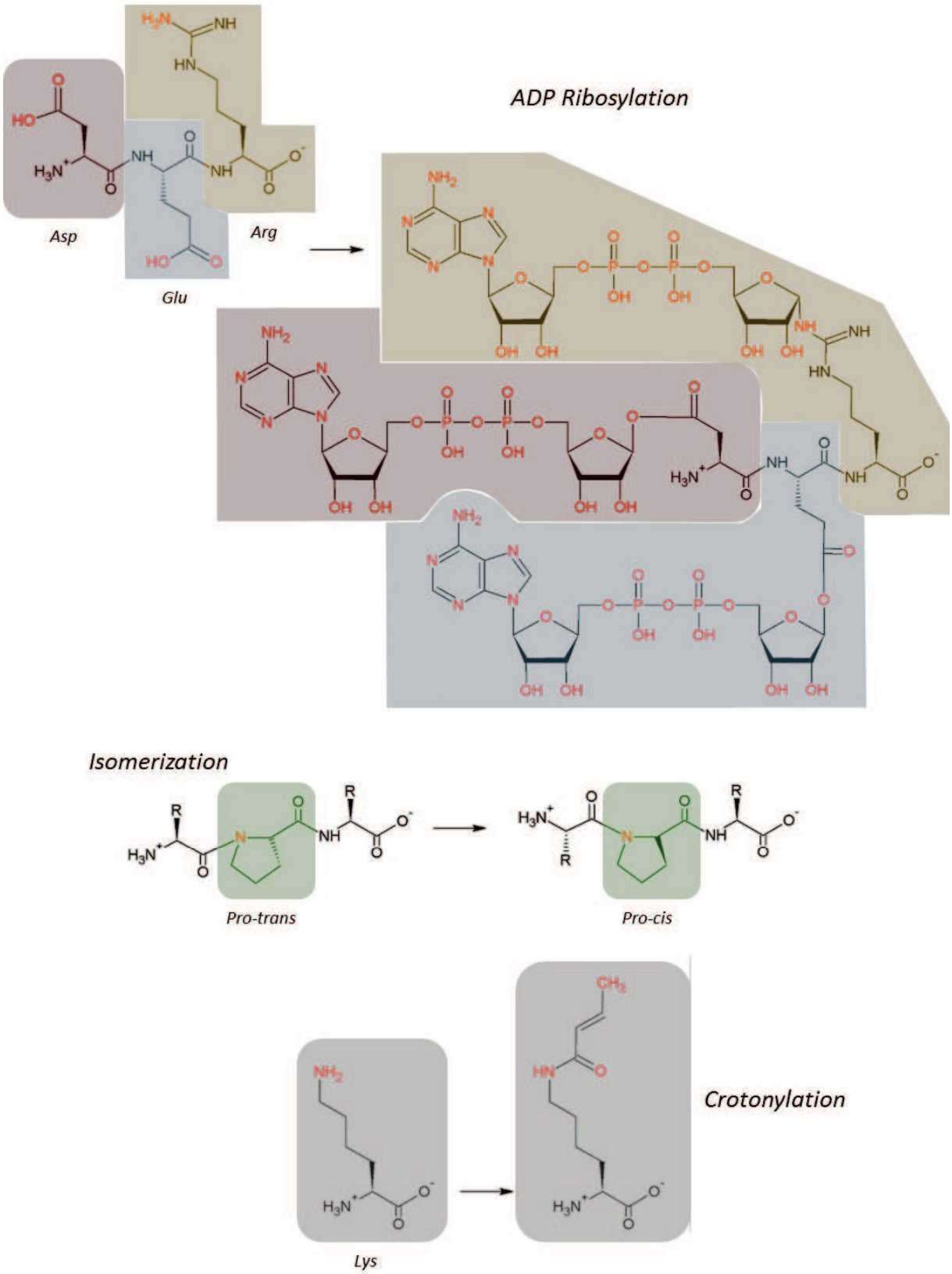


FIGURE 21

Structural overview of the different histone post-translational modifications

ADP-ribosylation is the addition of an ADP-ribose onto an arginine, glutamate or aspartate residue. This modification can be either mono- or poly-, and the enzymes responsible for this modification are then called MARTs (Mono-ADP-ribosyltransferases) and PARPs (Poly-ADP-ribose polymerases)[282]. The opposite reaction is achieved by PARGs (Poly-ADP-ribose-glycohydrolases). This modification is associated with a loosening of the chromatin, one negative charge being added with this modification.

Proline isomerisation consists in switching the conformation of proline residues from *cis* to *trans* and vice versa. This modification induces an important change in the structure of the peptidic backbone. Three families of enzymes are able to isomerize prolines: cyclophilins, FKBP (FK506-binding protein) and parvulins. For example, FPR4, belonging to the FKBP family, is responsible for *trans-cis* isomerization of H3P30 and H3P38 residues, regulating thus the methylation level of H3K36 through inhibition of Set2p[283,284].

Crotonylation is the addition of a crotonyl moiety, supplied by crotonyl-CoA, on ε-amines of lysine residues, similar to acetylation. This modification has nonetheless be discovered only very recently[285], and its role remains unanswered. It is noteworthy, however, that this crotonyl moiety is way more rigid than an acetyl moiety, owing to its double C-C bond. Thus, one can easily imagine that specific enzymes, different from the HATs and HDACs, are involved to set up or remove this modification. We can mention that crotonylation seems to be, at least partly, involved in X-related genes reactivation in spermatids. Pre-meiotic primary spermatocytes indeed undergo an inactivation of sex chromosomes, followed in post-meiotic spermatids by a general transcription inactivation, necessary for the replacement of histones by protamines[286,287]. It has however been shown that around 18 % of X-related genes were reactivated during spermatid elongation into spermatozoids[286,288], and that these genes were all crotonylated[285].

3.3 Reading methylated DNA

DNA methylation is generally linked to a repression of gene expression. This repression may be the consequence of three different mechanisms. First, methylation can have a direct structural consequence on chromatin, forcing nucleosome repositioning and modifying thus the ease of access for the transcription machinery. Also, CpG island methylation can inhibit direct DNA-protein interactions like CREB[289], c-Myc[290], E2F1[291] or UBF1[292]. Finally, 5mC can be directly recognised by methylated DNA-binding proteins.

Among the latter, three classes can be distinguished, classified upon their structural domains. The first family that was discovered, the MBPs (Methyl-Binding Proteins), have a MBD domain (Methyl CpG-Binding Domain). Later, a zinc finger protein, Kaiso, was identified, as well as the two related proteins ZBTB4 and ZBTB38. And more recently, the third family of SRA domain proteins (Set and Ring finger-Associated) was identified, with two representatives, UHRF1 and UHRF2.

3.3.1 The MBD-containing proteins

The study of MBPs started in 1989, after the fortuitous discovery of two proteins binding to methylated DNA. At this time, Bird and collaborators were seeking proteins able to bind to non-methylated DNA and likely to protect CpG islands from methyltransferases. Electromobility shift assays from mouse liver nuclear extracts showed the presence of two proteins, called MeCP1 and MeCP2 (Methylated CpG-binding Protein 1 and 2)[293,294]. MeCP2 was the first to be purified from mouse brain extracts. This 53-kilo Daltons protein exhibits what has then been described as a 90 residues N-terminal MBD domain (*figure 22*), as well as a C-terminal transcription repression domain (TRD)[295,296]. Sequence similarity searches in databases identified four other proteins, MBD1, MBD2, MBD3 and MBD4, all very conserved in vertebrates[297]. Among those MBPs, MBD2 and MBD3 share the highest sequence identity (77 %). Furthermore, a single homolog of these two proteins, MBD2/3, is found in invertebrates. It is encoded by a single gene, in contrast to vertebrates where this gene probably underwent a duplication event. Indeed, mbd2 and mbd3 genes have a very similar genomic structure, varying only by the size of their introns. This supports the idea that MBD2 and MBD3 are probably the ancestral representatives of this family[298,299]. Later, the protein MeCP1 initially discovered along with MeCP2 turned out to be a MBD2/HDAC1 complex[300].

MBPs are ubiquitous proteins, nevertheless exhibiting strong disparities depending on the cellular type and development stage. For example, in embryonic stem cells, MBD3 is the only predominantly expressed MBP. At the blastula stage of organismal development, MBD2 and MBD4 become detectable, and finally MeCP2 after the blastocyst implantation[301,302]. In adults, expression patterns depend on the cellular type: MBD3 (along with MeCP2 and MBD1) is highly expressed in the brain, notably in the olfactory bulb, cerebellum, hippocampus and prefrontal cortex[303,304], while MBD2 has an almost opposite expression pattern, with mRNA quantities up to twenty times higher in some tissues, such as breast cells or cultured HeLa cells[305].

One characteristic of MBPs is that they associate with chromatin remodelling complexes. MeCP2 can thus interact with the Sin3 complex to induce transcriptional repression, together with HDAC1 or HDAC2[296]. This protein was long studied during these last few years, since mutations of its gene are responsible for the Rett syndrome[306], lethal in men and causing neurological and psychiatric disorders in women. In 2008, Chahrour showed that MeCP2 was able to regulate a variety of genes, suggesting that this proteins plays more a global role than just specific in expression regulation[307]. This was confirmed in 2010 by works carried out by Skene, who showed that one MeCP2 protein was present every two nucleosomes in neural cells[308]. In this respect, Karolin Luger published in 2011 the SAXS envelope of a nucleosome in complex with MeCP2[309].

MBD1, of which the X-ray structure in complex with DNA is available[310], shows the very same recognition mode of methylated CpG islands, but can also recognise unmethylated CGCG motifs through its CXXC domain[311]. In 2010, Clouaire showed that MBD1 possesses, in addition, a sequence specificity since it is able to bind to T^{me}CGCA and TG^{me}CGCA motifs[312]. This protein proved to be an

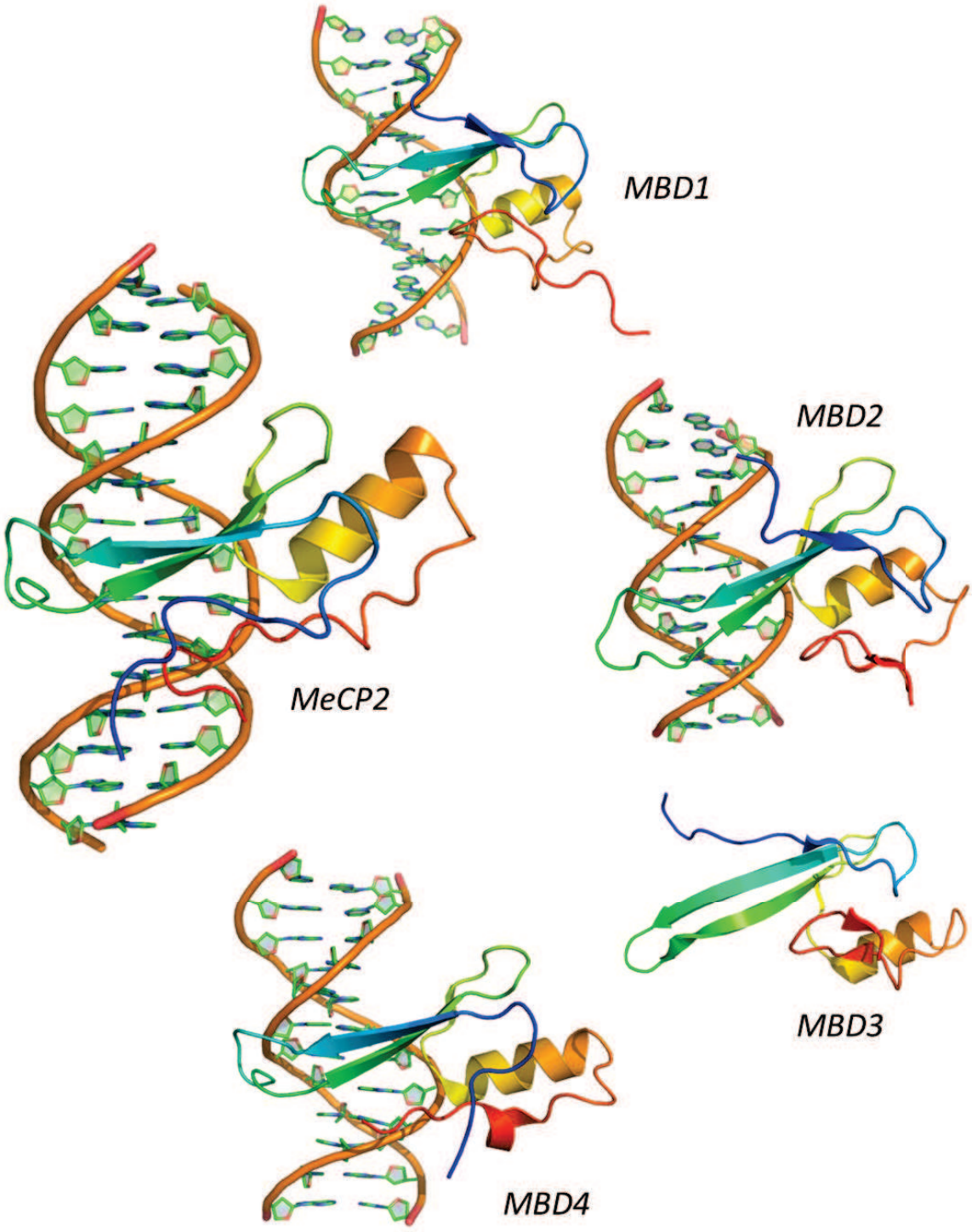


FIGURE 22

The MBD domain

Structural overview of the MBD domain of five MBPs. All MBDs apart from MBD3^{MBD} have been solved in complex with a methylated DNA oligo. MeCP2^{MBD} has been solved by X-ray crystallography at 2.5 Å resolution (*pdb: 3c2i*); MBD1^{MBD} has been solved in solution by NMR (*pdb: 1ig4*), as well as MBD2^{MBD} (*pdb: 2ky8*) and MBD3^{MBD} (*pdb: 2mb7*). Finally, MBD4^{MBD} has been solved by X-ray crystallography at 2.5 Å resolution (*pdb: 4lg7*).

inhibitor for miR-184 microRNA synthesis[313]. This microRNA is itself an inhibitor for stem cell differentiation into nerve cells.

MBD2 is associated with a repressive action on gene expression, since it is able, like MeCP2, to interact with several partners like HDACs. MBD2 knockout experiments in mouse led to an aberrant expression of pancreatic genes, suggesting a tissue-specific role for this protein[314]. Also, MBD2 intervenes in the regulation of several genes like BRCA1[305], pS2[315], TERT[316] or IL4[317].

MBD3 shares 77 % of sequence identity with MBD2, but because of two point mutations inside its MBD domain (Lys30His and Tyr34Phe), found in mammals, this protein lost its ability to bind to methylated DNA. Its role is still unclear, and hypothesis have been made, suggesting it could be, at least in part, not related to methylation levels of the DNA. It would then act indirectly, through other partners, but its implication within the NuRD complex leads to believe that it is still essential, as evidenced by the lethal effect of MBD3 knockout in mouse[318-320]. MBD3 being the central topic of this work, it will be described in more details in the coming pages.

Finally, MBD4 is the only MBD protein that is not associated with a histone deacetylase activity. And when the other proteins (with the exception of MBD3 in mammals) bind symmetrically methylated CpG islands, MBD4 shows a preference for monomethylated CpG islands, with a TpG mismatch. This mismatch is due to the deimination of a 5mC within a fully methylated CpG island. MBD4 ensures thus the recruitment of enzymes for this mismatch repair[321-323]. We can mention that mutations in the MBD4 gene lead to carcinomas, because of microsatellite instability, where these mismatches are frequently observed.

3.3.2 Kaiso and the zinc finger proteins

Kaiso was discovered in 1999 and described as a methylated DNA-binding factor, implicated in non-canonical WNT signalling pathway, essential for gastrulation[324,325]. Unlike the MBP family, Kaiso and the two related proteins ZBTB4 and ZBTB38 possess a POX domain (POXvirus and zinc finger) responsible for protein-protein interactions, as well as three zinc fingers, of whose two are essential for methylated DNA-binding[326]. Kaiso is also able to bind a non-methylated DNA sequence, called KBS sequence (Kaiso-Binding Site)[327]. This sequence (TCCTGCNA) contains a TpG, structurally similar to a methylated CpG island. In *Xenopus laevis*, Kaiso knockdown showed early transcriptional activity of numerous genes, involved in development and apoptosis; this same phenotype was observed upon DNMT1 knockdown, suggesting a close relationship between Kaiso and the methylation state of DNA[328].

ZBTB4 and ZBTB38 are able, like Kaiso, to bind to methylated DNA and to the KBS sequence. However, methylated DNA-binding is sequence-specific, highlighting thus a non-redundant but rather specific role for these two proteins[329].

3.3.3 The SRA-domain proteins

In mammals, UHRF1 and UHRF2 (Ubiquitin like containing plant Homolog domain (PHD) and RING Finger domains 1 and 2) are the two representatives of the SRA-domain family (Set- and Ring finger-Associated). UHRF1, also named ICBP90 (Inverted CCAAT box-Binding Protein of 90 kDa) or Np95 in mouse, possesses a UbL domain (Ubiquitin Like) involved especially in histone H3 ubiquitination[330], a leucine zipper, a tandem Tudor domain to interact with H3K9me3[331], and two zinc finger domains. Of these two, the first one is a PHD-like zinc finger, able to recognise methylated histones on lysine 4; and the second one is a RING-like zinc finger, guiding protein degradation by the proteasome. Finally, an SRA domain allows UHRF1 to bind to methylated DNA without sequence specificity[332] as well as hemimethylated DNA[333,334]. The structure of this SRA domain addressed the cooperation between UHRF1 and DNA methyltransferases[335-337]. Indeed, by interacting with a hemimethylated CpG island, UHRF1 allows the 5mC to flip out of the double helix and directs DNMT1 to methylate the cytosine on the complementary strand. UHRF1, together with DNMT1, is thus an essential component of the methylome transmission throughout DNA replication.

Historically, UHRF1 was identified in mouse, appearing specifically during the S phase in thymocytes but constitutively expressed in lymphocytes. Plus, sequence analysis revealed interactions with the A/E cyclin and the Rb protein, granting UHRF1 a role in proliferation and cell cycle[338]. This suggestion was validated a few years later, with the highlight of a post-translational regulation of UHRF1 by PKA and CK2 kinases[339,340], underlining its role in the G1/S transition, through repression of the genes p53, p21[341,342] and Rb[343].

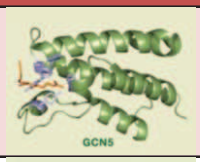
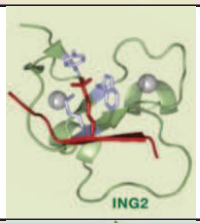
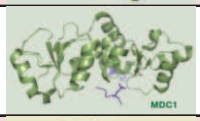
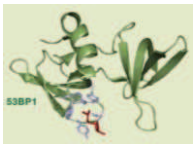
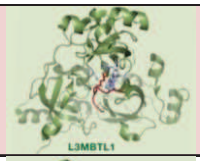
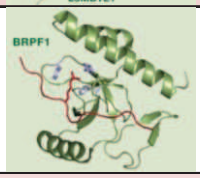
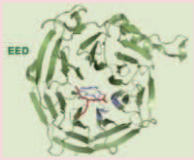
To finish, UHRF1 study in the tumour context allowed to qualify this protein as a proto-oncogene. It is indeed able to recognise methylated promoters of tumour suppressor genes like $p16^{INK4}$ and $p14^{ARF}$ and to induce their repression by recruitment of HDAC1[332]. Its interaction with DNMT1, noted above, allows also this protein to activate the expression of the growth factor VEGF, responsible for tumour proliferation[344].

UHRF2, also named NIRF (Np95/ICBP90-like ring finger) or Np97 in mouse, shares the same domains as UHRF1 and seems to be involved in the same mechanisms, namely the cell cycle regulation. Major differences can nonetheless be noted. First of all, overexpression of UHRF2 causes a blockade in cell cycle progression at phase G1, unlike UHRF1 which allows the G1/S transition[345]. Also, UHRF2 is expressed in differentiated cells, while UHRF1 rather is in pluripotent stem cells[346]. These observations suggest the existence of a feedback between the two, to regulate the activity of the cell cycle and differentiation.

3.4 Reading the histone code

In 1999, Christophe Dhalluin became interested in bromodomains, found in a large majority of HATs-associated coactivators. The structure and function of this 110-amino acid domain were then still unknown, and Dhalluin solved the first solution structure using NMR of the PCAF (p300/CBP associated protein) bromodomain[347]. He was then able to highlight the binding mode of

TABLE 9
Histone recognition modules
Modified from Tatiana Kutateladze. SnapShot: Histone Readers. Cell. 2011 September 02. (146):842

Recognition module		Substrate	Examples	Structure
Bromodomain		H3KAc H4Kac	BRG1 CBP/p300 GCN5 PCAF RSC4 $TAF_{II}250$	
PHD domain		H3K4me2/3 H3K9me H3	BHC80 CHD4 DNMT3L DPF3b ING1/2/3/4/5 JARID1A KDM7A RAG2 TAF3 TRIM24	
14-3-3 domain		H3S10ph H3S28ph	14-3-3-zêta	
BRCT domain		H2AXS139ph	BRCA1 MCPH1 MDC1	
The Royal family	**Chromodomain**	H3K9me2/3 H3K27me2/3 H3K27me2/3 H3K9me2/3 H3K36me3/2 H3K4me	CHD1 CHP1 HP1 Tip60	
	Tudor domain	H4K20me2 H3Rme2 H4Rme2 H3Kme3 H4Kme3	53BP1 CRB2 FXR JMJD2A UHRF1	
	MBT domain	H3Kme1/2 H4Kme1/2 H3K9me2/3	L3MBTL MBTD1 SCM SCML2 SFMBT	
	PWWP domain	H3K36me3 H4K20me1	BRPF1 DNMT3A PDP1	
WD40 motif		H3R2 H3Kme3 H4Kme3 H1Kme3 H4 (tail)	EED RbAp46/48 RCC1 WDR5	
UBD domain		H2Aub H2Bub	Cks1 Dsk2 Npl4 Rabex-5 S5a	

this domain with acetyllysine residues, similar to that of histone acetyltransferase with acetyl-CoA, and showed that bromodomain could directly interact with histone acetylated lysines, thus making it the first recognition module of an epigenetic mark.

Since then, several groups discovered recognition modules by studying conserved domains of chromatin-associated proteins (*table 9*). It is in this way that Jacobson discovered the role of the double bromodomain of TAF1[154]; and that Bannister studied the recognition of methylated lysines by the HP1 chromodomain[348]. Finally, the recent developments in high throughput screening offered the potential to rapidly accelerate the identification of reading modules, by using modified histone peptides.

3.4.1 The bromodomain

Recognising histone acetylated lysines, the bromodomain is a highly conserved structural motif, with a left-handed up-and-down four-helix bundle, namely α_Z, α_A, α_B and α_C and two loops ZA and BC. The whole constitutes an aromatic pocket which stabilises the motif and binds acetyllysines. Two conserved tyrosine residues, the first one in the ZA loop, and the other one at the C-terminal end of the α_B helix, are found in a majority of bromodomains[349], although they are not primary determinants for the recognition of acetylated lysines[350]. Another highly conserved residue, an asparagine found in the BC loop, is able to form a hydrogen bond between its lateral amine group and the carbonyl group of the acetyllysine[351] (*figure 23*). The adjacent residues within the ZA and BC loops are more variable, conferring the specificity of substrate. Indeed, these loops interact not only with the acetylated lysines, but also with the n-1, n+1, n+2 and n+3 residues around[351,352].

Sometimes, two bromodomains can be repeated within a protein, as is the case in $TAF_{II}250$[154], forming a U-shape structure (*figure 24*). In that case, each bromodomain folds independently, and the aromatic pockets are separated by around 25 Å, which corresponds to 7 to 8 residues on the target protein. Thus, $TAF_{II}250$ has a significantly higher affinity for di- or tetra-acetylated peptides like H4K5ac/K12ac, H4K8ac/K16ac or H4K5ac/K8ac/K12ac/K16ac.

3.4.2 The PHD domain

The PHD (Plant Homeo Domain) domain (or finger) is a motif composed of 50 to 80 amino acids, found in a multitude of proteins. It is a Cys_4-His-Cys_3 type domain, coordinating two zinc ions (*figure 25*). Initially, the PHD domain was considered as a protein-phospholipid or protein-protein binding domain. But its presence beside other domains like the bromodomain or the PWWP domain recently led to believe that this domain could also be a histone recognition module. And it was not until 2006 that several groups described the binding specificity towards methylated H3K4[353-358]. And then in 2008, other studies allowed to specify the valence state. The PHD domain thus showed a high specificity for trimethylated lysines[258,359-363].

Recently, the BHC80/PHF21A protein, composed of a PHD finger, was found to bind unmethylated lysines. This protein is part of the LSD1 complex which demethylates H3K4me2 into H3K4me0[364]. BHC80 probably plays a role in preventing remetylation of the residue.

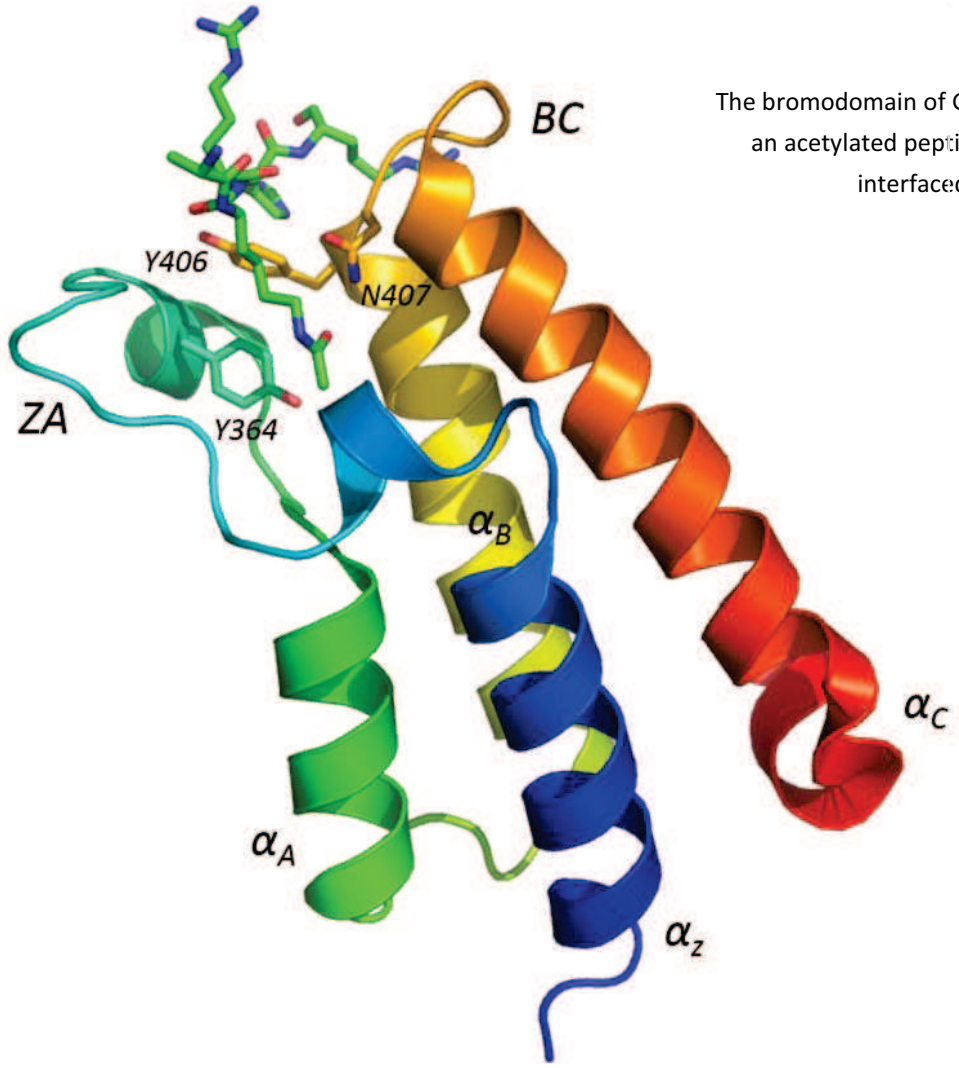


FIGURE 23

The bromodomain

The bromodomain of GCN5 in complex with an acetylated peptide of histone H4. The interfaced residues are shown.

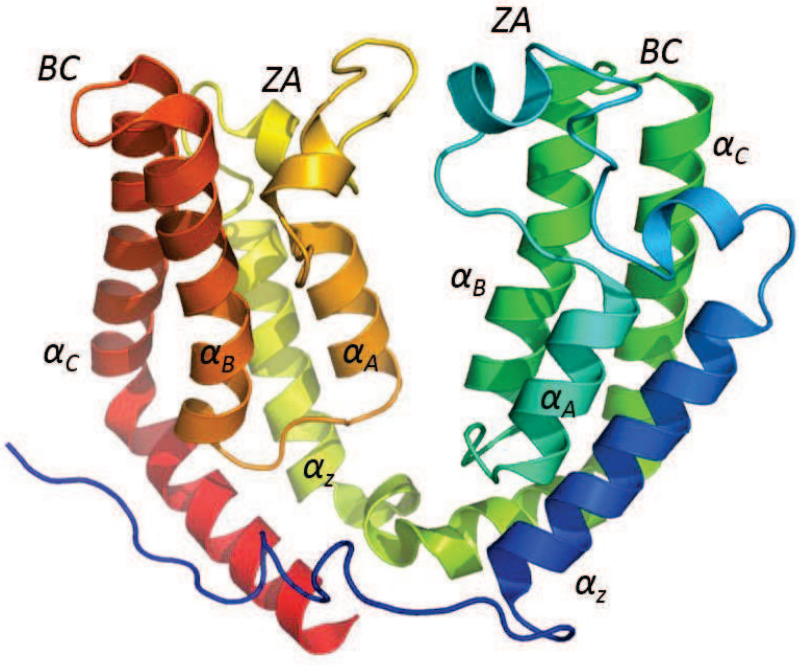
(*pdb : 1e6i*)

FIGURE 24

The tandem bromodomain

The tandem bromodomain of $TAF_{II}250$ and its U-shape structure.

(*pdb : 1eqf*)



Even more recently, the DPF3b protein found in the BAF complex showed its affinity for histone acetylated lysines[365], ensured by its tandem PHD fingers[366].

3.4.3 The 14-3-3 domain

The 14-3-3 domain is characteristic of a family of proteins called 14-3-3. This domain is able to bind phosphorylated serines and threonines. In particular, the protein 14-3-3-ζ is specific for the residue H3S10ph, phosphorylated during mitosis[367] (*figure 26*).

3.4.4 The BRCT domain

The BRCT domain was identified in the protein BRCA1 (Breast Cancer protein 1) and is often mutated or even deleted in breast cancers. This domain is implicated in cell cycle regulation and DNA-damage signalling pathways, through recognition of H2A residues when phosphorylated upon double-strand breaks[368]. It is made up with three α helices, surrounding a small sheet of four parallel β-strands (*figure 27*). This motif is generally repeated twice, in a head-to-tail fashion, although only the first domain binds the phosphoserine residue. The interface between both domains, nevertheless, allows recognition of a phenylalanine residue in carboxy position of the phosphoserine, for a higher binding specificity.

In the protein MDC1 (Mediator of DNA damage Check-point protein 1), the twin BRCT domain allows specific binding of the H2A.XS139ph residue[369,370].

3.4.5 The Royal family

The Royal family comprises structurally related domains, probably sharing a common ancestor, and presenting a specific affinity for methylated substrates[371]. They comprise a long and curved 3-stranded β-sheet and a small 3_{10} helix.

3.4.5.1 The chromodomain

The chromodomain (Chromatin Organization Modifier) is a domain composed of 40 to 50 residues, able to read methylated histones, especially H3K9 and H3K27[372]. The canonical structure of this domain is composed of a 3-stranded antiparallel β-sheet and a C-terminal α-helix[373,374] (*figure 28a*). This domain may be attributable to the OB-folds class (Oligonucleotide/Oligosaccharide)[375]. Furthermore, chromodomain proteins can be subdivided into three categories: those having a single chromodomain (Polycomb, Suv39, HP6, Tip60, Myst1, etc.); those with paired tandem chromodomains (CHD proteins); and those with an N-terminal chromodomain followed by a chromo shadow domain (HP1, Rhino).

The role of the chromodomain was highlighted in 2001: using *in situ* immunohistochemistry, several groups were able to show the close relationship between the methylation state of H3K9 and

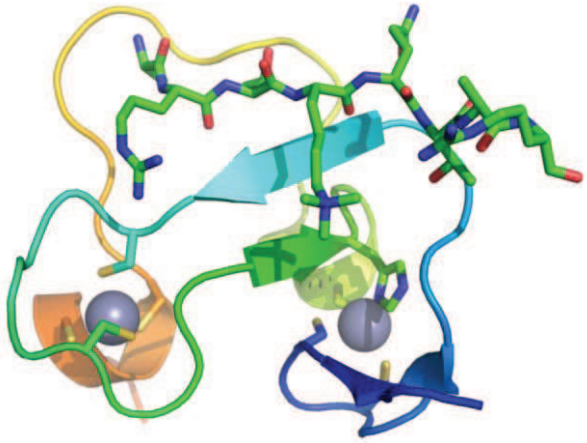


FIGURE 25

The PHD domain

The PHD domain of ING2 in complex with a K4 trimethylated peptide of histone H3. Zinc atoms are shown as purple spheres.

(*pdb : 2g6q*)

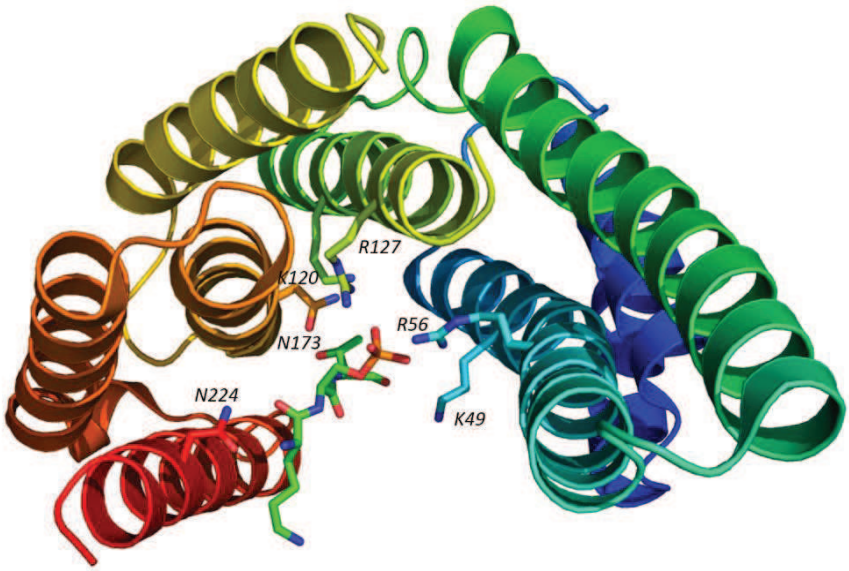


FIGURE 26

The 14-3-3 domain

The 14-3-3 domain of 14-3-3-zêta in complex with a phosphorylated peptide of histone H3. The interfaced residues are shown.

(*pdb : 2c1n*)

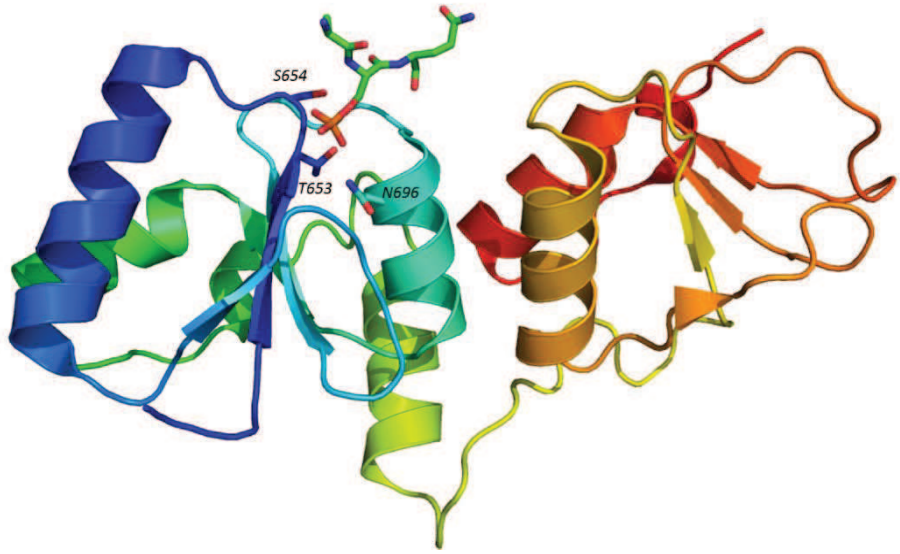


FIGURE 27

The BRCT domain

The tandem BRCT domain of MCPH1 in complex with a phosphorylated peptide of histone H2A.X. The interfaced residues are shown.

(*pdb : 3shv*)

the localisation of the HP1 protein[348,376-379]. Quickly, the first X-ray structure of the yeast HP1 chromodomain, in complex with a methylated peptide, was solved[380]; and affinity studies allowed to determine a K_D of around 10 µM for H3K9me2 and H3K9me3, ten-fold lower for H3K9me1, and above 1 mM for H3K9me0[381]. In 1992, Epstein pointed the internal homology between the HP1 chromodomain and the C-terminal sequence of the protein[382]. Three years later, the term "chromo-shadow domain" was proposed to name this similar but not identical structure[383]. Two distinct roles were suggested for this shadow domain: first, it allows the dimerisation of two HP1 proteins[384-387], in order to link two adjacent nucleosomes or the tails of H3 histones within a same nucleosome, to minimise their moving[388]. Then, it could be able to specifically recruit proteins exhibiting the PxVx[L/M/V] motif[389,390].

By sequence homology, the structure of the Polycomb chromodomain had to be similar to that of HP1. This was the case, with a comparable recognition mechanism, but an enhanced specificity for H3K27[391,392].

Finally, there is a third case in which two chromodomains are repeated in tandem, also called double chromodomains. This is the case in the CHD1 protein, an ATPase that interacts with the methylated H3K4 residue[393].

3.4.5.2 The TUDOR domain

In 1993, the Tudor domain was described in the same named protein in *Drosophila melanogaster*[394]. This domain comprises around 50 residues, and is often found in RNA-binding proteins[395]; but it is also able to recognise methylated histone residues. In human, Tudor, as described in the protein SMN (Survival of Motor Neuron) is made up with five β-sheets, folded so as to form a barrel-like structure (*figure 28b*).

3.4.5.3 The MBT domain

The structural MBT domain (Malignant-brain-tumour) is able to specifically bind histone lysines which are mono- or dimethylated[396]. Studies suggest nevertheless that there is no sequence specificity upon binding to the target protein. The interaction would then only be possible through the methylated lysine which would be strongly anchored inside a cavity of the domain, and stabilised by cation-π and van der Waals interactions[397] (*figure 28c*). This explains also the inability to bind to trimethylated lysines, due to the limited size of this cavity.

The L(3)MBTL1 protein, for example, has three MBT domains, able to interact with two nucleosomes at a time[398]. This protein aims thus at bringing closer together two nucleosomes, even if they are not physically adjacent.

3.4.5.4 The PWWP domain

The PWWP domain was named after the four conserved residues which constitute it: proline-tryptophan-tryptophan-proline. These are found within a five-stranded β-sheet lying on a

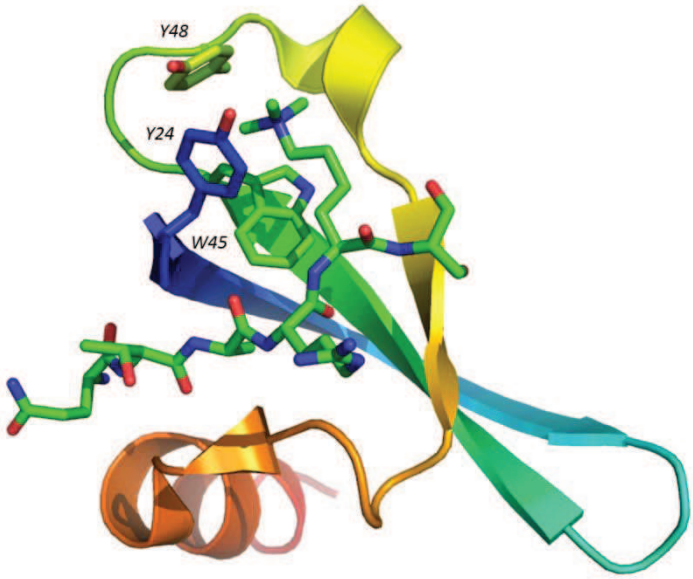


FIGURE 28a

The chromodomain

The chromodomain of HP1 in complex with a K9 trimethylated peptide of histone H3. The interfaced residues are shown.

(*pdb : 1kne*)

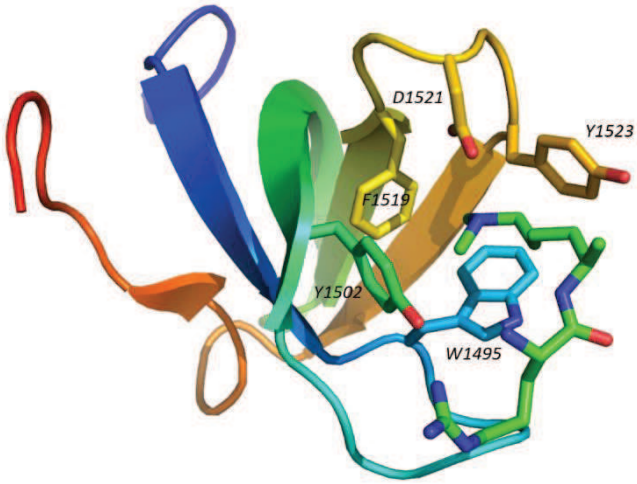


FIGURE 28b

The TUDOR domain

The TUDOR domain of 53BP1 in complex with a K20 dimethylated peptide of histone H4. The interfaced residues are shown.

(*pdb : 2ig0*)

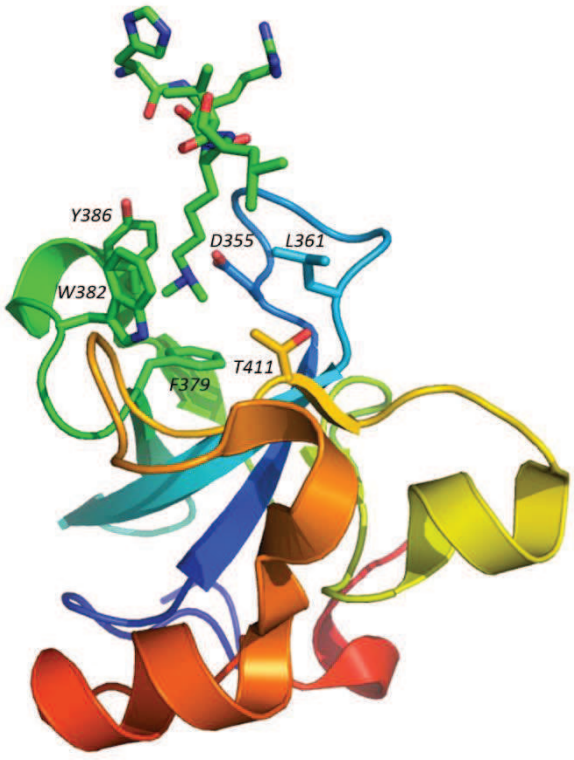


FIGURE 28c

The MBT domain

The tandem MBT domain of L3MBTL1 in complex with a K20 dimethylated peptide of histone H4. The interfaced residues are shown.

(*pdb : 2pqw*)

five-α-helix bundle (*figure 28d*). This motif was initially described as a non-specific DNA-binding domain[399-401]. Later however, its structural similarities with other domains of the Royal family suggested that PWWP was also able to bind methylated ligands[402].

The Peregrin protein (or Brpf1), which possesses a bromodomain and a PHD domain, also contains a PWWP domain which allows it to specifically bind H3K36me3[403].

3.4.6 The WD40 motif

The WD40 motif is one of the most predominant interaction motif in Eukaryotes. It comprises around 40 residues and often ends with the doublet tryptophan/aspartate. Four to sixteen of these motifs are repeated, forming a barrel-shaped domain, composed of several β-sheets blades, depending on the number of repeats (*figure 29*). This highly rigid structure serves as a stable platform for multiprotein complex formation, allowing protein-protein interactions as initially described, but also protein-methylated DNA interactions. The WDR5 protein has been shown to target the H3K4me2 residue, promoting its trimethylation[404]. But recent studies have suggested that WDR5 does not directly target H3K4 but H3R2 instead, and helps flipping H3K4 towards modification enzymes[405-408].

The EED protein, found in the PRC2 complex together with the EZH2 methyltransferase, seems to directly target methyllysines, unlike WDR5. Indeed, available structures of this protein in complex with methylated H1, H3 and H4 peptides show that the targeted methyl group points to the centre of the barrel, within an aromatic cage[409,410].

Finally, RbAp46 and RbAp48, which will be described later in this manuscript, also possess a seven-blade WD40 domain. The X-ray structures available to date show that these proteins have not one but several protein-protein interaction sites. Indeed, the central region of the barrel is involved in protein factor binding, like FOG-1; but a pocket, located on the side of the barrel, also allows interaction with non-modified histones[15,411,412].

3.4.7 The UBD domain

The UBD domain (Ubiquitin-Binding Domain) is able, as its name implies, to bind ubiquitinated histone residues. To date, eleven families of structurally different UBD domains were described. However, they all seem to share a common recognition mode of the ubiquitinated peptide, through binding on a hydrophobic region centred on residue Ile44[413].

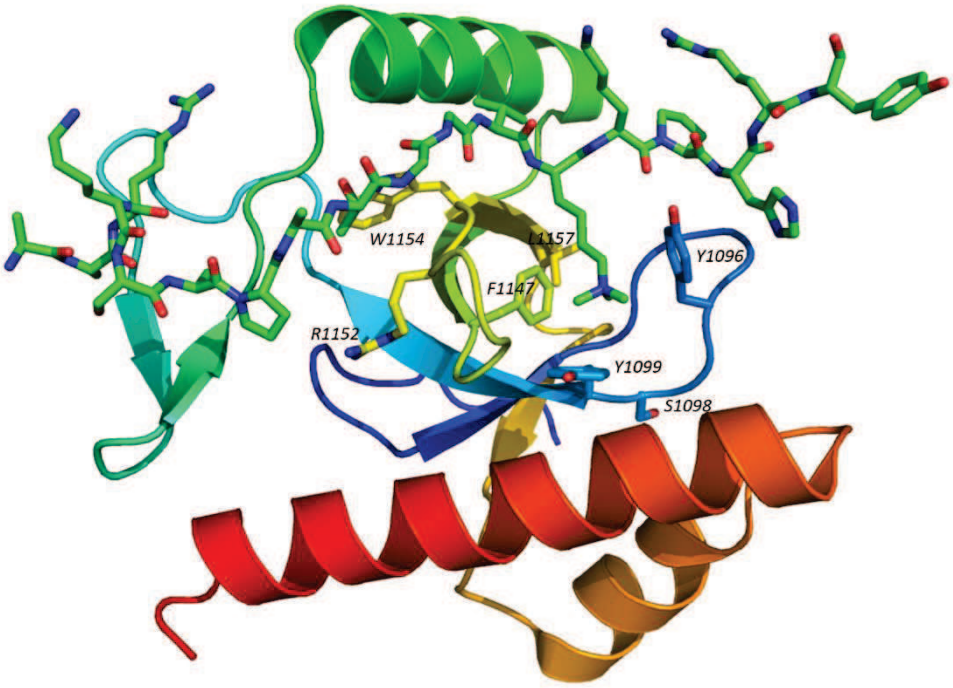


FIGURE 28d

The PWWP domain

The PWWP domain of BRPF1 in complex with a K36 trimethylated peptide of histone H3. Interfaced residues are shown.

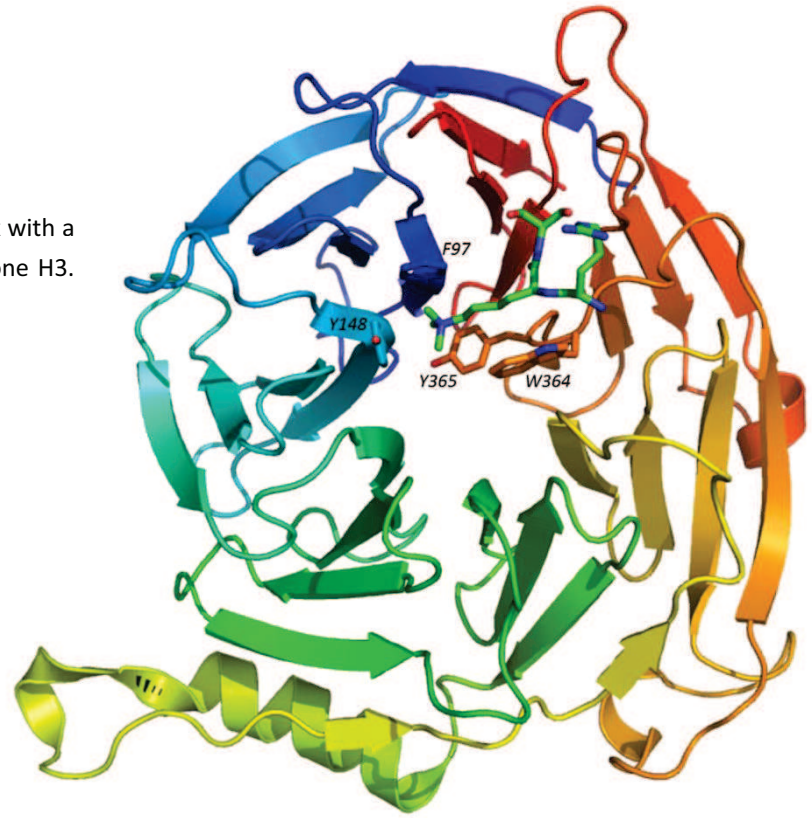
(*pdb : 2x4y*)

FIGURE 29

The WD40 motif

The WD40 motif of EED in complex with a K27 trimethylated peptide of histone H3. The interfaced residues are shown.

(*pdb : 3jzg*)



3.5 The ATP-dependent remodelling

Chromatin remodelling complexes utilize the energy brought by the hydrolysis of ATP, universal energy carrier, to destabilise the interactions between the histone octamer and the DNA[414]. The structure of the chromatin is thus altered, either by displacement of nucleosome along the DNA, to ensure access of a targeted sequence, or by eviction or replacement of histones (*figure 30*). The molecular mechanisms involved remain subject to debate, and several models are proposed[415]. Among those models, the twist-diffusion is based on the idea that the energy brought by ATP allows a rotation of the DNA at one extremity of the nucleosome, which would be little by little transmitted to the following nucleotides, making thus the DNA "crawl" on the nucleosome (*figure 31*). But studies showed that this phenomenon could not be inhibited in the presence of abasic sites[416], breaks or bulky molecules[417], which runs counter to this model. The second model to be proposed is the loop/bulge propagation model[418] (*figure 31*). ATP hydrolysis furnishes the energy to locally separate the histone octamer and the DNA. This results in the formation of a small bulge, which can propagate along the nucleosome, but also interact with transcription factors.

These remodelling complexes are all part of the SNF2 protein family, and depending on the subunits they contain, can be classified in several subfamilies described hereafter (*figure 32*).

3.5.1 The Swi2/Snf2 subfamily

The Swi2/Snf2 subfamily includes the ySWI/SNF and yRSC complexes from yeast, the Brahma complex from *Drosophila melanogaster* as well as the hBRM and hBRG1 complexes from human. All these complexes have a highly conserved subunit belonging to the Snf2 protein family, with an ATPase domain and a C-terminal bromodomain. The archetype of this subfamily, Swi2/Snf2, was discovered in *Saccharomyces cerevisiae* during genetic studies, because of its role in sugar metabolism (Sucrose Non Fermentation, SNF2) as well as mating-type switching (SWI2).

The ySWI/SNF complex was the very first ATP-dependent chromatin remodelling complex to be described[419-421]. It has eleven subunits, including Swi2/Snf2. Some of these subunits interact with transcription factors, allowing the specific recruitment of the complex to remodelling sites[422,423].

In *Drosophila melanogaster*, the Brahma complex includes eight subunits and is associated with a general transcriptional activation[424-426]. Knockouts of this complex showed the inability of the RNA Pol II to bind to chromatin[427].

In human, two 2-mega Dalton homologs of ySWI/SNF were purified. They contain DNA-dependent ATPase/helicase subunits and are called hBRG1 and hBRM. Although BRG1 and Brm share 75 % of identity, they play a different role as suggested by knockout experiments: Brm knockout mouse are viable[428], whereas BRG1 knockouts are not, from the embryonic stage[429]. Differential expression patterns were observed for these two proteins. Thus, BRG1 is mainly expressed in proliferating cells and fast-growing cells, while Brm is expressed in the brain, the liver or endothelial cells[430].

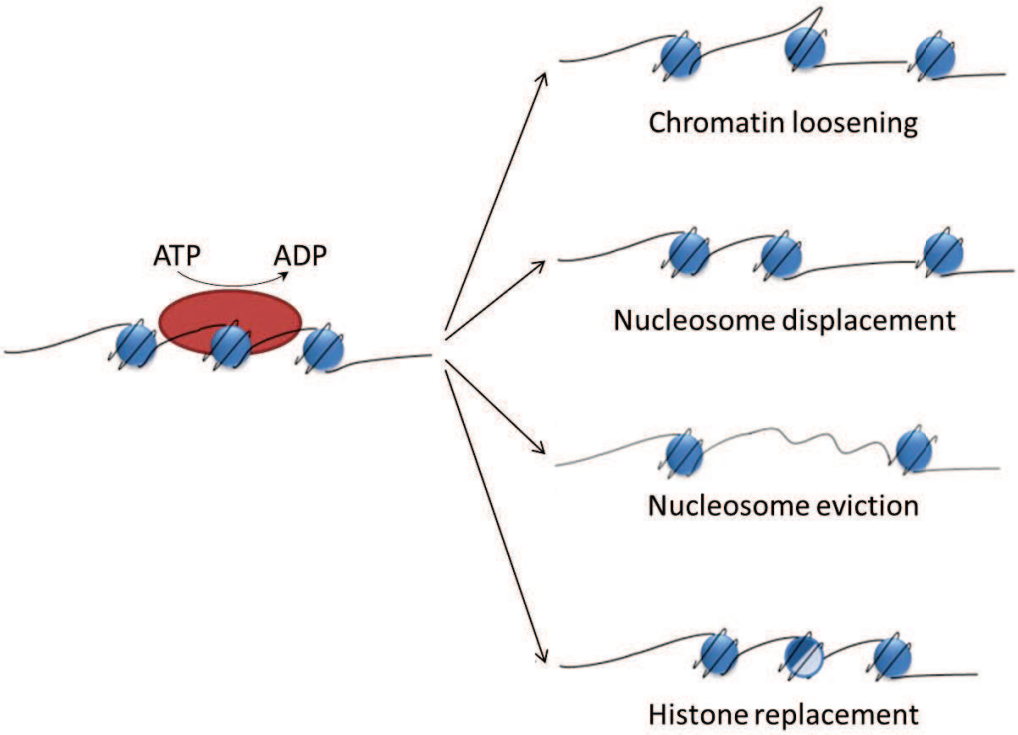


FIGURE 30

ATP-dependant chromatin remodelling

ATP-hydrolysis releases energy, which can alter the nucleosomal structure in several ways: loosening by histone-DNA contact dissociation, nucleosome displacement, eviction, and in some particular cases, histone variant incorporation.

Modified from Xu, Kanagaratham & Radzioch. Chromatin remodelling during host-bacterial pathogen interaction. Chromatin remodelling. 2013. Chapter 8.

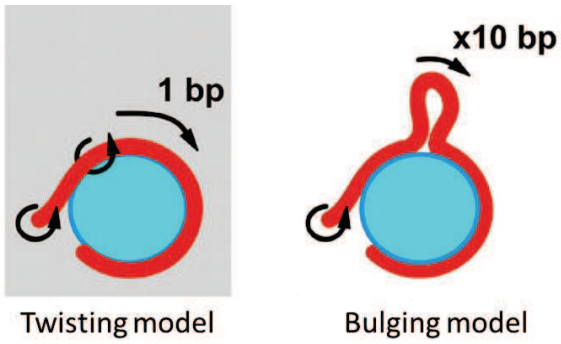


FIGURE 31

The chromatin remodelling mechanisms

Two different mechanisms are subject to discussion: the twisting model, allowing DNA to "crawl" on the histone octamer; and the bulging model, by local loosening of histone-DNA contacts and spreading.

Modified from Längst & Becker. Nucleosome mobilization and positioning by ISWI-containing chromatin-remodeling factors. J Cell Sci. 2001 Jul;114(Pt 14):2561-8.

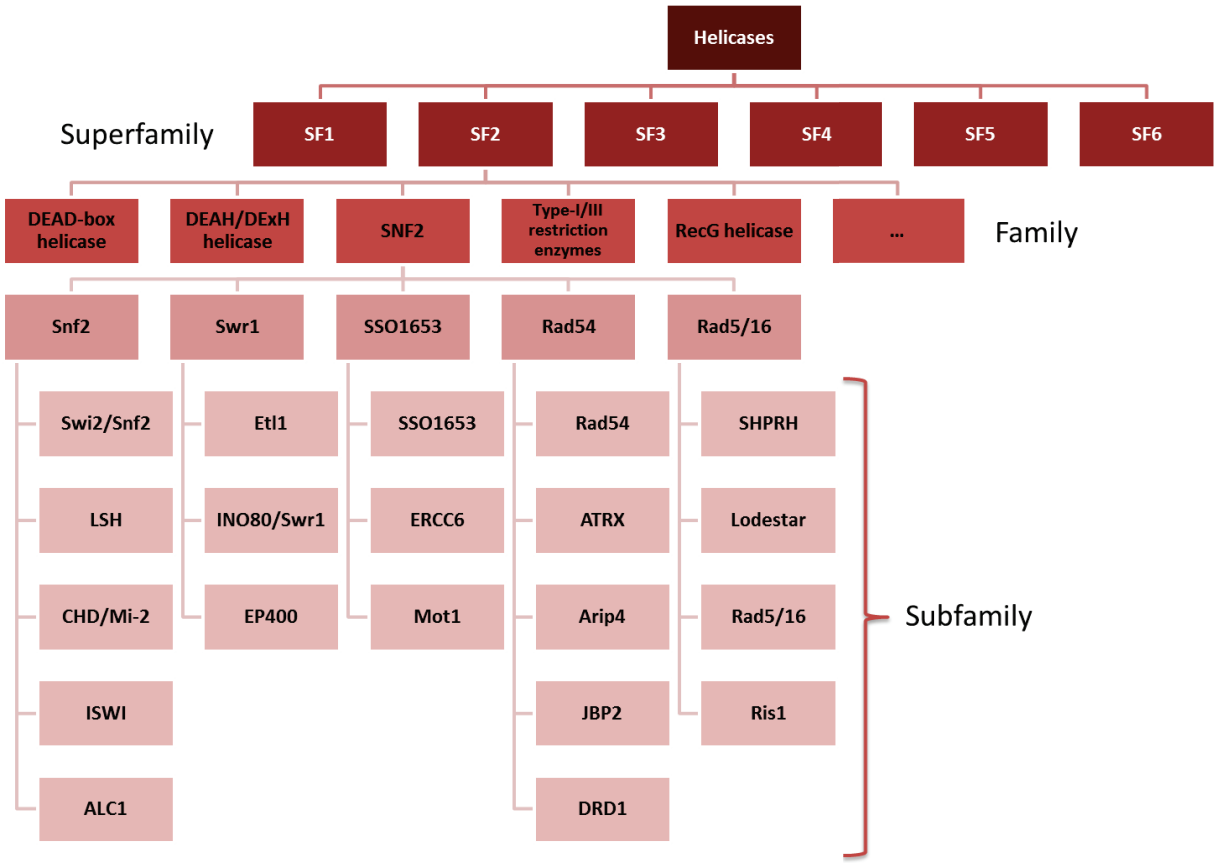


FIGURE 32

Helicases classification

Modified from Flaus A. et al. Identification of multiple distinct Snf2 subfamilies with conserved structural motifs. Nucleic Acids Res. 2006 May 31;34(10):2887-905.

3.5.2 The ISWI subfamily

The ISWI subfamily (Imitation Switch) brings together all the complexes containing an iSwi protein as ATPase subunit. This protein, discovered by sequence homology with Brahma[431], possesses two characteristic domains: a SANT domain (Switching-defective protein 3, Adaptor 2, Nuclear receptor co-repressor, Transcription factor IIIB) at the C-terminal end, responsible for non-modified histone tail binding; and a SLIDE domain (SANT-like ISWI) to interact with nucleosomal DNA[432-434].

The most studied complexes of this group –ACF, NuRF and CHRAC– were purified from *Drosophila melanogaster*[435-437]. Mammals have two homologs of iSwi: SNF2H (or SMARCA5), found in the CHRAF, NoRC and ACF complexes; and SNF2L (or SMARCA1), found in the NuRF and CERF complexes. These remodelling complexes seem to play an essential role as proven by studies carried out in *Drosophila melanogaster*[438].

The NuRF complex is composed of the iSwi protein, one catalytic subunit (NURF301), one pyrophosphatase (NURF38) and a homolog of RbAp46/48 (NURF55). NURF301 seems to play the role of organizational platform whilst interacting with transcription factors, allowing a specific and targeted recruitment of the complex[439].

ACF is a small complex composed of only two subunits, iSwi and Acf1[440]. It seems to play an important role in chromatin formation[441].

To finish, the CHRAC complex is identical to ACF, to which are joined two other small subunits, CHRAC14 and CHRAC16[442], which role could be to enhance nucleosome assembly and sliding[443].

3.5.3 The CHD/Mi-2 subfamily

The proteins of the CHD subfamily (Chromodomain, Helicase, DNA-binding domain) are composed of a characteristic pattern, with two tandem chromodomains at the N-terminal part, in addition to the ATPase domain. In yeast, only one CHD protein has been identified, yCHD1[444], while four in *Drosophila melanogaster* (dCHD1-4) and nine in mammals (hCHD1-9) exist. yCHD1 is closely related to d/hCHD1 and d/hCHD2 with approximately 35 % of overall sequence identity. The other CHDs on the contrary share only sequence similarity with yCHD1 within their defined domains, their extremities being highly variable.

CHD1 and CHD2 have a C-terminal DNA-binding domain, specific towards AT-rich regions[445]. CHD1 is a subunit of the SAGA (Spt-Ada-Gcn5-acetyltransferase) and SLIK (SAGA-like) complexes, in which it plays an important role in transcriptional activation[446], by keeping the chromatin in an open state. In *Drosophila melanogaster*, CHD1 has indeed been localised on chromosomes, at transcriptionally active sites[447]. Plus, the tandem chromodomain of hCHD1 binds selectively H3K4me3 at the 5' extremity of active genes[393].

CHD3 and CHD4, also called Mi-2α and Mi-2β, are the catalytic subunits of the chromatin remodelling complex NuRD. They are characterised by a PHD domain at their N-terminal end. They will be described in further details later in this manuscript.

The proteins CHD5 to 9 are the most variable of this family. They differ regarding C-terminal variable domains they possess, endowing them with various functions. For example, CHD7 is expressed in embryonic stem cells and seems to be responsible for the differentiation of certain cell lineages[448]. When mutated or deleted, it is also responsible for the CHARGE syndrome (Coloboma, Heart defect, Atresia choanae, Retarded growth and development, Genital hypoplasia, Ear anomalies/deafness)[449]. CHD9 on the other hand seems to have a specific and localised role in osteoblast differentiation[450].

3.5.4 The INO80/SWR1 subfamily

The INO80 protein (Inositol requiring 80) was fortuitly identified during a screening of proteins interfering in the inositol biosynthesis pathway in yeast[451]. It was then identified several

years later in human[452], as two orthologs: hINO80 and hSRCAP (SNF2-related CREB-activator protein). A broad range of functions was described for those complexes. INO80 thus seems to play a prominent role in repair mechanisms, as suggested by the hypersensitivity of knockout mutants to DNA damaging agents[453]. And a few years later[454-456], the SWR1 complex (NuA4 in human) showed its ability to replace canonical histone H2A with its variant H2A.Z within nucleosomes[456].

4. The NuRD complex

In 1998, several groups described a complex exhibiting an ATP-dependent remodelling activity, similar to that of ySWI/SNF from Saccharomyces cerevisiae, and coupled to a histone deacetylation function. This complex, called NURD, NRD, Mi-2 complex, and finally, NuRD, standing for "Nucleosome Remodelling and histone Deacetylation", is, to date, one of the two very unique complexes coupling two independent chromatin-remodelling activities[1,457-459], along with Tip60/p400 (NuA4/Domino)[460,461]. One possible reason for that could be that the ATP-remodelling activity is necessary for the HDAC subunits to access their target[462]. This idea is supported by the observation that in absence of ATP, deacetylation is only possible on histone octamers, and not on nucleosomes. The binding site of HDACs could be somehow protected by the DNA, and thus inaccessible. Experiments carried out to determine whether ATP could stimulate deacetylase activity did not show any significant effect on free histone octamers. By contrast, when nucleosomes were tested, ATP was shown to stimulate deacetylase activity by two-fold: without ATP, 30-35 % of acetylated H4 histones were deacetylated, while in the presence of ATP, 60-70 % were[1]. These data show that the ATP-dependent remodelling activity could facilitate the deacetylase activity of the complex, probably by exposing the substrates to the concerned subunits of the complex.

The NuRD complex is highly conserved among superior eukaryotes, and is expressed in a large variety of tissues. It forms a large macromolecular assembly that consists of different protein subunits (*figure 33*); however, different homologs and isoforms have been described for each of those subunits, leading to a horde of coexisting NuRD complexes, depending on the cellular, tissue, physiological or pathological context. Moreover, the stoichiometry of the different subunits remains an open question. Recently, the development of a new label-free quantitative mass spectrometry method, applied to the analysis of NuRD, suggested that it is composed of one CHD3 or CHD4 protein (Chromodomain, Helicase, DNA-binding domain), one HDAC1 or HDAC2, three MTA1/2/3 (Metastasis Associated), one MBD3 (Methylated CpG-Binding), six RbAp46/48 (Retinoblastoma Associated protein), two p66α or p66β and two DOC-1 (Deleted in Oral Cancer)[2]. Those data are nevertheless in contradiction with the structural analysis of the HDAC1/MTA1 complex showing a dimerisation of MTA1, suggesting the presence of two MTA1/2/3 and two HDAC1 or HDAC2 in NuRD[3].

Furthermore, it directly interacts with various partners, like the lysine specific demethylase 1 (LSD1/KDM1A)[4], Ikaros, Aiolos, Helios[5-7], B-cell lymphoma 6 (BCL6)[8,9], the oestrogen receptor α (ERα/NR3A1)[10-12] or Oct4/Sox2/Klf4/c-Myc (OSKM)[13,14]. This highlights the very broad and general role of NuRD, especially given that it is the most abundant form of deacetylase in mammals.

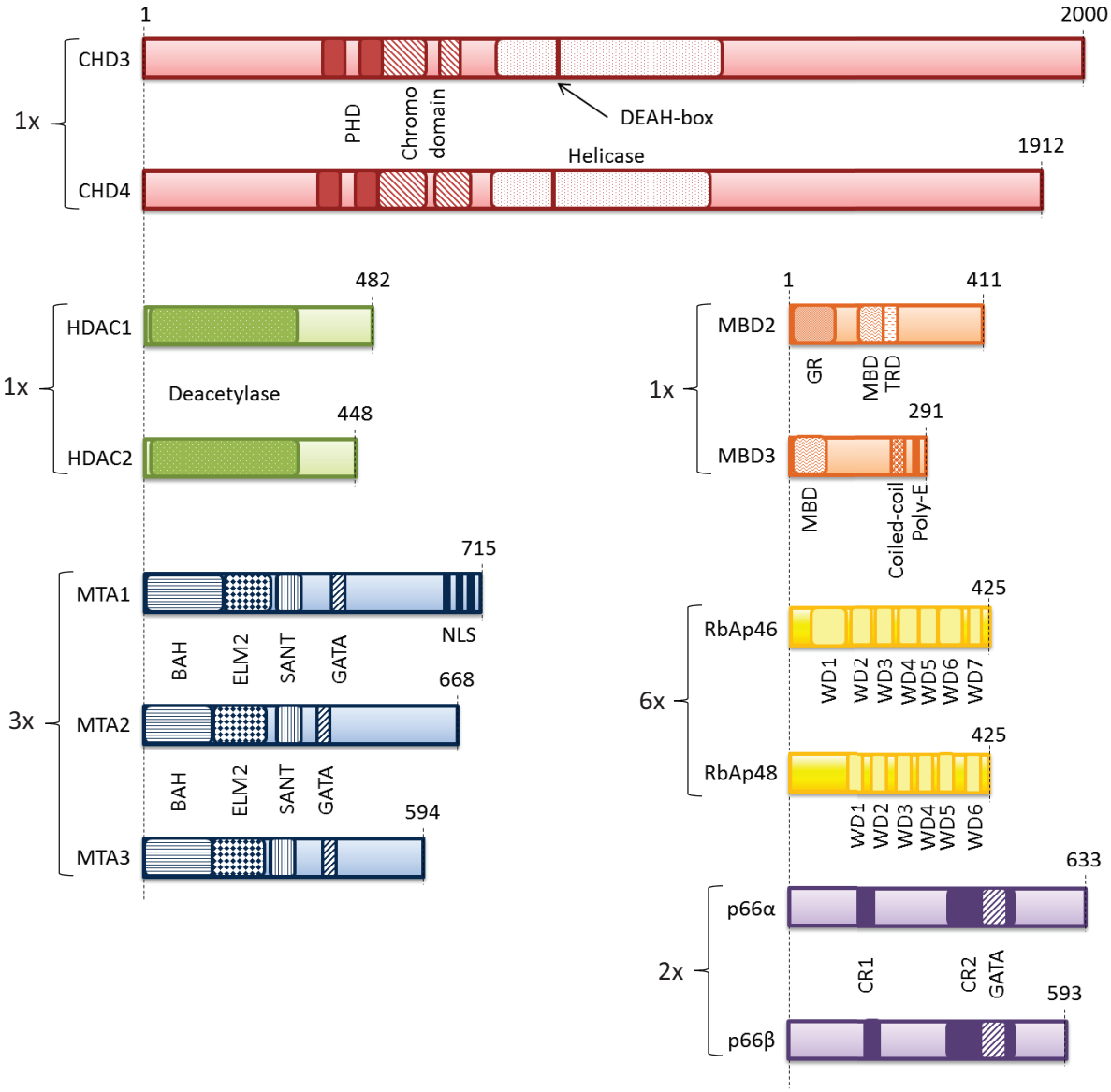


FIGURE 33

Schematic description of the NuRD subunits

The different NuRD subunits are schematically shown, with their main isoforms. Their size is indicated as the number of amino acids, and their characteristic domains are shown and annotated. The stoichiometry suggested by mass spectrometry analysis is indicated.

PHD: Plant HomeoDomain; Chromodomain: Chromatin Organization Modifier; DEAH-box: Asp-Glu-Ala-His-box; BAH: Bromo Adjacent Homology; ELM2: Egl-27 and MTA1 homology; SANT: Switching-defective protein 3, Adaptor 2, Nuclear receptor co-repressor, Transcription factor IIIB; NLS: Nuclear localization sequence; GR: Gly/Arg-rich region ; MBD: Methyl-CpG Binding Domain; TRD: Transcription Repression Domain; Poly-E: Poly-glutamate; WD : Trp/Asp-reich region; CR1: Conserved Region 1; CR2: Conserved Region 2.

(Uniprot - CHD3: Q12873; CHD4: Q14839; HDAC1: Q13547; HDAC2: Q92769; MTA1: Q13330; MTA2: O94776; MTA3: Q9BTC8; MBD2: Q9UBB5; MBD3: O95983; RbAp46: Q16576; RbAp48: Q09028; p66α: Q86YP4; p66β: Q8WXI9; DOC-1: O14519)

4.1 Detail of the components of the NuRD complex

4.1.1 CHD3/4: the ATP-dependent chromatin-remodelling

The ATPase activity within the NuRD complex in ensured by the Mi-2 subunits. This protein exists as two homologs: Mi-2α (or CHD3) and Mi-2β (or CHD4). The latter is the most abundant in the NuRD complex, although it seems that both proteins can coexist within the same complex. At least three molecular species can thus be found: Mi-2α/NuRD, Mi-2α/Mi-2β/NuRD and Mi-2β/NuRD. The Mi-2 protein was initially identified as an autoantigen in patients affected with dermatopolymyositis[463,464]. About one quarter of these patients are positives to anti-Mi-2 antibodies. While the correlation between those tumour developments and the presence of anti-Mi-2 antibodies hasn't been proven formally, in 20 to 25 % of the cases, the patients develop an ovarian, colorectal, lung, pancreatic, stomachic or lymphatic cancer[465,466].

CHD3 and CHD4 are large ATPases, with a molecular mass of about 220 kilo Daltons. Their domain organization comprises two conserved PHD fingers, two tandem chromodomains and a SWI2/SNF-like helicase domain[467]. They are thus part of a subclass of the Snf2 family[468] and are highly conserved among the yeast, plant and animal kingdoms, although absent in *Saccharomyces cerevisiae*. The activity of Mi-2 proteins from three different species (*Drosophila melanogaster; Xenopus laevis; Homo sapiens*) were shown to be stimulated by chromatin but not by free DNA or histones[458,469,470]. This implies that these enzymes are implicated in the recognition of the nucleosome rather than of its individual components. NMR solution structures of individual chromodomain and the two PHD domains have been determined, revealing a bivalent mode of binding to histone H3 tail[471,472]. Indeed, the two PHD domains of CHD4 are able to bind two distinct H3 tails, within a single nucleosome or on adjacent nucleosomes[473]. The post-translational modifications of those tails govern the binding affinity of CHD4: H3K9 trimethylation promotes the binding of the enzyme, while H3K4 methylation abolishes it[474].

Additionally, two isoforms of CHD3, CHD3.1 and CHD3.3, exhibit a C-terminal SUMO-interaction motif (SIM) allowing them to interact with the sumoylated form of the KRAB-associated protein-1 (KAP-1), a major component of heterochromatin. KAP-1 phosphorylation by the ataxia telangiectasia mutated protein (ATM), as observed in the case of DNA double strand breaks, inhibits this interaction with CHD3 and leads to chromatin decompaction[475-477].

The Mi-2 proteins have also shown their crucial role in the development of some model organisms. In *Caenorhabditis elegans*, both CHD3 and CHD4 are implicated in the Ras signalling pathway, regulating cell fate in the hermaphrodite vulva and male tail[478]. In *Arabidopsis thaliana*, the CHD3 homolog PICKLE is implicated in the auxin signalling pathway, required for lateral root initiation and development[479]. In human, both CHD3 and CHD4 interact with transcription factors Ikaros, Aiolos and Helios, and target NuRD to specific promoters involved in lymphocytic development and proliferation[5-7]. Among those genes, one could mention CD179b, for progenitor B cells to precursor B cells differentiation; dntt, required for the V-DJ recombination; or CD4 and CD8a, for thymocytes maturation. These data suggest that Mi-2 could also have an important role in

mammal development, but the lack of genetic models remains today a crucial bottleneck to further study these enzymes.

4.1.2 HDAC1/2: deacetylating histone lysines

The subunits ensuring histone deacetylation are HDAC1 and HDAC2. These 55 kilo Daltons proteins are highly conserved and ubiquitous in all Eukaryotes. They share 83 % of sequence identity, and their double knock-out in T-cells or embryonic stem cells leads to a decrease by half of the total deacetylase activity of these cells[480]. They are thus the two predominant enzymes in terms of histone deacetylation activity in mammalian cells. Though HDAC1 and HDAC2 don't exhibit any DNA-sequence specificity, it's been suggested that they could interact with coactivators and corepressors to target DNA in a more specific manner[481].

Sequence alignments of class I HDACs showed major differences in the C-terminal domain, which is entirely missing in HDAC8. This domain is required in HDAC1 and 2 to bind to partners in the context of protein complexes, and is furthermore post-translationally modified to regulate their catalytic activity[482]. Nevertheless, the first crystal structure of a HDAC, that of HDAC8 in complex with different inhibitors, paved the way for structural understanding of the class I HDACs[483,484]. These proteins are composed of a single α/β domain, consisting of an eight-stranded parallel β-sheet at the centre of thirteen α-helices. These secondary structures are connected through long loops, thus creating the catalytic core domain of these enzymes. The active site consists of a long tunnel with a mini-mum depth of 8 Å, also referred to as lipophilic tube leading to the catalytic machinery. This tunnel is occupied by the four carbons of the side chain of the acetylated lysine, stabilised by hydrophobic contacts with residues G151, F152, H180, F208, M274 and F306 (HDAC8 numbering). All these residues are con-served among the class I HDACs, with the exception of M274 being a leucine in all other class I HDACs. Finally, the end of the tunnel accommodates a zinc ion, chelated by five coordination bonds in a trigonal bipyramidal fashion, and stabilised by the carboxylic oxygen of residues D178 and D267, and by the Nδ1 atom of the H180 side-chain (*figure 34*). The carbonyl oxygen of the acetyl moiety carried by the acetylated lysine, as well as a water molecule, occupy the two other coordination sites. More recently, the structures of HDAC2 in complex with inhibitors[485,486], and the one of HDAC1 in complex with the ELM and SANT domains of MTA1[3] (described later in this manuscript) shows the same global structure of the core HDAC protein.

It has been observed that inhibitors of the hydroxymate class, in the manner of SAHA (Suberoylanilide hydroxamic acid) or trichostatin A, bind to the catalytic site in roughly the same way as acetylated lysines, with fast binding kinetics and nanomolar K_d range over a large majority of class I and II HDACs. This is explained by the direct access of the ligand through the lipophilic tube, chelating the zinc ion with its hydroxamic group. In contrast, inhibitors of the benzamide class, like Entinostat and Mocetinostat, are also located in the lipophilic tube, but their thiophene group is accommodated in a deeper pocket, named "foot pocket". This pocket is formed by flipping and shifting of the two M31 and L140 residues (HDAC2 numbering). Those two residues are conserved among HDAC1-HDAC3 but not HDAC8 and class II HDACs, giving rise to a higher specificity of this class of inhibitors. Finally, the central secondary amide moiety of these inhibitors chelates the zinc

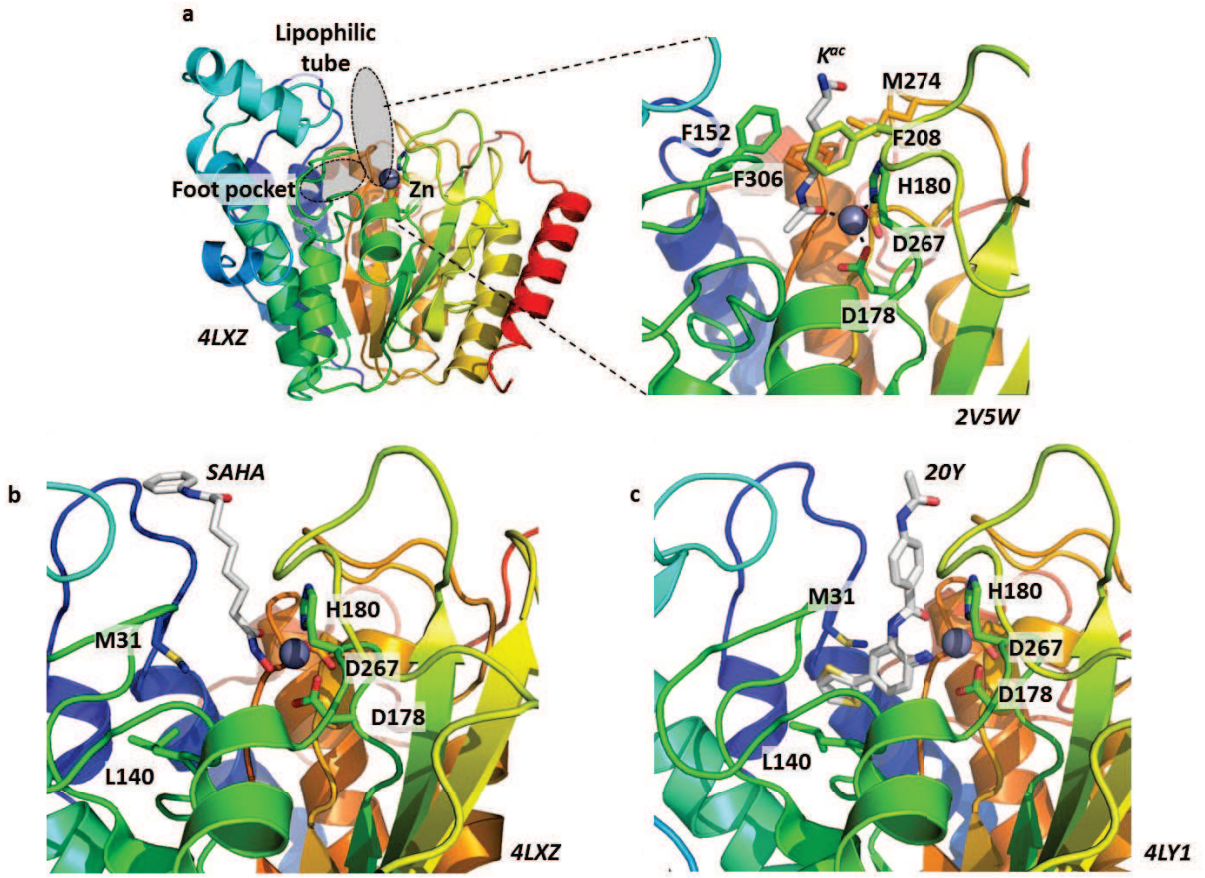


FIGURE 34

Crystal structure of class-I HDACs

a) The HDAC8 structure, in complex with an acetylated lysine, is shown. The lipophilic tube as well as the foot pocket are marked. They contain a zinc atom, shown as a purple sphere. The interfaced residues are shown (*pdb : 2v5w*)

b) The HDAC2 structure in complex with the SAHA inhibitor shows a similar recognition mode to that of its natural substrate, acetylated lysine (*pdb : 4lxz*)

c) The HDAC2 structure in complex with the 20Y inhibitor (4-acetylamino-N-2-amino-5-thiophen-2-ylphenylbenzamide) shows the foot pocket opening by flipping out of the M31 and L140 residues (*pdb : 4ly1*)

ion, locking the molecule in place. This explains the slower kinetics of benzamides, compared with hydroxymates, together with the higher specificity for class I HDACs, and in particular, HDAC1 and HDAC2.

The biochemical and genetic properties of HDAC1 and 2 make it difficult to understand their specific functions within the NuRD complex. Indeed, although deacetylation is largely associated with gene repression, knock-out experiments showed that several genes become repressed in the absence of HDAC 1 or HDAC2. This suggests that these two enzymes could also have a role in gene activation[487-489]. Further studies carried out by treating embryonic stem (ES) cells with trichostatin A showed both a decreased expression of pluripotency-related genes and an increase of lineage-

specific genes, indicating a negative as well as positive regulation activity. By chromatin immunoprecipitation (ChIP), it has been shown that these enzymes can localise at some transcriptionally active loci in human[490], mouse[491] and yeast[492], corresponding to DNase I hypersensitive sites. In particular, HDAC1 has been detected in promoter regions, on pluripotency genes in ES cells (like fgf4, mbd3, nanog, oct4, sox2, tbx3 or zfp42) and trophoblast-lineage genes in trophoblast stem cells (like bmpr1a, cdkn1c, cdx2, elf5, hand1, msx2 or tcfap2c)[491], while HDAC2 is present in both promoters and gene bodies.

A commonly observed phenomenon when knocking-out HDAC1 and HDAC2 is the decrease of cell proliferation[488,493-496]. The loss of these enzymes induces an overexpression of the kinases p21/WAF1/CIP1[493,496] and p57/Kip2[488] inhibitors, preventing G1/S phase transition. HDACs inhibitors have been tested in numerous cases of cancers, with the aim of limiting tumour growth[497], but most of these inhibitors, in the manner of SAHA (approved and commercialized under Vorinostat or Zolinza) are large-spectrum inhibitors of class I and II HDACs and therefore lead to significant side-effects. Studies carried out on mice showed that the use of specific HDAC1 and/or HDAC2 inhibitors, like benzamides described above are equally efficient with respect to antiproliferative effects, but with potentially reduced side-effects[494,498].

Given their biochemical and genetic identity, it is not surprising that HDAC1 and HDAC2 are redundant enzymes: knock-outs of these showed no deleterious phenotype, the remaining enzyme complementing the missing one[480,487-489,493,499-502]. Why this redundancy exists remains to be addressed.

4.1.3 MTA1/2/3: reading histone tails and promoters

MTA proteins were the last ones to be characterised within the NuRD complex. The first representative in this family, temporarily called p70, then MTA1, was isolated after the observation of its differential expression pattern observed by cDNA library screening using the 13762NF rat mammary adenocarcinoma metastatic system[503]. But despite the overexpression of this protein, one had to wait for the discovery of NuRD and the presence of MTA proteins in this complex to start understanding the role of this family[1,459].

Phylogenetic studies suggested that the mta gene underwent duplications to lead to the three loci found in vertebrates (mta1 on chromosome 14q, mta2 on chromosome 11q and mta3 on chromosome 2q), mta2 being the nearest relative to the ancestral non-vertebrate gene[504]. Those three genes encode the three proteins MTA1, MTA2 and MTA3, and also three alternative-splicing products: MTA1S, MTA1-ZG29p and MTA3L[505].

The three canonical MTA proteins have a molecular weight of 80, 70 and 65 kilo Daltons, respectively, and share 68 % of sequence homology between MTA1 and MTA2 and 73 % between MTA1 and MTA3. This strong homology is especially due to the N-terminal domains, the C-terminal parts being more variable. With the exception of MTA1-ZG29p, all the MTA proteins possess various highly conserved domains: a bromo adjacent homology domain (BAH; 70 % of identity between

MTA1^{BAH} and MTA2^{BAH} and 76 % of identity between MTA1^{BAH} and MTA3^{BAH}), an Egl-27 and MTA1 homology domain (ELM; 76 % of identity between MTA1^{ELM} and MTA2^{ELM} and 78 % of identity between MTA1^{ELM} and MTA3^{ELM}) and a SANT domain (87 % of identity between MTA1^{SANT} and MTA2^{SANT} and 94 % of identity between MTA1^{SANT} and MTA3^{SANT}). The role of these domains has not been fully studied yet in the context of MTA proteins. Nevertheless, some functional insights come from related proteins. For example, the SANT domains in Ada2 and SMRT seem to interact primarily with unmodified histone tails[433,434,506], and BAH of Rsc2 is implicated in histone H3 binding[507], while ORC1^{BAH} recognises H4K20me2[508]. The short MTA1S isoform is produced by alternative splicing inside a cryptic site of exon 14, resulting in a shift in the reading frame, involving the addition of 33 new amino acids. The C-terminal end of this isoform is thus unique within the MTA family, and shows no sequence homology with other proteins within the GenBank[509]. Finally, isoform MTA1-ZG29p is a product of the mta1 gene, including only the seven last exons. For this reason, it doesn't exhibit the three domains described previously, and its location seems to be restricted to zymogenic granules in the pancreas[510].

Expression regulation for the mta genes is still little-known to date, however, preliminary results are available. For example, heregulin, a growth factor which binds to the human epidermal growth factor receptors 3 and 4 (HER3 and HER4) transmembrane receptors, is able to induce MTA1 expression in breast cancer cells[11]. It has also been shown that the c-Myc proto oncogene could bind directly to the mta1 gene to activate its expression[511]. Moreover, MTA1 is overexpressed in hypoxia, and is responsible for hypoxia inducible factor 1 (HIF-1) stabilisation by deacetylation, becoming then resistant to degradation by the 26S proteasome[512].

Additionally, MTA proteins are intimately linked to the oestrogen receptor ER[10,513], in breast cancer and mammary gland development[514]. For example, the short isoform MTA1S directly interacts with ER and is responsible for its sequestration in the cytoplasm[509]. MTA1 also blocks ER-driven gene activation, by antagonising the effect of oestradiol[11], while MTA2 can make breast cancer cells insensitive to oestrogens and tamoxifen, by deacetylation of ER itself[10]. Finally, the promoter of mta3 is directly activated by ER-α[515-517], thanks to the presence of a half response-element ERE, and MTA3 seems to be involved in repression of some genes involved in invasive growth, like Snail[515] or Wnt4[518]. Consequently, MTA1 and MTA3 seem to have an opposite role. Expression patterns of those two proteins support this idea: MTA3 is largely expressed in healthy epithelial cells, and its expression decreases along with tumour growth, until complete shutdown at the carcinoma stage; on the contrary, MTA1 is gradually expressed, concomitantly with tumorigenesis.

Recently, a first 3-Å-resolution crystal structure of HDAC1 in complex with MTA1 has been published[3,519] (*figure 35*). It shows the ELM and SANT domains of MTA1 (residues 162-335, i.e., one quarter of the protein), wrapping around HDAC1, with an interaction interface of 5185 Å² surface area. Three regions can be distinguished: the first one corresponds to the SANT domain of MTA1, composed of three α-helices (H1 to H3). The interface with HDAC1 forms a positively charged pocket, which can accommodate an inositol tetraphosphate molecule (Ins[1,4,5,6]P4) to stabilise

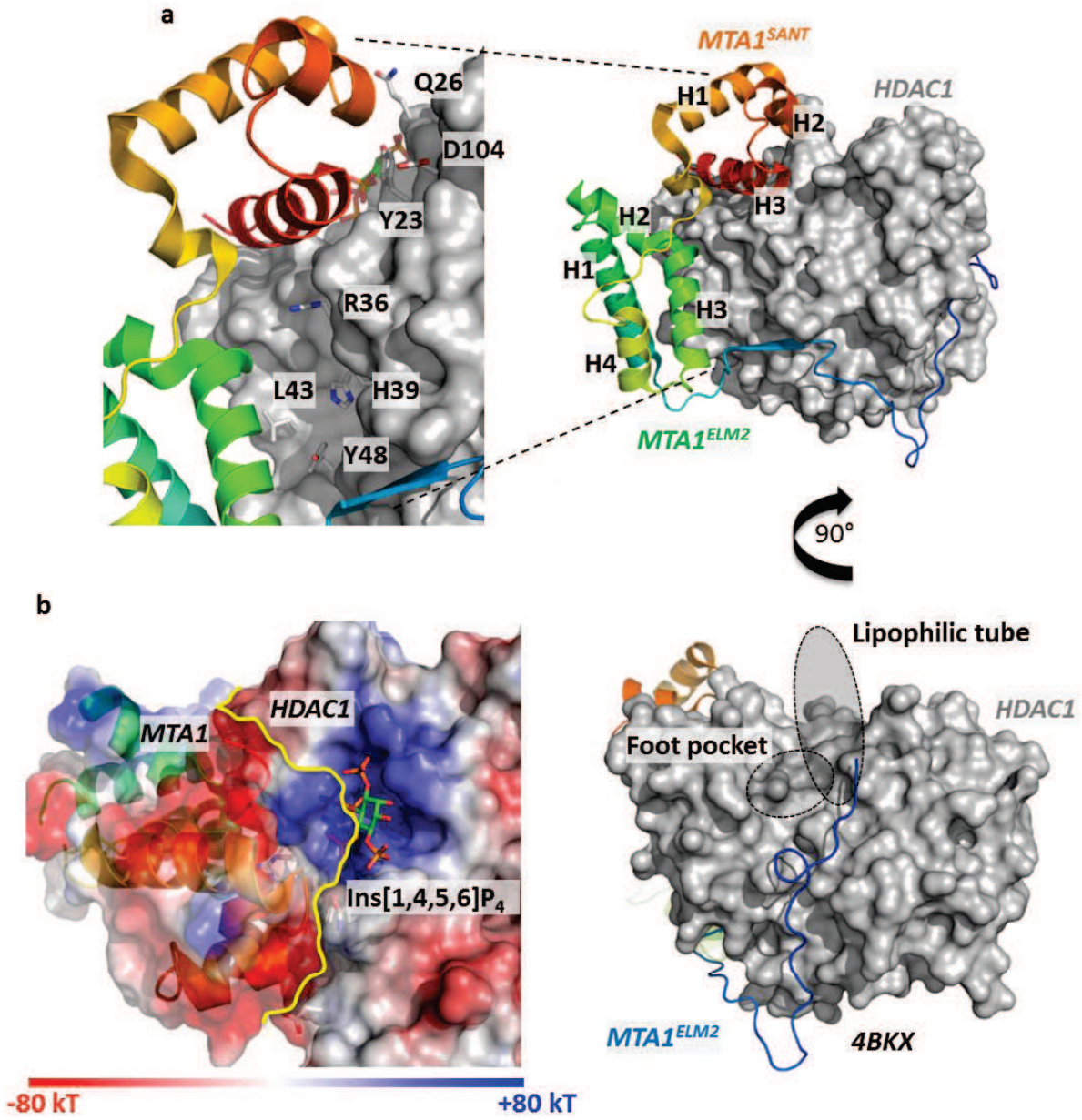


FIGURE 35

Crystal structure of HDAC1 in complex with MTA1

a) The HDAC1 structure (in grey) in complex with the SANT and ELM2 domains of MTA1 is shown. One can notice how MTA1 is wrapped around HDAC1. Crucial interface residues are marked.

b) This model shows how an inositol tetraphosphate molecule can accommodate inside the basic pocket (coloured in blue) at the interface between HDAC1 and MTA1. In yellow, the interface between both proteins is shown. This model was made by superimposing two X-ray structures: the HDAC3-SMRT complex containing an IP4 molecule (*pdb: 4a69*) and the HDAC1-MTA1 complex (*pbd: 4bkx*). Electrostatic potentials are shown in red (negative), white (neutral) and blue (positive).

this highly basic interaction, through residues K31, R270 and R306, among others. This observation had previously been made on a HDAC3-SMRT^{SANT} complex, copurified from mammalian cells[520]. Further studies showed that mutations of the MTA1^{SANT} residues involved in coordination of Ins[1,4,5,6]P4 lead to a reduced interaction between the SANT domain and HDAC1. However, MTA1 can still be tethered to HDAC1 in absence of Ins[1,4,5,6]P4, through interaction of the ELM domain as described later. Studies on the HDAC3-SMRT showed a link between ageing of the complex, loss of Ins[1,4,5,6]P4 moiety and decreased HDAC activity. However, addition of exogenous Ins[1,4,5,6]P4 recovered the HDAC activity with level higher than endogenous complexes. Similarly, the same observation has been made on the HDAC1-MTA1 complex, with an activation K_d around 5 µM. These elements tend to confirm Ins[1,4,5,6]P4 as having a regulatory role of class I HDACs *in vivo*.

The second region correspond to three-quarters of the C-terminal region of the ELM domain, folded with four α helices (H1 to H4). The isolated ELM domain shows no folded secondary structure in circular dichroism, implying a radical structural reorganization upon binding to HDAC1[3]. Helices H1 and H3 mediate the interaction interface with HDAC1 (1278 Å²). Simultaneously, this region is responsible for dimerisation of two MTA1 proteins, mediated by interactions between helices H1 and H4, and to a lesser extent, H2, of the two MTA1 molecules. Up to twenty-eight apolar residues (fourteen for each monomer) are involved in this dimerisation, with an important interaction interface of 2332 Å². This is a rather clear confirmation that this dimerisation interface is physiologically relevant, and that in terms of stoichiometry, the NuRD complex probably contains two MTA proteins, as well as two HDAC proteins. Finally, a third region corresponds to the N-terminal part of the ELM domain. It comprises a specific and conserved motif (EIRVGxxYQAxI), and forms an extended loop conformation. This long thirty-amino-acid-chain runs on the surface of HDAC1, inside a long apolar groove.

4.1.4 MBD2/3: DNA-binding and the connexion to methylation

The smallest subunit in the NuRD complex is represented by a MBP[481,521]. With a mass of approximately 43 and 33 kilo Daltons respectively, MBD2 and MBD3 are the two exclusive yet interchangeable MBPs within NuRD[522]. While MBD2 binds to methylated DNA[297], MBD3 has lost this ability in mammals. Indeed, the appearance of this class was accompanied by two point mutations in the mbd3 gene, leading to the incorporation of two new amino acids in positions 30 and 34 (a histidine and a phenylalanine, instead of a lysine and a tyrosine, respectively). This abolishes the selectivity of this protein for methylated DNA[523-525]. While the very first studies fifteen years ago credited MBD2 with only a transient role in the complex, being in particular a NuRD recruiter to methylated DNA before its eviction and replacement by MBD3[481,521], other studies since have shed light on a MBD2/NuRD complex, biochemically and functionally distinct from the MBD3/NuRD complex[522,526]. In that sense, MBD2 knock-out experiments showed only little effects at the phenotype level, whereas MBD3 knock-out leads to embryonic lethality[319].

Recently, it has been proposed that MBD3 and, to a lesser extent, MBD2, were able to specifically bind to hydroxymethylated CpG islands. Notably, MBD3 seems to colocalise with TET1

(ten-eleven translocation methylcytosine dioxygenase 1), the protein responsible for hydroxylation of methylcytosines[527]. Additional experiments however failed to show an interaction between MBD3 and hydroxymethylated DNA[528]. Instead, MBD2 and MBD3 appear to be preferentially localised at CpG-rich transcription start sites (TSS). At TSS's, MBD2 predominantly binds methylated CpG islands, leading to a repression of gene expression; whereas MBD3 binds to non-methylated DNA, and is associated with active transcription[17,529]. Recently, NMR spectroscopic dynamic analysis suggested that MBD3 could have a counterbalancing role, binding in a competitive manner to non-methylated CpG islands, avoiding thus an abusive repression of those active genes by MBD2[530].

Several X-ray and NMR structures of MBDs in complex with DNA have been solved, revealing a common interaction pattern for all the MBPs[310,530-535] (*figure 22, 36*). In particular, two solution structures of MBD2 and one solution structure of MBD3 have been solved, highlighting a quasi-structural identity between the two[530,534,536]. The MBD is characterised by an α/β sandwich, composed of an N-terminal four-stranded antiparallel β-sheet (β1: residues 6-8 in MBD3; β2: residues 15-20; β3: residues 32-37; β4: residues 41-43), and a C-terminal α-helix (residues 47-53). This α-helix is kept antiparallel against the β4 strand by hydrophobic contacts. Furthermore, the MBD exhibits three loops L1, L2 and C-terminal hairpin. L2 connects the α-helix and the C-terminal hairpin and is well defined in solution. In contrast, the long L1 loop between β2 and β3, composed of a dozen of residues, is more flexible. This appears to be a necessary prerequisite for binding to DNA. Indeed, seven residues of this loop make contacts with one of the DNA strand, at the level of the major groove. The other DNA strand interacts mainly with residues in the α-helix and L2-loop.

Recognition of a CpG island is independent for each methylcytosine, as guessed by the absence of symmetry in the interaction domain. Arginines 22 and 44, which are conserved among all MBPs, interact with symmetrically arranged guanines inside a CpG island. The guanidinium group of R22 makes a hydrogen bond with the O6 and N7 atoms of the guanine base of the first DNA strand; while the guanidinium group of R44 exhibits the same interaction pattern with the guanine base of the second DNA strand. Both arginines lie in a plane with their interacting guanines, stabilised and locked by direct hydrogen bonding of residue D32 and water-mediated hydrogen bonding of residue Y34. This flat orientation allows the two arginine residues to pack against the methylated cytosine bases neighbouring their interacting guanines, and permitting weak van der Waals interactions. Finally, the carbonyl group of R44 forms a weak CO-HC hydrogen bond with the methyl group of the cytosine base on the second DNA strand; while Y34 forms a water-mediated hydrogen bond to recognise the methylated cytosine of the first DNA strand. The integrity of those residues is crucial to ensure the binding to methylated DNA, as proven by mutagenesis experiments. In particular, Y34 turned out to be a key-residue in the recognition of the methylation state. Its mutation into a phenylalanine, as found in mammals, leads to a loss of affinity of methylated CpG islands. On the contrary, *Xenopus laevis* MBD3 doesn't exhibit this evolutionary mutation, and is thus still able to bind to methylated DNA. Also, the crystal structures of MBD4MBD in complex with different modified DNA show that Y96 (Y34 in xMBD3, T34 in m/hMBD3) is flipped out of the DNA interface, and is only making water-mediated hydrogen bonds with the phosphate backbone of the first DNA

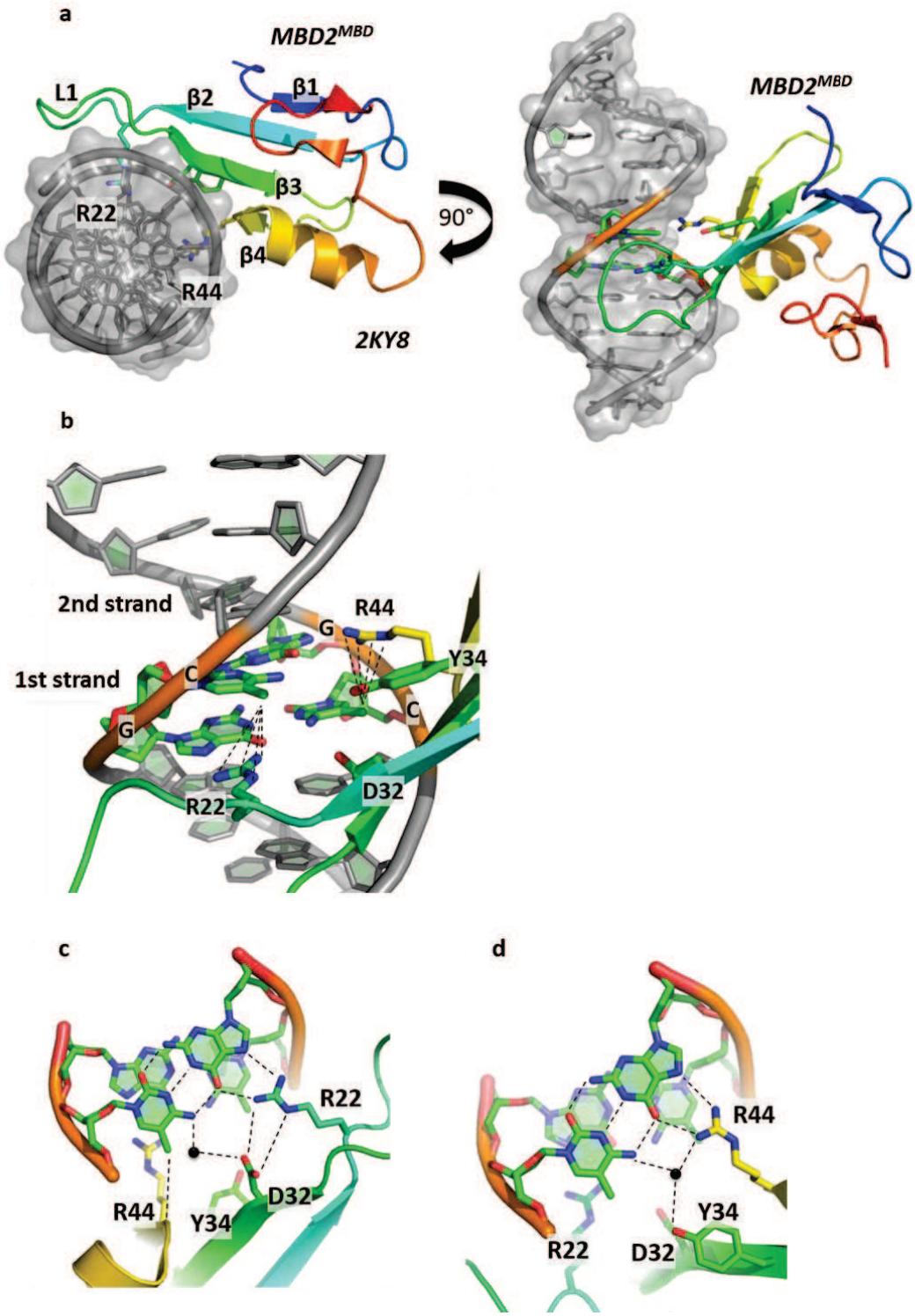


FIGURE 36

Recognition mode of methylated DNA by MBD2 and xMBD3

a) The MBD domain of MBD2, in complex with a methylated CpG island is shown.

b) A detailed view of the binding interface shows the crucial residues involved in the interaction, as well as the van der Waals forces (black dotted lines) between arginine residues and the methylated cytosines.

c) The C-G base-pairs and their specific hydrogen bonds with MBD2 are shown. Water molecules involved in indirect hydrogen bonding are shown as black spheres (*pdb: 2ky8*).

strand. This leads to a loss of specificity of MBD4 towards methylated DNA, at the cost of an increased binding of 5mCG/TG and 5mCG/hmCG islands[533].

Recently, the central role of MBD3 in somatic cell reprogramming and cellular differentiation was suggested[537]. Researchers at the Weizmann Institute indeed showed that, when knocking-out the mbd3 gene by RNA interference, the reprogramming process of mouse and human somatic cells into pluripotent stem cells was highly facilitated, with success rate ranging to 100 % in some cases, compared with 0.01-5 % in general, and with an average speed significantly improved, since reprogramming only requires a week compared with more than a month with classical methods. This discovery is a follow-up to the highlighted interaction of MBD3 with OSKM proteins[13,14] (Oct4, Sox2, Klf4 and Myc), transcriptions factors responsible for maintaining totipotent state until blastocyst stage. MBD3, when binding these proteins, could thus recruit the NuRD complex to totipotency genes to repress their expression; or act independently and induce a conformational change of these proteins, preventing their binding to the DNA. These results corroborate the embryonic lethality observed in MBD3 knock-outs, the embryo being then unable to properly differentiate. However, opposite data obtained out of two different reprogramming systems suggest a context-dependent role of MBD3 in reprogramming, albeit further studies will be needed to confirm this theory[538].

In a completely different context, the role of MeCP2 in the Rett syndrome, as mentioned earlier, led researchers to seek for mutations in other MBPs that could also be linked to neurologic disorders. In this respect, the DNA of 226 Caucasian and Afro-American autistic patients and their relatives was thus analysed and alterations were found in mbd1-4 genes in 198 of them[539]. Interestingly, one of those alterations was found in exon 1 of the mbd3 gene. It corresponds to a point mutation (G>T at 1,543,563 in locus 19p13.3), leading to the incorporation of a new amino acid inside the MBD domain (R23M). This mutation, inducing the loss of a positive charge, has been observed in two Afro-American half-brothers, displaying late and unfunctional language acquisition. This mutation seems to be inherited from their disease carrier maternal grandmother, suggesting a sex-related effect. This residue is semi-conserved in MBD2 were it correspond to K167. Though published structures haven't shown any relevant role of this arginine in DNA binding, it is located right after the crucial R22 residue binding the CpG island.

4.1.5 RbAp46/48: ensuring a stable platform and binding histones

RbAp46 and RbAp48 (also called Rbbp7 and Rbbp4, respectively), were first identified because of their interaction with the tumour suppressor factor Rb[540-542]. Later, studies showed their affinity for histones, and their presence in various deacetylation and remodelling complexes[504,543-545].

Although those two proteins share 90% of sequence identity[541], they exhibit different biochemical activities. Thus, RbAp46 associates with other proteins, notably histone acetyltransferase 1 (HAT1), involved in *de novo* histone H4 acetylation, on its lysine 5 and 12 residues[462,546]. This acetylation pattern is conserved among all Eukaryotes, from yeast to human[547]; whereas RbAp48 is an essential chaperone for histone H3-H4 tetramer deposition on newly

replicated DNA[548], and is especially found in the assembly complex CAF-1 (Chromatin assembly factor 1), with p150/CHAF1A and p60/CHAF1B. Nevertheless, RbAp46 and RbAp48 can be jointly found inside complexes, for example in association with HDAC1 and/or HDAC2, within the Sin3A or NuRD complexes, where they promote gene repression, including the one regulated by Rb[481,542,549] ; they are also found within the Polycomb repressive complex (PRC2 and PRC3), with the histone-lysine N-methyltransferase EZH2, to methylate H3K27 or H1K26[550] ; or in the nucleosome remodelling factor (NURF) complex, along with ISWI (SNF2L in human), where RbAp proteins are called NURF55[551].

RbAp46/48 are 48-kilo Dalton proteins that share a WD40 repeat sequence. The great stability of those proteins allowed to date to solve seven crystal structures: two RbAp46 structures in complex with a histone H4 peptide, at 2.4 and 2.6 Å resolution[15] ; a structure of RbAp48 alone, at 2.3 Å resolution[552] ; a structure of RbAp48 in complex with a FOG-1 (Friend of GATA) peptide at 1.9 Å resolution[412] ; and three structures of RbAp48 in complex with an MTA1 peptide at 2.5 and 2.15 Å resolution, respectively[553]. Predictably, the RbAp proteins showed a structure similar to other WD40 proteins: a donut-shaped seven-bladed β-propeller, with a long N-terminal α-helix (residues 9 to 28), lying on the seventh blade of the barrel, and a short C-terminal α-helix (residues 405 to 409), which is placed above and seems to extend the N-terminal helix. Finally, one particularity of these WD40 proteins is the presence of a seventeen residues loop, negatively charged, inside the sixth blade of the barrel, called PP loop (because of two successive prolines P362 and P363) (*figure 37*).

The crystal structure of RbAp46 in complex with a small histone H4 peptide shows an interaction interface of approximately 700 Å². This H4 peptide corresponds to residues 25 to 42 of the human isoform, i.e., the first α-helix of the histone fold and a part of the N-terminal tail. Though the structures previously described in other WD40 proteins highlightened an interaction interface on the front of the barrel, or even sometimes, inside it, histone H4 preferentially binds in a unique pocket located on the side of the barrel, and formed by the PP loop and the long N-terminal helix. Thus, hydrophobic residues I34, L37 and A38 in helix α1 of histone H4 interact with a hydrophobic patch composed of residues F29, L30, F367, I368 and I407 of RbAp46. A complex network of salt bridges and hydrogen bonds is also described between Q27, K31, R35, R36, R39 and R40 of histone H4; and E356, D357, D360, G361, P362, L365, N406, I407 and D410 of RbAp46[15]. All these residues are conserved in RbAp48 and the yeast homolog p55, suggesting that the binding mechanism of these three proteins with histone H4 is similar. Finally, it has been shown that, in order to pro-mote a proper interaction with RbAp46, the α1 helix of histone H4 must partially open, abolishing interactions with the α2 helix as well as those with histone H3, in particular through residues I34, L37 and A38. This observation raises thus the question of whether RbAp proteins interact with the nucleosome inside protein complexes, or they suggest an accrued flexibility of the nucleosome[15]. Recently, pulsed electron-electron double-resonance (PELDOR) experiments have shown that RbAp48 can interact with a H3-H4 dimer, but not with a $(H3\text{-}H4)_2$ tetramer[554].

The structure of RbAp48 in complex with the fifteen N-terminal amino acids of the GATA-1 cofactor FOG-1, involved in erythroid and megakaryocytic cell differentiation, shows a binding interface located on the face of the barrel, which extends into the central channel[412]. This interaction

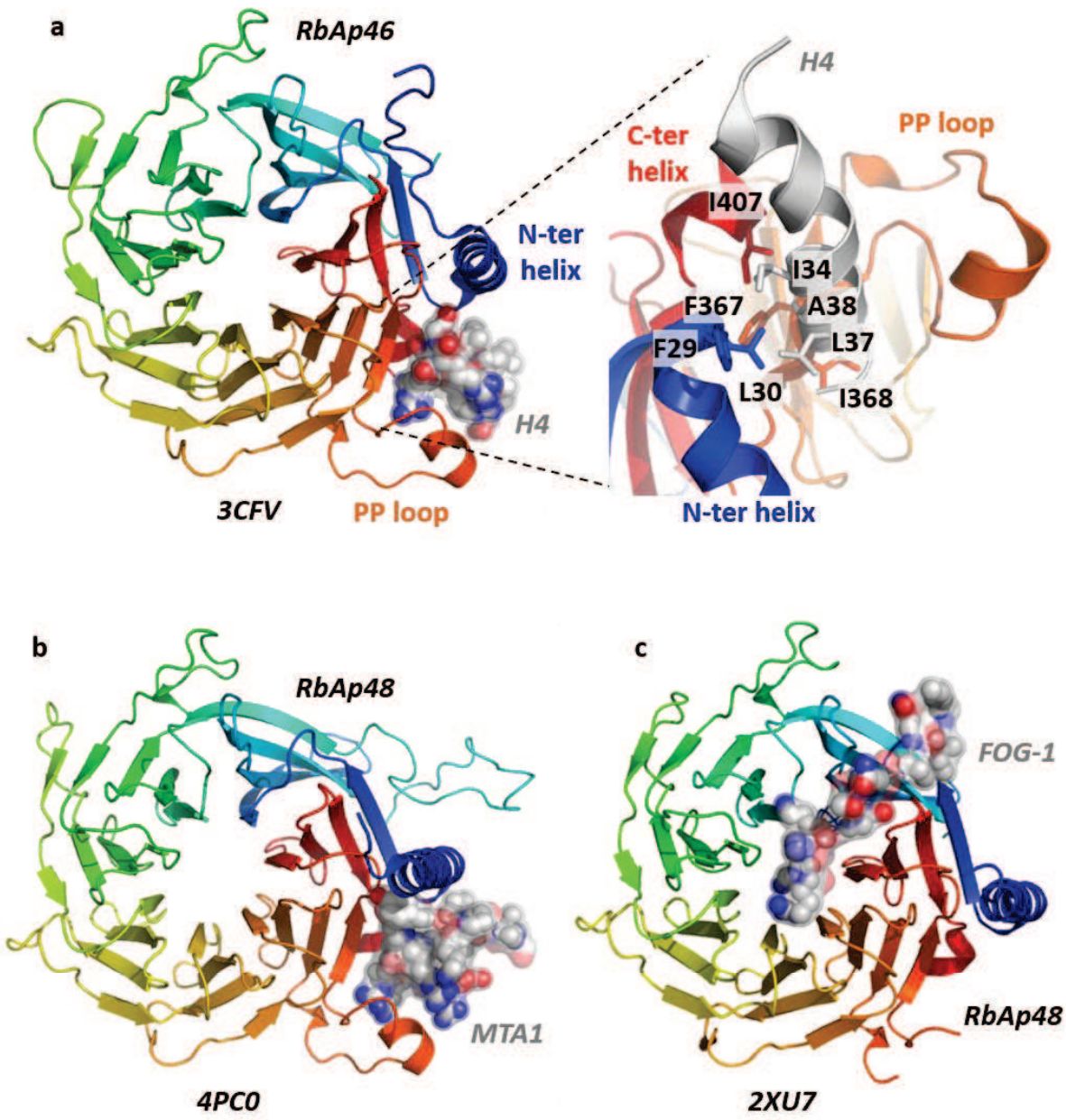


FIGURE 37

X-ray structures of RbAp46 and RbAp48 in complex

a) RbAp46 in complex with a peptide of histone H4 N-terminal tail shows a binding interface located on the side of the barrel, within a pocket formed by the PP loop and the long N-ter helix. Crucial hydrophobic residues are shown (*pdb: 3cfv*).

b) RbAp48 in complex with a peptide of MTA1 C-terminal end shows a similar binding interface than that with histone H4 (*pdb: 4pc0*).

c) RbAp48 in complex with a peptide of FOG-1 shows a different binding interface, located on the face of the barrel and extending inside the central channel (*pdb: 2xu7*).

is different from that observed in the RbAp46/H4 complex, and is highly specific, because eight out of the thirteen residues in FOG-1 are involved in hydrogen or ionic bonds with RbAp48. In particular, this interface is composed of numerous acidic residues of RbAp48 (E231, E319, E179, E126, E395, E41), allowing the stabilisation of a basic triade of FOG-1 (R3, R4 and K5). This interaction pattern can be extrapolated to RbAp46, as it shares those same conserved residues.

Finally, the RbAp48 structure, in complex with a short peptide of the C-terminal end of MTA1, shows a very similar binding site to the one observed in the complex with H4[553]. This suggests that RbAp46/48 cannot simultaneously interact with MTA1 and histones.

Misregulations of RbAp46 and RbAp48 seem to be linked to tumorigenesis in several localisations, among which mammary and cervical tissues[555-557]. They indeed were shown to directly interact with the nuclear receptor ERα, and to affect ERα-regulated-gene expression[558]. For example, siRNA silencing experiments in MCF-7 cells were carried out, and gene activity was recorded for the progesterone receptor (PR) and pS2 genes (activated by oestradiol), and the cyclin G2, Sox9 and ERα genes (repressed by oestradiol). Interestingly, RbAp46 was shown to enhance the activation of oestradiol-repressed genes and attenuates the activation of oestradiol-activated genes, even in presence of oestradiol; and RbAp48 ensures the basal repression of oestradiol-repressed genes in absence of hormone. Furthermore, a prolonged oestradiol exposure of those cancer cells leads to a two to three-fold increase of RbAp46 levels. Together, these data suggest that RbAp46 could be a mediator favouring a continue ERα activity, while RbAp48 could ensure the basal repression of these genes in the absence of a ligand. Previous studies corroborate this idea, showing that repression of RbAp48 is involved in cervical cancer formation[559] ; and that an increase of RbAp46 levels prevents breast cancer development[555,556,560]. RbAp48 therefore appears to be a key therapeutic target for cervical cancer treatment[561]. Indeed, it has been shown that RbAp48 expression is favoured by radiotherapy irradiations, and that SiHa, HeLa and Caski cells were radiosensitive, the more the level of RbAp48 is high. Thus, adenovirus-induced overexpression of RbAp48, combines with radiotherapy, show a convincing antiproliferative effect in athymic mice.

In another context, a recent study carried out on eight human beings, aged 33 to 88, showed a differential expression pattern of RbAp48 in their brain, reduced along with the age, specifically in the dentate gyrus, a subregion of the hippocampus known for its lifelong neurogenesis, and foreseen to be the seat for episodic memory[562]. Additional studies carried out on mice confirmed the role of RbAp48 in the memorization process. A young knock-out mouse has indeed less potential in memorizing new objects and environments; on the contrary, lentivirus-induced re-expression of RbAp48 in old mice helped increase their cognitive capacities. Those phenomena seem to be closely related to the activity of the RbAp48-binding partner CREB-binding protein(CBP)/p300, nuclear receptor-bound transcription coactivators increasing gene expression through their intrinsic histone H4 and H2B acetyltransferase activities.

4.1.6 GATAD2A/B: potentialising repression

In 2002, a two-hybrid screening on MBD2 highlighted in 2002 the interaction of two proteins, baptized p66α and p66β, and later, GATAD2A and GATAD2B, respectively (GATA Zinc Finger Domain Containing 2A/B)[563]. These two proteins, initially thought to be two isoforms of the same genes, appear to derive from an ancestral gene duplication, undergone with the emergence of the mammal class. Indeed, a unique orthologs, named p66, is found in *Drosophila melanogaster*, *Caenorhabditis elegans* and *Xenopus laevis*[521,564]. The human gene p66α could be localised on the chromosome 19p13.11, while the p66β gene is localised on the chromosome 1q23.1.

These proteins have shown to interact and colocalise with MBD2 and MBD3[563]. Later, functional assays showed that GATAD2A/B were recruited through two domains: on one hand, to MBD2, via their CR1 domain; on the other hand, to DNA and deacetylated histones, via their GATA zinc finger-like CR2 domain[564]. Moreover, the overexpression of both p66 proteins induces an increase of repressive action by MBD2; whereas p66 knock-outs allow a partial recovery of MBD2-repressed genes[565].

GATAD2A/B can be targeted by post-translational modifications, like sumoylation. Thus, residues K30 and K487 of p66α, and K33 of p66β, when sumoylated, enhance the interaction of these proteins with other partners within the NuRD complex, like HDAC1 or RbAp46[566].

4.1.7 DOC-1: the overlooked tumour-suppressor

Recently, copurification experiments carried out on recombinant MBD2 and MBD3-expressing stable cell lines revealed the presence of a new 12 kilo Daltons subunit called CDK2AP1 (Cdk2-associated protein 1) or DOC-1 (Deleted in oral cancer-1) inside both NuRD/MBD2 and NuRD/MBD3 complexes[522]. As its name suggests, this protein, a putative tumour suppressor interacting with CDK2, is inhibited in oral and colorectal cancers[567,568]. Later, mass spectrometry experiments confirmed the presence of this protein in NuRD[2,569].

The role of DOC-1 is still unclear; nevertheless, it was shown that overexpression in 293T cells would lead to a partial arrest of the cell cycle phase G1/S, together with a significant growth retardation[570]. This can be offset against the consequences of MBD2 overexpression promoting cell proliferation. This suggests thus that an opposite role for those two proteins exists inside the NuRD complex.

4.2 NuRD functions: history and current believes

When the NuRD complex was discovered in the late 1990's, the knowledge available at that time regarding the role of chromatin remodelling led researchers to define this complex as a general transcriptional repressor[571]. Moreover, its recruitment was supposedly driven by protein-DNA interactions (in particular through MBD2/3) or protein-protein interactions (with transcriptional corepressors).

In the following years, and due to a lack of genetic models to study it, the role of the NuRD complex was described primarily based on expression patterns and isolated-subunits data. In that sense, the MTA1 subunit was for example known to be overexpressed in some cases of breast cancer. The NuRD complex was thus appointed to the position of transcriptional regulator in breast tumour cells[11,572]. These studies also showed that MTA1 expression was enhanced in ERBB2+/HER2+ cells, and that it interacts directly with ER to repress ER-dependent transcription, in particular that of the BRCA1 gene, causing invasive cell growth. This was the very first evident confirmation of the direct-recruitment driven repressing role of NuRD[11].

This first study was followed by numerous of others, each awarding NuRD a repressor role in specific cellular processes: MTA3, responsible for Snail repression, to prevent invasive growth in breast cancer[515]; CHD4, via its interaction with NAB2, to corepress EGR transactivators (Early Growth Response) which are responsible for prostate cancer progression[573]; MBD3, which interacts with the unphosphorylated oncoprotein c-JUN, to repress its transcriptional activity in the context of rectal cancer[574]; MTA3, to regulate the outcome of B-cells, via direct interactions with BCL6[9]; CHD4 to inhibit the activation of the mb-1 promoter by EBF and Pax5 in lymphoblasts[575]; MTA1 and MTA2, which interact with BCL11b, to repress the HIV-1 LTR expression in infected T-cells[576,577]; MBD2, in partnership with GATAD2A, by direct recruitment of NuRD to methylated CpG islands to repress the embryonic and fetal β-globin gene expression[536]; etc. In parallel, early biochemical experiments with histone peptides showed an enrichment of NuRD in the histone H3 tails region, but this interaction is inhibited by H3K4 methylation, an epigenetic mark associated with transcriptional activity[578-580]. All these examples taken together, and others, strengthened the position adopted then. The NuRD complex was a transcriptional repressor, involved in a variety of signalling pathways, and covering a very broad panel of biological contexts.

Around 2004, the first genetic models were developed. The CHD4 protein was the first target of these experiments[581]. A conditional knock-out mouse was designed to study the consequences of CHD4 invalidation in T-cells, demonstrating the importance of the NuRD complex in several stages of the cell development. The next very interesting discovery was the role of NuRD in activating the CD4 gene. It was then the first statement of an activating role of NuRD, until then considered as a repressor only. A few years later, the same team showed how important CHD4 and the whole NuRD complex are for maintenance of pluripotency in hematopoietic stem cells. Silencing of CHD4 in those cells would lead to a differentiation into erythroblasts, at the expense of lymphoblastic and myeloblastic cell lines. The expression patterns in those cells, after CHD4 invalidation, highlighted an equivalent number of abnormally repressed and abnormally activated genes [582]. NuRD had then definitively lost its general repressor title.

In the past five years, NuRD was shown to be active in other cell processes than transcription, like DNA damage response, or assembly and maintenance of chromosomal structures. For example, silencing of the NuRD complex leads to hypersensitivity and accumulation of DNA damages. NuRD seems to be recruited to damages sites, either by CHD4 interaction with PARP1

(Poly ADP-ribose polymerase 1) [583], or by CHD4 interaction with RNF8 (E3 ubiquitin-protein ligase), itself recruited by MDC1 (Mediator of DNA damage checkpoint) interacting with the histone variant γH2AX [370,584,585].

In the unique case of fast proliferating T-lymphocytes, strong accumulations of NuRD have come to light, called NuRD foci, and localised in hypermethylated pericentromeric heterochromatin on chromosomes 1, 9 and 16 during the S phase of the cell cycle[586]. These foci replace the Polycomb PRC1 foci, observed in every other cell types, suggesting thus a unique role of NuRD in lymphocytes proliferation. It is plausible that NuRD be recruited at these hypermethylated sites through MBD2, to ensure chromatin assembly during and/or after replication. No similar mechanism has yet been described in fast proliferating tumour cells.

This review on the different functions of the NuRD complex and its subunits is not intended to draw up an exhaustive list of all the processes in which NuRD is involved. First of all because this list is constantly evolving, but most of all because we lack the needed hindsight to fully understand the importance of this complex. This is clearly illustrated by the lack of unity between the different functions outlined above. Of course, it seems rather undeniable that the NuRD complex plays a quasi-ubiquitous role in our cells; however, we still lack biochemical, genetic and structural data to understand the precise role of a given NuRD complex, *in vitro* but also *in vivo*, in its environment. Also, more than 15 years after its discovery, the composition of this complex is still unclear. Which isoforms make up this complex, and in which stoichiometry? Are these different NuRD complexes systematically found in every cell types or are they specific to a given cell type, development stage, pathology, etc.?

From a functional point of view, it remains unclear today why evolution chose to assign two enzymatic activities within a single complex. Indeed, even though HDACs have proved their capacities to activate a subset of genes, these subunits still persist in being considered as general repressors, which raises the question of the apparent contradiction with the ATP-dependent remodelling activity of CHD3 and CHD4, known to allow breathing of the chromatin and thus, potentially activate gene expression. A long date proposal suggests that ATP-dependent remodelling of the chromatin is a prerequisite to allow other subunits of the NuRD complex, in particular HDACs, to access their substrate. However, this has never been clearly proven, and further experiments will be needed to confirm the mechanism underlying the function of NuRD.

Finally, from a structural point of view, it is intriguing how so many different proteins can interact with a complex. Structural studies of the whole NuRD complex will be needed to address the accessibility of factors to this macromolecular complex, and determine the molecular basis of inter-protein interactions, such as with factors involved in cancer progression. Furthermore, relatively little is known about the intramolecular interactions within the entire NuRD complex, as illustrated by the remaining open question of the stoichiometry. Some works nevertheless constitute the blueprint for a better comprehension of the NuRD architecture, in the manner of the HDAC1/MTA1 complex or RbAp46/H4 complex structures. This indeed suggests the presence of two MTA and two HDACs subunits within the complex, as well as potentially two RbAp46/48 per nucleosome. Whether

the latter work in synergy with CHD3/4 to destabilise histone octamer, as suggested by the binding of RbAp46 to H3-H4 dimer only, is also a remaining question to be answered.

RESEARCH QUESTIONS AND OBJECTIVES

The study of the chromatin remodelling NuRD complex is a project initiated when I arrived in the laboratory lead by Bruno KLAHOLZ, in collaboration with Ali HAMICHE.

NuRD, standing for “Nucleosome Remodelling and histone Deacetylation”, is, to date, one of the two known complexes coupling two independent chromatin-remodelling activities. This highly conserved complex forms a large macromolecular assembly that consists of different protein subunits: an ATPase for ATP-dependent chromatin remodelling (CHD3 or CHD4) and histone deacetylases (HDAC1 or HDAC2); but also auxiliary proteins that stabilise and target the complex, and regulate its activity: the histone chaperones RbAp46 and RbAp48, the DNA-binding proteins MBD2 or MBD3 and the histone-binding proteins MTA1, MTA2 and MTA3; this large number of homologs and isoforms of each subunit leads thus to a horde of coexisting NuRD complexes, depending on the cellular, tissue, physiological or pathological context. The aim of this project was thus to study, from a structural point of view, the organization of this multi-subunit complex.

When I started to work on this project in 2010, only few data were published. From a functional point of view, the NuRD complex still had a lot to reveal about itself; as for the structural aspect, only few structures were available, like the NMR solution structures of the chromodomain and one of the PHD domain of CHD4, and the MBD domain of MBD1 and MeCP2; and the crystal structure of RbAp46. Our goal was to lift the veil on several important aspects, in particular through structural analysis of the isolated NuRD subunits for which no structure was known; the stable subcomplexes of NuRD, composed of two, three or four subunits; and finally, the whole NuRD complex. Two different approaches are taken in this manuscript: on one hand, the production and purification of recombinant proteins for the study of isolated subunits or subcomplexes by X-ray crystallography and/or cryo-EM, and on the other hand, the purification of the endogenous NuRD complex, produced in human cells, for its study by cryo-EM.

My work during my PhD has focused on the implementation of this ambitious project, starting with the cloning of all the subunits of the complex, design of the protocols for production and purification and biophysical and structural study of several isolated subunits. Having obtained a master degree in biological and organic chemistry, I thus had the opportunity to train myself to a very wide range of methods and to run a full project.

The NuRD complex binds to the nucleosome. This structure, composed of DNA and histone proteins, is the basic unit of DNA compaction in cells, and above all, the target of a vast majority of compaction regulation processes, making it a partner of choice for structural studies. Several biochemical and functional published data show that CHD3/4, HDAC1/2, MBD2/3 and RbAp46/48 are able to directly interact with nucleosomes or their components: DNA on one hand, and histones on the other hand. The latter can furthermore be post-translationally modified by covalent and reversible modifications like acetylation, methylation, phosphorylation, etc. As described later in this manuscript, most of the studies carried out on NuRD subunits were done in complex with home-made reconstituted nucleosomes. This choice allowed a more or less significant stabilisation of the studied proteins, which is favourable for structural studies, including in some cases, stabilisation of the nucleosome itself, as observed by cryo-EM.

During these four years of PhD, I focused on three subunits: MBD3, RbAp46 and RbAp48. These three proteins are indeed of particular interest within the NuRD complex. The first one is with little doubt the most interesting of all: MBD3 indeed remains very little studied and poorly understood. It belongs to the MBD family (Methylated CpG-Binding Domain), binding methylated DNA. But due to a mammalian-specific point mutation (Y34F), MBD3 turns out to be an exception in this family and shows a loss of specificity towards methylated DNA. This mutation is however not observed in *Xenopus laevis*, suggesting a redundant role of the two paralogues xMBD2 and xMBD3; in *Drosophila melanogaster*, only one protein, namely dMBD2/3, exists. It is thus a peculiar protein, perfectly illustrating an example of recent evolution and gain of function. The RbAp46 and RbAp48 proteins, for their part, are both histone chaperones, found within the NuRD complex but also other chromatin-associated complexes such as HAT-1, CAF1, Sin3A, Polycomb, EZH2/EED and NURF. Their chaperoning role is however poorly studied and these subunits are most often described as a stable structural platform within these complexes. Their study in complex with the nucleosome would allow shedding light on new information about their function, especially with regards to the molecular basis of nucleosome recognition. Plus, this type of approach is more likely to provide functionally relevant insights, in contrast to publications in 2008 and 2013, suggesting that RbAp proteins can only bind free H4 histone, but not nucleosomes. It was however only assumptions based on an X-ray structure of RbAp46 in complex with a small peptide of histone H4 and not the entire H4 protein or nucleosome.

The limited amount of biochemical data and the instability of the MBD3 protein made this work long and tedious, as it will be illustrated in the next chapter. Furthermore, a structural study requires high-quality samples, in large quantities, as well as a fine biophysical analysis prior to conducting the structural study itself. Within a multiprotein complex like NuRD, the subunits are stabilised through their intra-complex interactions, but also with other partners such as DNA, histones or other protein factors. Taken out of their natural context, the isolated subunits can thus show some significant instability, making their study more complicated. This case will in particular be illustrated with MBD3, which exhibits many partners in addition to DNA, and for which very few biochemical data are available. *In vitro*, this protein doesn't fold properly and hence becomes extremely instable as described in this work. Despite of designing sample preparation optimisation, the purified protein is very quickly destabilised by minor temperature changes, requiring working constantly below 4°C.

MATERIAL AND METHODS

The study of the NuRD complex, carried out during my PhD, used an integrated approach combining different fields in molecular biology, biochemistry, biophysics and structural biology (*figure 38*). In this chapter, I describe in detail the techniques applied in this work.

The first step was the isolation of the human genes of the different subunits of the NuRD complex and to clone them in expression vectors for recombinant protein overexpression and purification. Two systems were used during this study: the bacterial *Escherichia coli* and baculovirus-infected insect cells. Particular attention was paid to three proteins of the NuRD complex: MBD3, RbAp46 and RbAp48. The expression and purification was standardised for each of these proteins, using, in particular, liquid chromatography, based on various physicochemical properties of proteins: affinity, ion exchange, size exclusion. Thereafter, biophysical methods, including mass spectrometry, light scattering, analytical ultracentrifugation or differential scanning fluorimetry were used to characterise the samples and verify their quality, especially the size, mass, stability and aggregation state.

In parallel, nucleosome particles were reconstituted *in vitro*, using purified histones and DNA, following the protocol established by Karolin Luger in 1999[45,587]. The genes of the four canonical histones in humans (or humanized in the case of histone H4) were cloned in bacterial expression vector and expressed in *E. coli*. The proteins were then purified separately, in denaturing conditions, and assembled while refolding. In parallel, optimised DNA fragments of 145 to 147 bp were amplified and purified. Altogether, the four histones and the DNA were mixed to reconstitute functional and high quality nucleosome particles.

The structural study of nucleosomes in complex with subunits of the NuRD complex requires a high level of purity and stability of the samples. A major part of my work consisted of standardising the expression and purification of the three proteins MBD3, RbAp46 and RbAp48. In the case of MBD3, in particular, the lack of preliminary studies together with its highly unstable behaviour, required putting in place strict and rigorous working conditions, thereby making this study difficult and time-consuming. Finally, the stability of the nucleosome-protein complexes also had to be optimised to get stable and homogenous samples. The two major structural techniques, X-ray crystallography and cryo-EM were used for this study, because of their complementarity and their suitability for macromolecular complexes studies.

X-ray diffraction data were collected at synchrotron sources (SLS, Villigen, Switzerland and Diamond, Oxfordshire, England). Electron micrographs were, for their part, collected on the in-house transmission electron microscopes Tecnai F30 Polara and Titan Krios (FEI).

Among these different techniques, I should stress out the implication of:

- the molecular biology service of the IGBMC which carried out clonings of MBD3 mutants F34Y and R23M during my last year of PhD, from plasmids and oligos I supplied;
- the proteomic platforms of IGBMC and IBMC which carried out all mass spectrometry experiments;
- the baculovirus service which took care of virus generation from bacmids I designed and supplied, as well as cell infection and cultures. Harvesting and expression tests were carried out by myself.

- Dr. Kareem Mohideen Abdul, with the help of Isabelle Hazemann, Sinthuja Peiris as well as mine, for recombinant histone production and purification, DNA production and purification and *in vitro* nucleosome reconstitution.
- Dr. Bruno Klaholz and Dr. Kareem Mohideen Abdul, as well as myself, for X-ray data collection at the SLS and Diamond synchrotrons. Dr. Kareem Mohideen Abdul has also been involved in X-ray data analysis.
- Dr. Jean-François Ménétret and Dr. Alexander Myasnikov who froze the cryo-EM grids and collected data on our in-house F30 Polara and Titan Krios microscopes. Jean-François Ménétret has also been involved in cryo-EM data analysis and reconstruction of the first MBD3-Nucleosome density map.

The remaining work, including molecular biology, bacterial expression, purification, optimisation processes, complex formation, biophysical assays, crystallisation and data collection at the SLS has been achieved by myself.

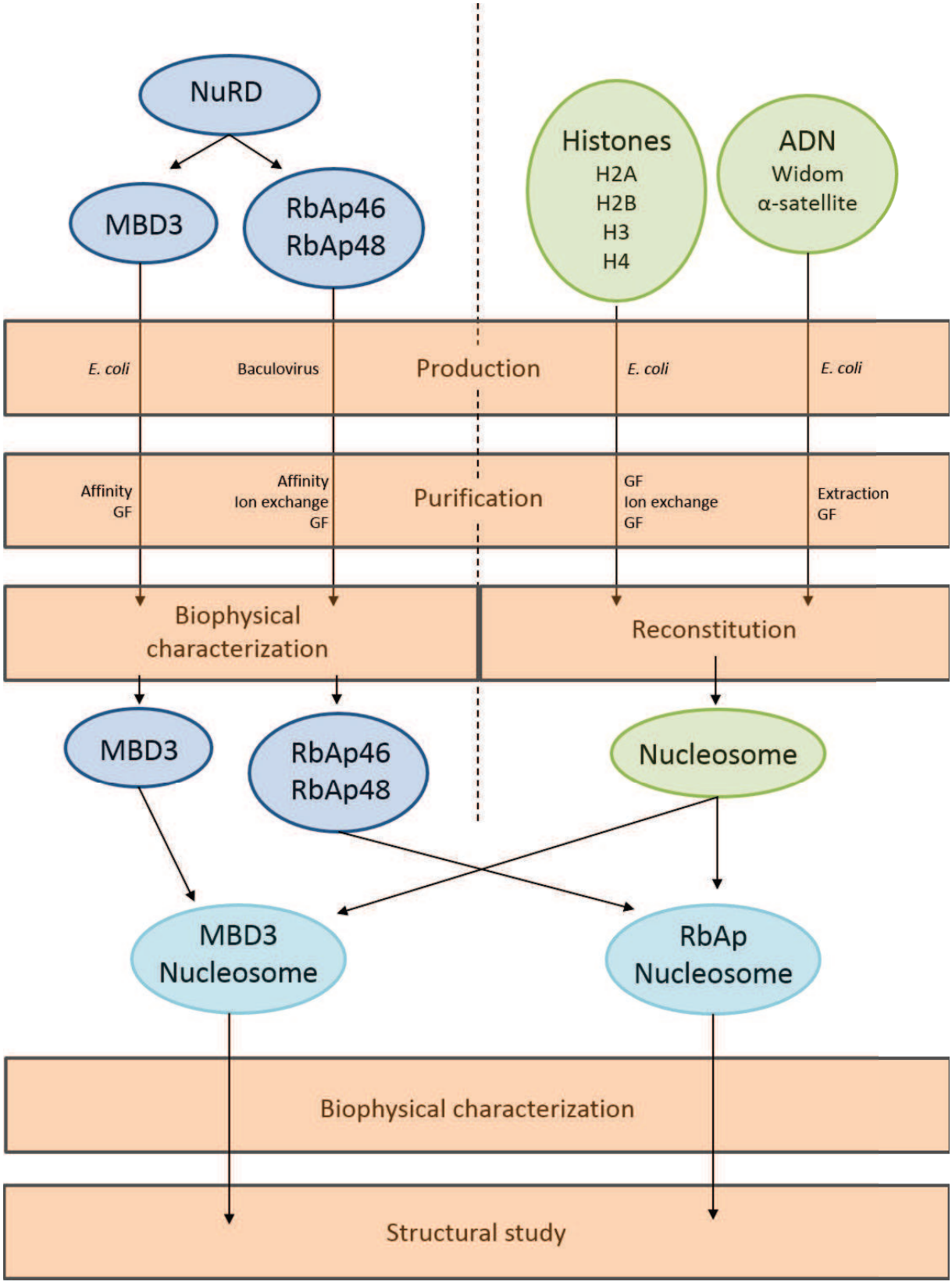


FIGURE 38
Workflow

1. Molecular biology methods

In order to produce a protein of interest in a given expression system, its gene must first be cloned into an expression vector. Once this recombinant vector is designed, it is incorporated into the expression system, usually a cultured cell strain that is able to transcribe the gene of interest and translate it into a protein. Several expression systems are commonly used in the laboratory, such as the bacteria *Escherichia coli*, the yeast strains *Saccharomyces cerevisiae* or *Pichia pastoris*, the lepidopterian cells *Spodoptera frugiperda* and *Trichoplusia ni* infected with a baculovirus, and a wide range of mammalian cell lines.

1.1 Cloning

1.1.1 The vectors

The typical cloning vector used in bacterial system is a plasmid. It is a double-stranded circular DNA molecule, with a size ranging from 1- over 1000 kbp. Due to the presence of a replication origin (*oriV*), plasmids have a particular property to replicate independently of the genomic DNA in the bacteria (*figure 39*).

Though plasmids were described for the first time in 1952 by Joshua Lederberg[588], it was not until 1977 that the first artificial plasmid was designed, from a natural plasmid[589,590], named pBR322. A 4.3 kb plasmid is the first representative of the second generation. However, this generation of plasmids offer, relatively limited if any capacities of selection, by insertional inactivation of the gene of interest within the antibiotic resistance gene.

The pUC family of plasmids (*plasmid « University of California »*) marks the starts of the third generation[591], that is still widely used in laboratories. These plasmids were designed to solve the selection issue encountered with pBR322. Vieira and Messing have created a system combining a negative selection through the antibiotic resistance with a positive selection, by gathering most of the restriction sites within a cloning cassette itself inserted within the *lacZ* gene, coding for β-galactosidase. This enzyme is able to degrade a lactose analogue, 5-bromo-4-chloro-3-indolyl-β-D-galactoside (XGal) into 5-bromo-4-chloro-3-indolyl, a blue product. When a gene of interest is cloned in the plasmid, the lacZ gene is inactivated and XGal degradation is thus not possible anymore, leading to a loss of the blue coloration of the colonies. Furthermore, being smaller than pBR322, the pUC plasmids (2.7 kb) allow insertion of longer DNA fragments. But because of the absence of regulatory components for expression, these plasmids are commonly used for DNA amplification only, and not recombinant protein production.

To produce recombinant proteins, expression plasmids were developed. Like the pUC plasmids, they were designed after pBR322 and are a part of the third generation. Besides the already described elements, these plasmids also include a strong promoter, a ribosome binding site (RBS), a stop codon and a transcription termination site. The pET plasmids (Novagen), for example, include the promoter of the bacteriophage T7 gene 10, specific for the T7 RNA polymerase. For this reason, bacterial strain with the genotype λDE3(lacI, lacUV5-T7 gene 1, ind1, sam7, nin5) is required: as it allows the expression of the T7 RNA polymerase after induction with IPTG of the $lacI^q$ operon.

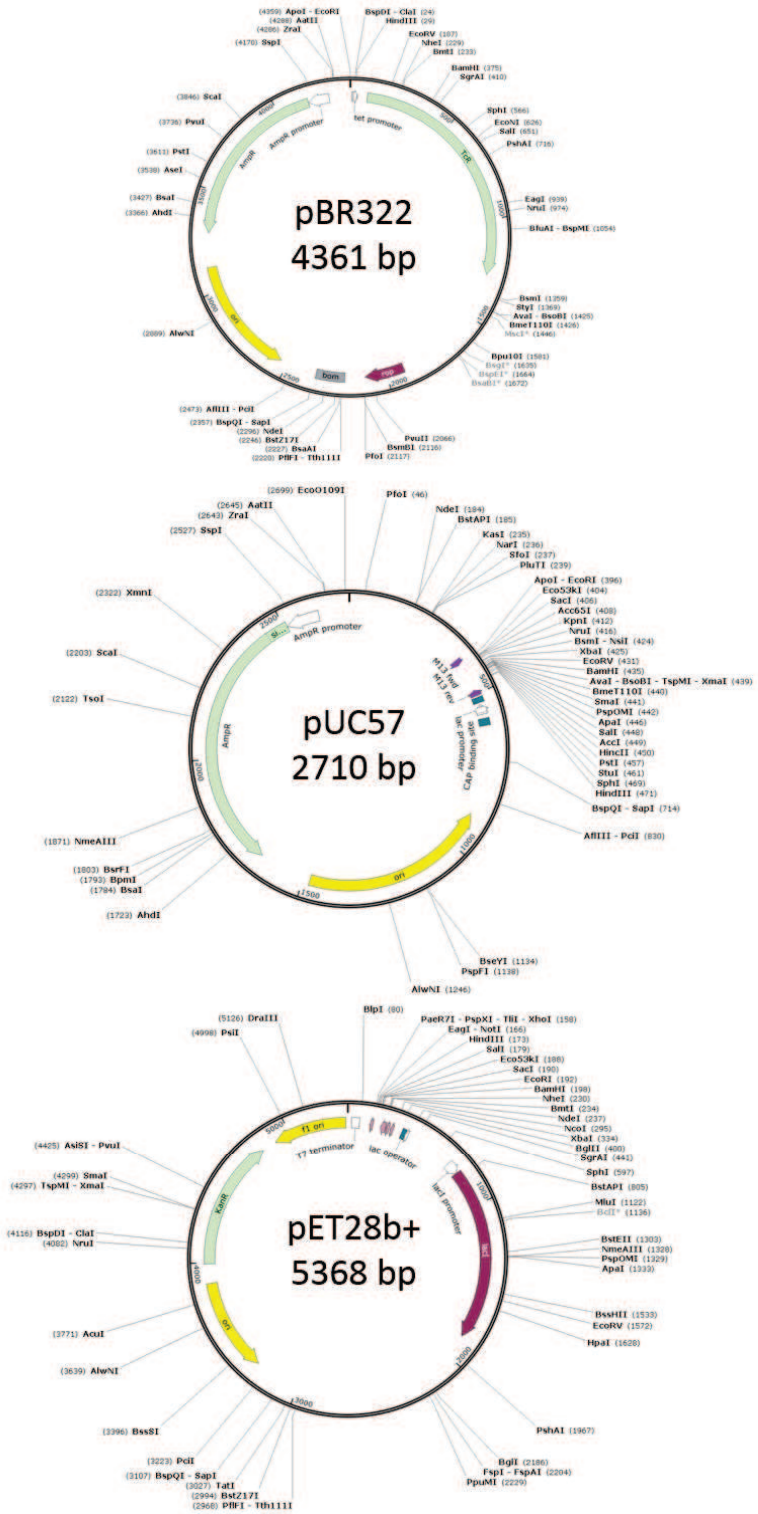


FIGURE 39

Bacterial cloning vectors

The maps of the second-generation plasmids (here shown pBR322) and third-generation plasmids (here shown pUC57 and pET28b+) highlights the evolution of these vectors. One can note in particular the appearance of a cloning cassette in the third generation, which contains a large number of restriction sites within a limited area, as well as the lactose operon, which allows regulation of recombinant gene expression.

pET plasmids are, by far, the most commonly used for recombinant protein production in bacterial system.

Finally, the fourth generation of plasmids was developed recently for the Gateway® system (Invitrogen), based on directional and conservative recombinational cloning[592,593]. These plasmids contain recombination sites, or "attachment sites" named *att*: the recombination of an attB site with an attP site gives rise to an attL site and an attR site, and vice versa. In order to make this reaction directional, the sequence of the att sites has been slightly modified, into att1 and att2. These sites recombine specifically to each other, in such a way that attB1 reacts only with attP1, to form attL1 and attR1. Gateway plasmids are further categorised into "donor vectors" which contain attP sites to recombine the gene of interest, flanked with attB sequences, to form "entry vectors" with attL sites. Further different are the "destination vectors" containing attR sites, which recombine the gene of interest included in the entry vector with its attL sites, to form the final expression vector with attB sites. This technology also allows multiple cloning, by using specific *att3*, *att4*, *att5 sites*, etc.

For the baculovirus system, the expression vector used is the bMON14272 bacmid. It is a large synthetic plasmid of around 136 kb, containing the genome of a baculovirus: Autographa californica multicapsid nucleopolyhedrovirus (AcMNPV). When the insect cells are infected by these bacmids, *in vivo*, these viruses express their genetic material and multiply to form virions, between 24 and 72h after the infection. The virion formation is dependent, in particular, on the expression of two proteins, polyhedrin and p10, which together represent up to 50% of the total protein production in the late phase. The baculovirus system takes advantage of the strong promoters of these two proteins, the pPolh and p10 promoters, effective only during *in vivo* infection. The gene of interest is thus transposed over the polyhedrin or p10 genes, to ensure a massive production of recombinant protein in the late phase of infection.

1.1.2 The cloning and mutagenesis techniques

The gene of interest can be isolated either from a complementary DNA (cDNA) bank, from genomic DNA (gDNA) or even be synthesised *in vitro*. The latter option offers the added advantage of optimising the codon bias in accordance with the organism that is chosen for protein expression. The codon bias is the preferential usage of some codons over others, because of the differential expression of some tRNAs and it is specific for each organism.

In practice, a vector can be designed using three different methods: restriction-ligation, which allows to subclone a DNA fragment, like a gene of interest, by using restriction sites; the Gateway® technology (Invitrogen); or the Bac-to-Bac technology, devised by the baculovirus system. I mainly used restriction-ligation and the Bac-to-Bac technology.

Besides, deletion mutations or point mutations can easily be designed from a vector already containing the wild-type gene of interest, with the SLIC method (Sequence and Ligation Independent Cloning).

1.1.2.1 Restriction-Ligation

During this work, all the vectors were designed using the restriction-ligation method, based on the use of restriction enzymes, bacterial endonucleases that possess the ability to cut a double-

stranded DNA, in a sequence-specific manner (typically, 6 to 8 nucleotides). Two types of restriction enzymes have been characterised based on the end products: “blunt ends” and “sticky ends”. The former ones cut the two strands of the DNA at the same level in the sequence, whereas the latter ones cut each strand around four base-pairs from each other, creating 5’ overhangs of unpaired nucleotides.

Nine enzymes were mainly used during this work: XhoI (from *Xanthomonas holcicola*), NotI (from *Nocardia otitidiscaviarum*), HindIII (from *Haemophilus influenza*), BamHI (from *Bacillus amyloliquefaciens H*), BglI (from *Bacillus globigii*), KpnI (from *Klebsiella pneumoniae OK8*), EcoRI (from *Escherichia coli RY13*), EcoRV (from *Escherichia coli J62 pLG74*) and HinfI (from *Haemophilus influenzae Rf*) (*figure 40*).

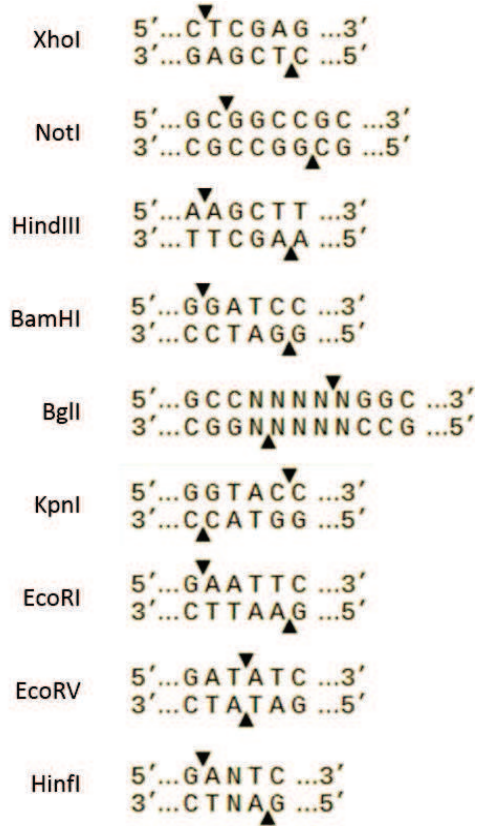


FIGURE 40

Restriction sites

Nine restriction sites were largely used during this work. Their specific cleavage sequence is shown here. Besides EcoRV, all these sites have sticky ends.

In order to insert a DNA fragment in a plasmid, like a gene of interest or any other nucleotide sequence, both must be treated with the same restriction enzyme. As mentioned earlier, the third generation plasmids have a cloning cassette inserted downstream of the promoter, containing a multitude of unique restriction sites that can be used to insert gene of interest. The gene itself has to be amplified by PCR with DNA oligos to add restriction sites, at each of its extremities.

Three important parameters have to be taken into account: the direction and the orientation of the insert in the plasmid, and the reading frame. Concerning the direction, it is a question of cloning the insert in such a way that it's coding coincides with the coding strand of the plasmid, downstream from the promoter. The use of an enzyme producing "sticky ends" is thus necessary to solve this first obstacle. With regards to the orientation of the insert, it should be inserted from 5' to 3'. Previously, this was achieved by using only one restriction enzyme, dephosphorylating the digestion products, and sequencing the vectors after ligation. This way, a majority of the products would have inserted the DNA fragment in the good orientation. But this solution is time-consuming and was replaced by using two restrictions enzyme that allows achieving 100 % of correct orientation products. Finally, the apt reading frame is essential to produce an mRNA that is functional and encodes the recombinant protein. Adding one or two nucleotides downstream of the restriction sites in the DNA insert is thus sometimes required to overcome this problem.

Ligation of both digestion products – the DNA insert and the linearized plasmid – involves formation of a phosphodiester bond between a 5' phosphate and a 3' hydroxyl. Two enzymes can catalyse this reaction: the DNA ligase from *E.coli*, and that of the T4 bacteriophage. The latter is commonly used, due to its higher efficiency.

1.1.2.2 Bac-to-Bac

In 1993, researchers from Monsanto designed a fast and powerful tool to generate recombinant baculovirus[594]. This new technology is based on the site-specific transposition of an expression cassette containing the gene of interest, into a bacmid, carried by the bacterium *E. coli* DH10Bac. This autoreplicative bacmid bMON14272 contains, among others, a kanamycine resistance gene, the lacZ gene from the pUC plasmid, and the attachment sequence "mini-*att*Tn7", inserted within the lacZ gene.

A pFastBac plasmid is used as donor vector. It contains a "mini-Tn7" element, with the Tn7R and Tn7L sites at each extremity, including a gentamycin resistance gene, a cloning cassette to insert the gene of interest by restriction-ligation and the pPolh or p10 promoter. After transformation of *E. coli* DH10Bac with the recombinant pFastBac vector, the mini-Tn7 element is transposed onto the "mini-*att*Tn7" attachment site in the bacmid, by *trans*-complementation. The lacZ gene is thus inactivated and XGal degradation is not possible anymore, leading to a loss of the blue coloration of the bacterial colonies. This transposition is assisted by the transposase encoded by the pMON7124 "helper" plasmid, also carried by the DH10Bac bacterium (*figure 41*).

This method offers the advantage of selecting bacterial colonies and purifying a single clone for the recombinant bacmid. Until a few years ago, the transposition was directly ensured in the insect cell, by co-transfection with the wild-type AcMNPV virus and a baculoviral transfer vector containing the gene of interest. A mix of wild-type and recombinant virus was thus obtained, and 4 to 6 weeks were necessary to isolate the recombinant baculovirus. The Bac-to-Bac technology gets rid of this time-consuming step and shortened this period to 7-10 days.

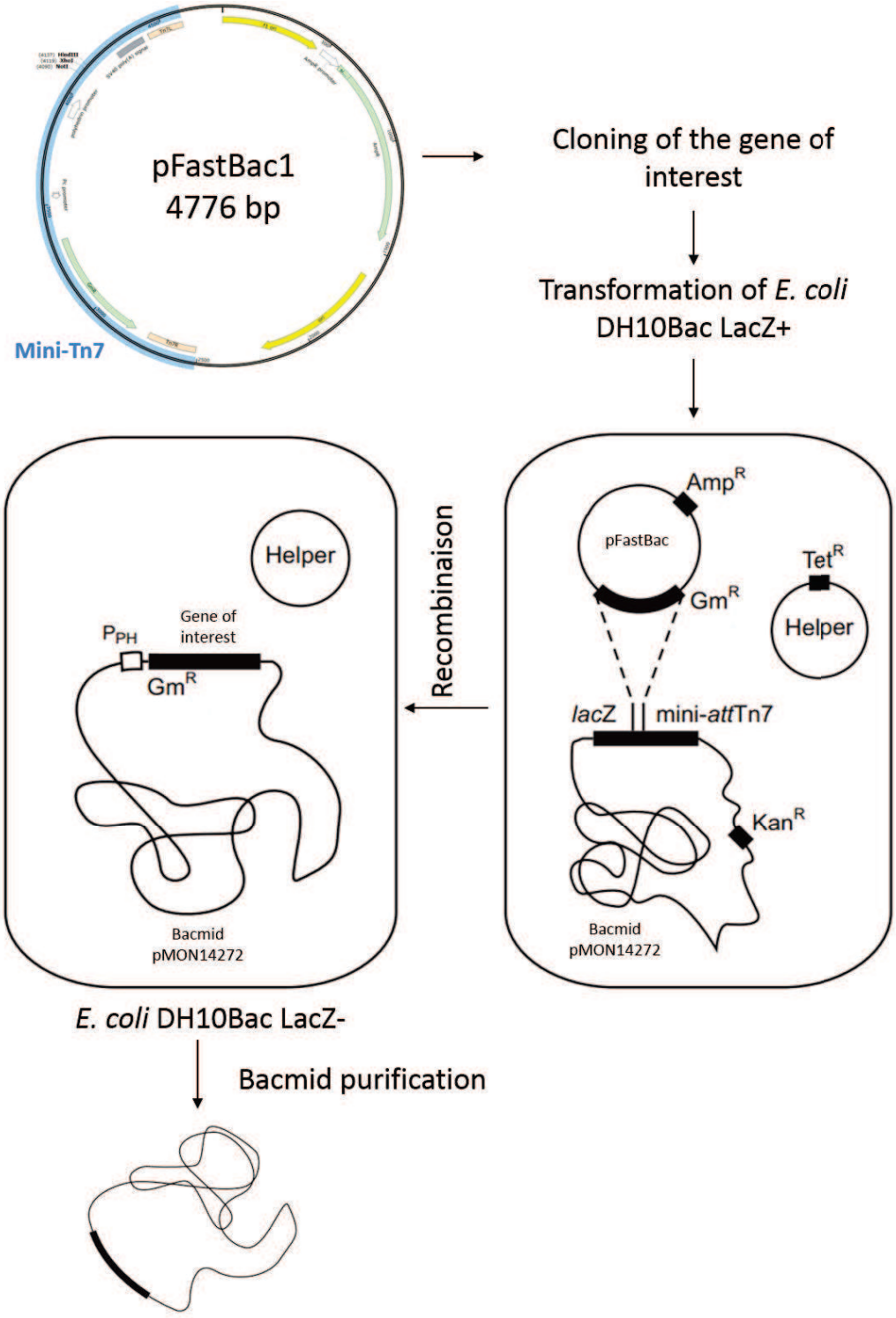


FIGURE 41

Recombinant bacmid generation using Bac-to-Bac technology

The Bac-to-Bac technology takes advantage of the site-dependent transposition of a mini-Tn7 cassette (in blue on the pFastBac1 donor plasmid), marked out by mini-attTn7R and mini-attTn7L sites. This transposition occurs *in vivo* in the bacterial strain *E. coli* DH10Bac, which contains an artificial pMON14272 bacmid as well as a helper plasmid encoding a transposase. The transposition efficiency is evidenced by the loss of the blue coloration of bacterial colonies in presence of X-Gal, due to the inactivation of the *lacZ* gene.

Modified from Bac-to-Bac® TOPO® Expression System, User Manual, Version A (15 December 2008), Invitrogen

1.1.2.3 SLIC

The SLIC method[595,596] is a relatively easy, ligase-independent cloning technique that requires no restriction enzyme. It is based on the homologous recombination of a destination vector and a DNA insert that takes place in several steps: creation of a blunt-ended double-strand break in the vector backbone, the generation of single-stranded overhangs by an exonuclease, and lastly, the hybridization of complementary ends between the vector and the insert (*figure 42*).

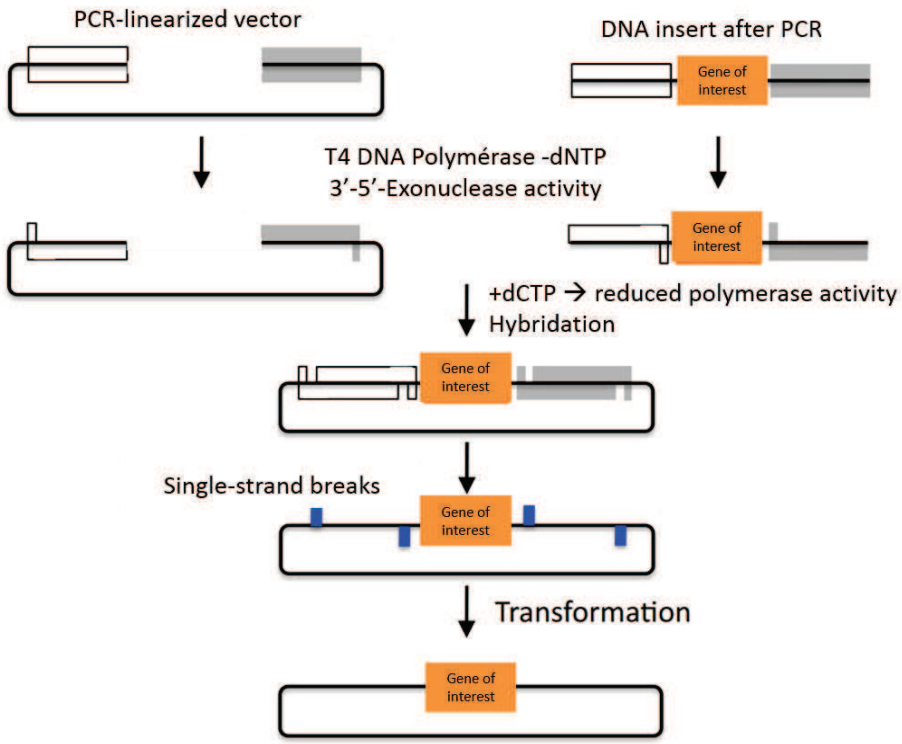


FIGURE 42

The "SLIC" method

This cloning method was used to design point or deletion mutants. First, the destination vector is linearized to the desired position by rolling circle PCR. Using a high-fidelity polymerase allows amplification of the whole vector while minimizing the risk of unwanted mutations. In parallel, the gene of interest (or only a part of the gene in the case of a deletion mutant) is amplified by PCR, with respect to the sites used for vector linearization. A 30-minutes treatment with T4 DNA polymerase, in absence of dNTP, allows digestion of the first nucleotides on the 3'-5' strand and creates sticky ends, with complementarity between the vector and the insert. After hybridisation, a direct transformation of *E. coli* DH5α allows amplification of the final vector, with single-strand breaks repair.

The linearization of the vector is achieved by a rolling-circle amplification using a high fidelity DNA polymerase like pfu or Phusion. Once the vector is fully synthesised, DpnI treatment allows degradation of the initial methylated vector and keeps intact the neo-synthesised unmethylated vector. In parallel, the DNA insert is also amplified by PCR to insert the desired mutations and modify its extremities in order to make it complementary to the vector.

Finally, the linearized vector and the insert are separately treated with the T4 DNA polymerase in the absence of dNTPs. Its 3'-5'-exonuclease activity allows nucleotide digestion, creating overhangs. Addition of a small concentration of dCTP can stop this exonuclease activity and restore the 5'-3'-polymerase activity, albeit blocked by the absence of the three other dNTPs.

Hybridization of the vector with the insert occurs at 37°C by simply mixing both DNAs. This hybridization product can then be directly used to transform *E. coli* DH5α cells. The ligation occurs naturally in the bacteria during the replication process.

1.2 The *Escherichia coli* expression system

Expression of recombinant proteins in *Escherichia coli* is the easiest, cheapest and most commonly used technique. Therefore, it is generally considered as a first step, before trying other expression systems in case of bad yield or failure. This system, however, presents some limitations, for high molecular weight proteins for example, or production of proteins which require post-translational modifications, unachievable in bacteria.

In our lab, several *E. coli* strains are available. The most frequently used in this work is *E. coli* BL21(DE3). It is characterised by its genotype:

F^- dcm ompT hsdS(r_B^- m_B^-) gal λDE3(lacI lacUV5-T7 gene 1 ind1 sam7 nin5),

where F^- means that the bacterium doesn't carry the F plasmid to conjugate;

dcm is the gene encoding cytosine methylase, an enzyme which methylates specifically the second cytosine in the CCWGG motif (with W = A or T);

ompT indicates that the VII protease is mutated to offer reduced proteolysis of the overexpressed proteins;

hsdS(r_B^- m_B^-) ensures that the exogenous DNA won't be digested by endogenous restriction enzymes;

gal means that the bacterium is not able to metabolize galactose as a carbon source;

and, λDE3(lacI, lacUV5-T7 gene 1, ind1, sam7, nin5) indicates that the bacterium carries the λDE3 prophage, that allows control of the T7 RNA polymerase expression by adding a $lacI^q$ operon inducer like IPTG. When it's expressed, the T7 RNA polymerase can transcribe genes under the control of the T7 promoter, like the gene of interest in the expression vector.

During this work, other strains from BL21(DE3) were also tested, in particular BL21(DE3)pLysS and BL21(DE3)pRARE. The pLysS plasmid encodes the T7 phage lysozyme, which can inhibit the T7 RNA polymerase in the absence of IPTG, to avoid basal expression, or "leaking", of the gene of interest; while the pRARE plasmid includes the genes encoding the $tRNA^{Arg/Ile/Gly/Leu/Pro}$ for rare codons. These two strains, however, were only little used since all the expression vectors I used didn't show any sign of leaky promoter, and the bacterial codon bias was adjusted for each sequence. Lastly, a HB101 strain was used to amplify plasmid DNA. It was chosen for its genotype RecA13, which ensures a low occurrence of exogenous DNA recombination.

The composition of culture media is a primary determinant for the good bacterial growth and numerous different media have been described. In this work, five of those were mainly used:

- The LB medium: this medium is a standard for bacterial culture. It was developed in 1951 by Giuseppe Bertani[597], then student under the supervision of Salvador Luria. For this reason, the LB medium is also called Luria-Bertani medium, wrongly since Bertani himself named it, in his very first publication, "Lysogenic Broth".

  This medium is easy to prepare and provides a complete set of nutrients: peptides, vitamins, minerals, trace elements, salts, etc. As described by Bertani, one litre of liquid LB medium contains 10 g of tryptone (a trypsic digest of casein), 5 g of yeast extract (a yeast autolysate, providing vitamins and trace elements necessary for growth) and 10 g of sodium chloride.

  However, the bacterial growth in LB medium remains limited, because of the shortfall of carbon sources. Carbohydrates are indeed negligible, and the other carbon source, peptides and amino acids, are not adequately used by bacteria. This is, in part, due to the size of the peptides: 75 % of these are too big to go through the porins with size-limit 650 Daltons[598].

  Later, it was observed that *Escherichia coli* starts a diauxic growth very quickly, around OD_{600nm} = 0.3, accompanied by a net decrease of the bacterial size[599] The amino acid catabolism, indeed, shows two distinct phases, depending on whether the amino acid is an easily degradable carbon source or not[600]. Thus, aspartate, arginine, serine and cysteine are catabolized first, followed by the others, especially the aliphatic ones. For these reasons, the growth of *Escherichia coli* is generally limited to an OD_{600nm} around 2-4. This medium had, in fact, been developed by Bertani to study lysogeny at low bacterial density.

  A solid version of the LB medium exists, called LB-Agar, simply supplemented with 15 g of agar per litre of medium. It is mainly used for Petri-plate culture.

- The 2xLB medium: similar to the classic LB medium, the 2xLB medium contains twice the quantities of tryptone, yeast extract and salt for the same final volume. This offers more initial input of nutrients compared with the LB medium and allows a higher bacterial growth, yet with the same limitations about carbon catabolism.

- The TB medium: this medium, named "Terrific broth", was specially developed to optimise bacterial growth and keep the cells in an exponential growth stage for a longer period (Tartoff, 1987). For one litre of medium, it comprises 12 g of tryptone, 24 g of yeast extract, 17 mM final of KH_2PO_4, 72 mM final of K_2HPO_4 and 0.4 % final of glycerol. It is, thus, an enriched medium, offering more vitamins, minerals, trace elements, and another carbon source with glycerol.

- The 2xYT medium: this medium is named according to its composition. It has twice more yeast extract than the LB medium, but the other components are not doubled. Thus, it contains 16 g of tryptone, 10 g of yeast extract and 5 g of sodium chloride for one litre of medium. It is, thus, enriched compared with the LB medium and can handle a longer bacterial growth.

- The AI medium or Auto-Inducible: this medium was designed for inducible recombinant protein production[601]. One litre of medium contains 10 g of tryptone, 5 g of yeast extract, 25 mM final of $(NH_4)_2SO_4$, 50 mM final of KH_2PO_4, 50 mM final of Na_2HPO_4, 0.5 g of glucose, 2 g of α-lactose, 2 mM final of $MgSO_4$ and trace elements.

  This medium is based on the observation made by Studier, that glucose prevents lactose-induced expression. Thus, a small concentration of glucose helps the bacteria metabolize this energy source and grow until reaching quasi-saturation. After glucose depletion, lactose can be metabolized and converted into allolactose, by the β-galactosidase. Allolactose is an analogue of IPTG, able to induce T7 promoter-controlled gene expression.

  This medium has the advantage of being self-sustaining, and does not require any human intervention of expression induction. Furthermore, the glucose, carbon source, allows achieving a higher bacterial saturation than in the LB medium. However, proteins produced in this medium might be subjected to degradation. This is due to the fact that lactose induction is much faster and harsh than with low concentrations of IPTG.

1.3 The "Baculovirus" expression system

In bacterial system, recombinant protein production can sometimes be problematic, leading to a bad expression or protein insolubility. This can be due to the lack of post-translational modifications or protein chaperones, often required to ensure a proper folding of the protein. In 1981, Miller hypothesised the use of baculovirus to produce recombinant proteins[602]. In 1985, this system was used for the first time to produce the IL-2 protein on a large scale[603]. Since then, this system has been used widely for protein expression. Recombinant baculovirus can be generated by transfection of insect cells with the recombinant bacmid. Thereafter, the amplified virus is purified and used to infect *en masse* insect cells (*figure 43*).

In the lab, three different strains of insect cells are available: Sf9 and Sf21 from *Spodoptera frugiperda*, and Hi5 from *Trichoplusia ni*. A study carried out on 23 different strains shows that none of them allows a maximal yield for every protein[604]. On the contrary, according to the recombinant protein that needs to be expressed, some cell lines might be more favourable than others, and vice versa. All three cell lines can be cultivated in suspension in Erlenmeyer flasks (for protein production), or as adherent monolayers in culture T-flasks (for maintenance of the cell lines, production of a viral stock or expression tests). The optimal growth temperature for insect cells is 27°C +/- 1°C, without CO_2 supply, with constant stirring, and the cultivation time varies between 48 and 72h, after infection.

- The Sf21 strain, or IPLB-Sf21AE, initially came from the ovarian tissue of the fall armyworm, *S. frugiperda*[605].
- The Sf9 strain is a subclone of the Sf21 strain and is the most commonly used nowadays. Like Sf21, they have a generation time of about 24-30h.
- The Hi5 strain, or BTI-TN-5B1-4, comes from the ovarian tissue of the cabbage looper, *T. ni*. This strain is more and more commonly used for its increased capacity of production as compared to other insect cell lines. These cells can be cultivated as adherent cells but the monolayer formation is irregular. Suspension is thus the preferred mode for their culture. Besides, these cells offer the advantage of a reduced generation time, to 18-24h.

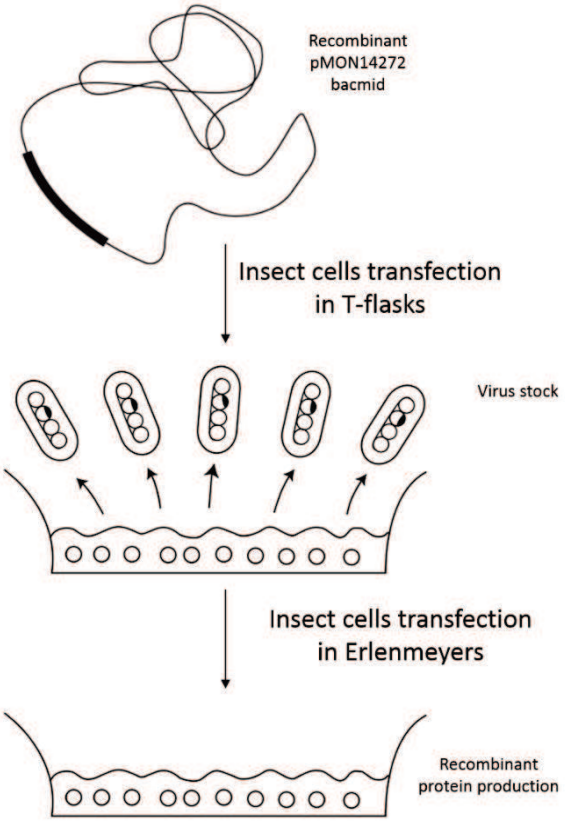


FIGURE 43

The "baculovirus" expression system

The recombinant bacmid is used to transfect insect cells, cultivated in T-flasks. After 72 hours, the first viral capsids are released in the culture medium. These virus are harvested and constitute the P1 stock, which can be further amplified on a wider scale (P2 stock, P3 stock, etc.) or directly used to infect large volumes of insect cells, for recombinant protein production.

Modified from Bac-to-Bac® TOPO® Expression System, User Manual, Version A (15 December 2008), Invitrogen

---

The culture medium that is typically used for Sf9 and Sf21 cells is Grace's Insect Cells medium (Sigma), supplemented with 10% of fetal bovine serum and 28 mg/L of culture of gentamycin, to prevent bacterial growth. This medium can be prepared up to one month in advance, and stored at 4°C for future use. It contains salts including 55 mM of KCl, 11 mM of $MgCl_2$ and 11 mM of $MgSO_4$, amino acids, vitamins, and 80 mM of sucrose. In the case of Hi5 cells, a serum-free medium is preferred, like Express Five® SFM medium, supplemented with 18 mM of L-glutamine. Generation and maintenance of the viral lineages from recombinant bacmids, as well as the cell cultures, are carried out by the common "baculovirus" service of IGBMC (Illkirch).

2. Biochemistry methods

After cultivation, cells are harvested by centrifugation (from 1000 Xg for insect cells to 4000 Xg for bacteria). The cell pellets obtained can then either be directly processed or stored at -20°C for a couple of weeks or -80°C for longer term.

In order to isolate the expressed recombinant protein, several biochemistry steps are needed. This starts with the cell lysis, an essential prerequisite to release the protein if it's not excreted in the culture medium. Thereby, the soluble content from the cytoplasm is released and treated by chromatography methods to purify the protein of interest, i.e., to clear it out from all the contaminants, including endogenous proteins.

2.1 The cell lysis

2.1.1 Lysis buffer

Lysis is performed by resuspending the cell pellets in an appropriate lysis buffer. The choice of this solution is crucial to keep the protein soluble. Several parameters have to be taken into account while optimising the lysis conditions according to the protein of interest, including the pH of the buffer, the type of buffer, the salt concentration, reducing agents, stabilising agents and protease inhibitors.

The first parameter to be considered is the pH of the buffer, chosen according to the global isoelectric potential (pI) of the protein to purify. This pI can be calculated from the primary sequence of the protein, with bioinformatics tools like ProtParam (Expasy). A common observation is that a protein is rarely soluble when the pH of the solution is close to its pI, which means that it will have a global net neutral charge. This would abolish all the electrostatic interactions, leading to structural instability. This rule is however, biased by the calculation of the pI, which considers all the amino acids of the protein, and not just those which are exposed to the solvent, at the surface. For this reason, it is common to use as a first trial a buffer at pH 7.5, like Tris-HCl or HEPES, to imitate the physiological environment of the cell. Using a strong buffering agent helps maintain the lysate at constant pH and the choice of the buffering agent itself is important. In particular, it must be used in its buffering range, i.e., the range of pH with optimum buffering capacity. Furthermore, some buffering agents are sensitive to temperature. This is the case for Tris-HCl: when buffered at pH 8.0 at room temperature, its pH will decrease to 7.7 at 4°C and increase up to 8.6 at 37°C. Tris-HCl remains however the most commonly used buffering agent.

The salt concentration is also a parameter that has to be taken into account. Here again, 150 mM of NaCl is often used as a working base to emulate the physiological conditions.

In the cell, redox potential is globally reducing, mainly due to high concentrations of glutathione (up to 5 mM in hepatic cells). But the atmosphere itself is rather oxidative. Thus, after the cell lysis, the protein of interest must be protected against excessive oxidation. This oxidation targets in particular cysteine residues, leading to inappropriate disulphide bond formation. Addition of a reducing agent is thus recommended such as: β-mercaptoethanol (BME), dithiothreitol (DTT) and tris-carboxyethyl-phosphin (TCEP). The choice of the reducing agent is made depending upon its stability (BME has a half-life of only a couple of days in aqueous solutions; just barely more for DTT; and TCEP is stable for a couple of weeks), its reducing power (roughly, 10 mM BME are equivalent to 5 mM DTT and 1 mM TCEP), its compatibility with chromatography techniques (at high concentration, BME and DTT are not compatible with nickel ions used in affinity chromatography), and finally, the cost (from 8.5 € for one mole of BME to over 10000 € for one mole of TCEP). It must be noticed that the usage of a reducing agent at reasonable concentration does not destabilise the disulphide bonds required for protein folding, since they are generally buried in the core of the protein and thus inaccessible to the solvent.

Several stabilising agents can be used, for example, 10 % glycerol is commonly used for protecting exposed hydrophobic regions. Addition of an inert protein like BSA can also, in some cases, help stabilise the protein of interest. Finally, some detergents or ionic compounds like sulphates, arginine or citrate, can be used at low concentrations to protect surface electrostatic interactions.

Lastly, another factor that is crucial during cell lysis is the presence of proteases, which, once released, are an immediate threat to the protein of interest. Their activity can be reduced by working at a constant low temperature, like, 4°C and maintaining physiological or basic pH, but this is not always compatible with the protein stability. Addition of protease inhibitors can, thus, limit their activity. In our lab, two commonly used inhibitors are: phenylmethanesulfonyl fluoride (PMSF), a serine protease inhibitor, against trypsin, thrombin, papain, etc.; and protease inhibitor cocktails (cOmplete EDTA-free, Roche). Also, EDTA can be used as a metalloprotease inhibitor, due to property to chelate divalent cations like Mg^{2+} or Ca^{2+}. Its usage is, however, not compatible with certain techniques such as nickel affinity chromatography.

2.1.2 Extraction techniques

Once resuspended in the lysis buffer, the cell membrane must be disrupted. The extraction techniques can be categorised into, either mechanical or chemical.

Among the mechanical techniques, I used during this work, one can mention sonication, Dounce homogenisation and Aminco-French press. Sonication is the most widespread technique in the lab. It is based on the emission of ultrasonic waves, alternating high-pressure (compression) and low-pressure (rarefaction) cycles. During low-pressure cycles, small vacuum bubbles are produced in the solution, and implode during high-pressure cycles. This implosion, also called cavitation, is accompanied by a sudden local temperature increase, up to 5000°C, and pressure increase, up to 2000 bar. Also, this cavitation results in liquid jets approaching the speed of sound (around 300 m/s), which causes strong shearing forces, able to break the cell membranes. This technique however presents local heating problems, if not global when the sonication is continued for several minutes, which can be harmful for the protein of interest. The Aminco-French press raises the same problem. In spite of its name, this press is nothing French but was developed by Charles Stacy French. It is a hydraulic pump-based technique, a piston pushed inside a cylinder containing the sample. The applied pressure reaches 400 to 700 bar, and the sample is suddenly decompressed and drawn down past a thin pipe, often thinner than the cell diameter itself, causing the disruption. Lastly, the Dounce homogenizer looks like a graduated cylinder, from one mL to a couple of hundreds, containing the sample, and inside which a tight piston is manually pushed. The disruption of the cells is caused by their passage in the limited space between the cylinder wall and the piston. This manual technique generates a smaller pressure increase as compared to the other two techniques. It is thus, a less efficient technique, but also less traumatic for the sample, especially when working with proteins prone to degradation.

Among the chemical techniques, detergents and lytic enzymes were used. CHAPS was the most effective detergent for my work. It is a non-denaturing detergent (unlike SDS, anionic) and zwitterionic (unlike other detergents like Triton, Tween and glycosides, which are non-ionic) as it carries both positive and negative charges, but has a net neutral charge. This property is crucial for solubilizing membranes and abolishing non-specific protein-protein interactions thereby, reducing the aggregation risk. It is, however, less efficient than Tween or Triton, but has the quasi-unique advantage of being easily dialyzable.

Lysozyme, the most commonly used lytic enzyme, is an acidic hydrolase that can digest the bacterial polysaccharide membrane. It is usually used at concentrations around 15 mg/L of culture, and incubated in the cell lysate for a few minutes. However, unlike mechanical techniques, it

releases the chromatin without destroying it. This highly viscous sample must thus be treated with DNase or sonication before clarifying the lysate.

Finally, post-lysis, the lysate obtained is named “total extract” that must then be clarified by ultracentrifugation, thereby allowing to separate the “soluble extract” (soluble proteins, nucleic acids, etc.) from the “insoluble extract” (insoluble proteins, cell debris, organelles, etc.). Several steps of purification are then required to isolate the protein of interest and get a high purity level as described below.

2.2 Chromatography techniques

The premises of chromatography date back to the early 20th century, when Mikhail Tswett, a Russian botanist, managed to separate chlorophyll and carotenoids, plant pigments, by straining the mix through an adsorption column made of calcium carbonate. He named this new technique, chromatography, in 1906, from old Greek χρῶμα/khrôma, “colour” and γράφειν/graphein, “to write”. However, this technique was quickly forgotten. Tswett indeed published his invention in Russian only, at the dawn of the 1917 Russian Revolution. Plus, Willstätter, Nobel Prize laureate in chemistry in the year 1915 for his work on chlorophyll, and Stoll, Swiss chemist who founded Sandoz (today, Novartis), discredited Tswett’s invention after they failed to reproduce his results. Their mistake was simply the use of a corrosive adsorbent, which destroyed chlorophyll.

It was not until 1931 that chromatography was established as a technique, in Heidelberg, Germany. Edgar Lederer, who had heard about this method, managed to separate carotenoids from egg yolk. From then on, chromatography has undergone a remarkable development, becoming a powerful and essential tool in biochemistry and organic chemistry, still widely used nowadays.

The principle of chromatography, whatever the method, remains the same: a mix of molecules is applied onto a stationary phase (the adsorbent), and driven by a mobile phase, generally liquid (the eluent). the most commonly used is the FPLC technique (Fast protein liquid chromatography) where the absorbent is fine grain size resins, generally made from a polymer of agarose or cross-linked dextran, and packed in a column. The sample is driven through this resin by the eluent, at low-speed and low-pressure (between 0.5 and 5 mL/minute, for a working pressure of 3 to 15 bars, according to the resin resistance). At the end of the column, two detectors measure the sample’s conductivity and absorbance at 280 nm. It is thus, possible to plot a chromatogram to follow the purification. This technique, especially developed for protein biochemistry, was designed in 1982 by the Swedish company Pharmacia (today, GE Healthcare).

Several types of resin are available, each having different properties. For example, affinity resins bind specifically to a protein motif or a given amino acid sequence, (*figure 44*); the ion-exchange resins, which are charged and can bind proteins depending upon their isoelectric potential (*figure 45*); hydrophobic resins, which bind protein through hydrophobic interactions; or size exclusion resins, which separate biomolecules according to their hydrodynamic volume (*figure 46*). With the exception of hydrophobic resins, all the others were used during this work and are described hereafter.

2.2.1 Affinity chromatography

Since the 1970's, with the growth of biotechnology, affinity chromatography gained popularity as a main step in protein purification. The protein of interest can be expressed in fusion with a specific tag, which has affinity for an immobilized ligand on the resin, i.e., a derivatised resin. The benefit of this method is to get, theoretically, a pure protein after only one step of chromatography.

Tags often consist of short polypeptide sequences, like the flag-tag DYKDDDDK, specifically recognised by the anti-flag antibody; the strep-tag WSHPQFEK, which binds streptavidin. It can also be entire proteins like the Glutathione S-transferase (GST-tag), a 211-amino acid protein, whose natural substrate, glutathione (Glu-Cys-Gly), can be immobilized on an affinity resin. Lastly, the most commonly used tag is the 6x-His-tag, including a succession of six histidine residues. Through their imidazole ring, histidines exhibit a binding affinity to divalent metals like nickel, cobalt, copper, zinc or iron.

I exclusively used this tag to purify my proteins of interest with nickel affinity purification. The stationary phase used is Ni-NTA resin (Qiagen), composed of modified Sepharose-6B beads (6% of agarose), and covered with nitrilotriacetic acid (NTA) capable of complexing nickel ions. This resin can be used in two different ways to allow binding with tagged-protein:

- Using a chromatography column containing the Ni-NTA resin like 1 or 5 mL HisTrap columns (GE Healthcare) connected to a chromatography workstation (Äkta Purifier UPC 10, GE Healthcare).
- Using the "batch" method, which requires incubating the sample directly with the resin, with slow stirring.

After binding the protein of interest to the affinity resin, the latter is washed to remove as many contaminants as possible, and the protein is recovered by an elution step, using an eluent acting as a competitor. It is often a molecule exhibiting a higher affinity for the resin than the tag itself. In the case of the 6x-His-tag, this competitor is imidazole, a metabolic precursor of histidine, presenting the same heterocyclic aromatic structure (*figure 44*).

Each of the above-mentioned methods has pros and cons. Using a chromatography column, the sample is injected at a speed of 1 to 5 mL/minute. Depending on the protein and the purification conditions, this method is not always optimal because, affinity chromatography on column requires that the tag be well exposed in order to bind quickly and efficiently in spite of the fast flow rate. The pH conditions must also be optimal since at low pH, the histidine residues of the 6x-His-tag are protonated (pKa = 6.8) and thus lose their affinity for nickel. The batch method can, in some cases, increase the yield of purification due to a prolonged binding time. But it also has the disadvantage of a "blind" manual work. No absorbance or conductivity follow-up is possible, and elution with an increasing gradient of competitor is hardly feasible manually.

Another affinity chromatography technique that I used is based on heparin-affinity. Heparin is a glycosaminoglycan, which, because of its structure and negative charge, is able to mimic DNA. It is thus an ideal method to purify DNA-binding proteins. HiTrap Heparin 5 mL (GE Healthcare) chromatography columns were used, consisting of modified Sepharose-6B beads, covered with

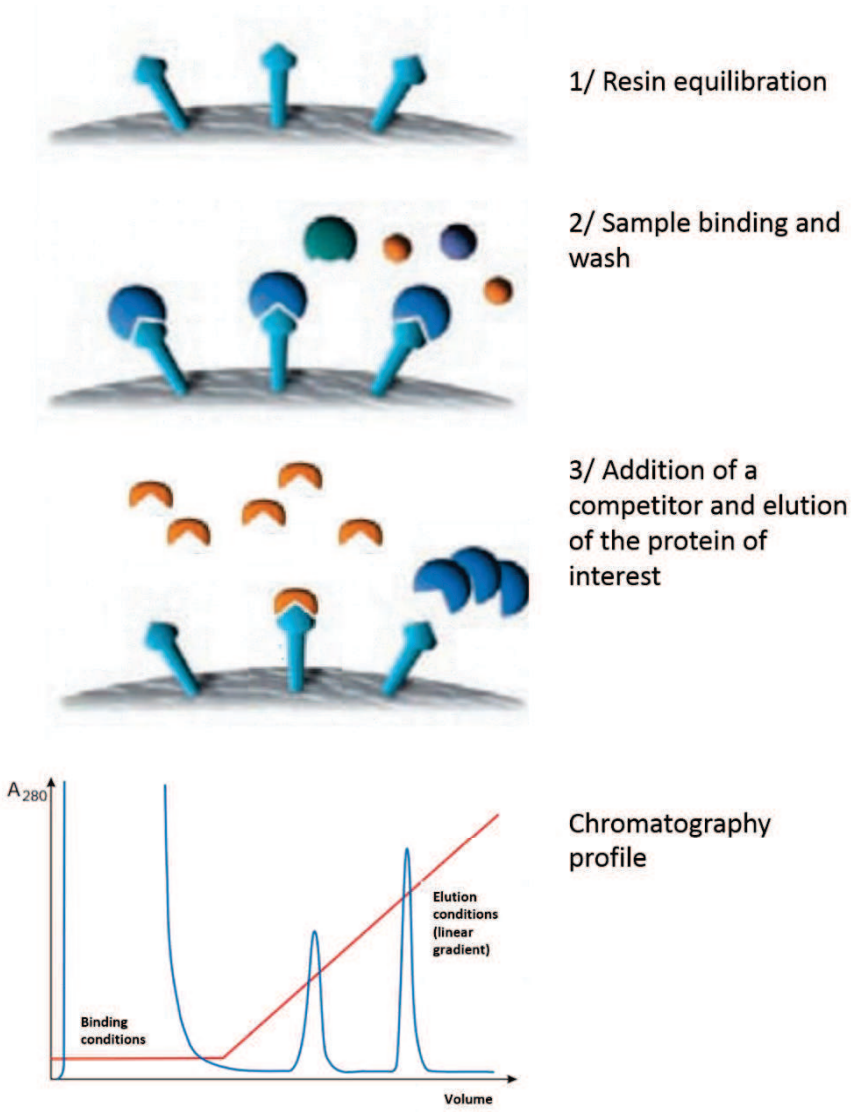


FIGURE 44

Affinity chromatography

Affinity chromatography is based on a specific interaction between a modified resin and the protein of interest. This protein can be expressed in fusion with a tag for specific resin affinity, or can be naturally affine towards a given resin.

Two types of affinity were used during this work: the 6xHis tag affinity towards Ni^{2+} ions; and the naturel DNA-binding affinity towards heparinized resin. In both cases, the protein is retained on the resin, washed and finally eluted using a competitor (imidazole in the first case; ionic strength in the second case).

Modified from Affinity chromatography handbook, 18-1022-29, GE Healthcare

heparin. The sample was applied onto the resin, washed and finally, the protein of interest was recovered by increasing ionic-force gradient, to destabilise the heparin-protein interactions.

2.2.2 Ion-exchange chromatography

This technique is based on the reversible interaction between a charged biomolecule and a stationary phase, of opposite charges. It thus allows to separate the molecules according to their net surface charge (*figure 45*). Developed in the 1960's, it remains one of the most advantageous chromatography technique, offering high resolution to separate particles with only slight difference in the net charge and a strong binding capacity. Like affinity chromatography, very large volumes of sample can be applied onto these resins. The protein of interest is retained during the whole process, and its final elution acts as a "concentrator", the final volume being reduced to a few mL.

Ion-exchange chromatography takes advantage of proteins being amphoteric molecules where their net charge depends upon pH. This pH/net charge relationship is protein-specific, because of their amino acid composition, their structure and their chemical microenvironment. If the pH of the working solution is below the pI of the protein, the latter will be globally positively charged, and thus can be retained on a negatively charged matrix, the cation-exchanger. On the contrary, if the pH is above the pI of the protein, it will be negatively charged and will bind a positively charged matrix, the anion-exchanger.

Several ion-exchange matrices are available, with components cross-linked to 6% agarose beads or poly(styrene/divinyl benzene) beads, packed in chromatography columns, and named after the charged group they carry:

- Q, for quaternary amine, a strong anion-exchanger;
- SP, for sulphopropyl, a strong cation-exchanger;
- DEAE, for diethylaminoethyl, a weak anion-exchanger;
- CM, for carboxymethyl, a weak cation-exchanger;

The terms "strong" and "weak" do not reflect the binding force of these resins, but rather their capacity to maintain their ionic state depending on the pH. Strong groups have a weak buffering capacity, meaning their net charge doesn't vary in a large range of pH. On the contrary, weak groups have a very strong buffering capacity. Proton exchange with the environment is rapid when the pH varies, leading to a loss of capacity of these resins. Despite these obvious disadvantages, using weak resins can nevertheless be beneficial since they exhibit a different selectivity for biomolecules compared with their strong equivalents.

After binding to the protein of interest, the resin is washed to get rid of the unbound contaminants, and the protein is finally recovered by elution. Most frequently, an ionic force gradient is used, consisting of an increasing salt concentration (typically from 0 to 1-2 M NaCl). By competition, weakly charged proteins are eluted first, whereas the ones that are strongly charged are eluted last, at high salt concentration of the eluent.

An alternative method consists of applying a pH gradient, to modify the net surface charge of the proteins retained on the resin. This technique, called chromatofocusing, allows separating, at constant ionic force, proteins with a pI which differs by only 0.02 pH units. However, one major drawback is the aggregation risk when proteins cross the pH ≈ pI border and their net charge becomes zero.

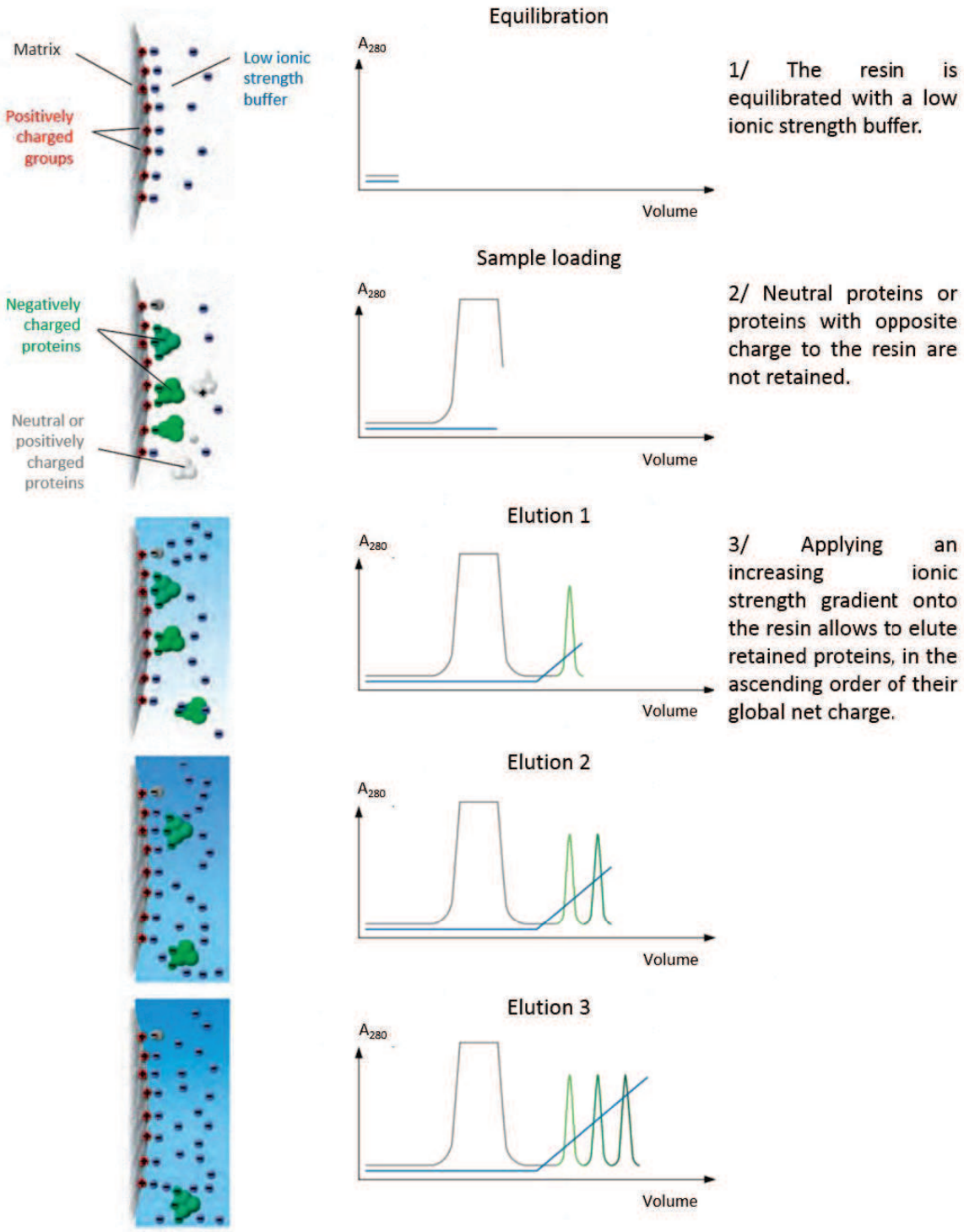


FIGURE 45

Ion-exchange chromatography

Ion-exchange chromatography is based on the electrostatic interaction of a charged molecule (protein, nucleic acid, etc.) with an opposite-charged resin. The more the surface is important, the stronger the interaction is. Elution is carried out by ionic strength gradient, allowing thus to fractionate the sample in accordance with the charges of its different components.

Modified from Ion exchange chromatography and chromatofocusing handbook, 11-0004-21, GE Healthcare

2.2.3 Size exclusion chromatography

In 1955, Grant Lathe and Colin Ruthven published their work on separation of molecules according to their molecular weight, using hydrated starch columns[606,607]. This invention earned them the John Scott medal in 1971. In 1959, Jerker Porath and Per Flodin developed, for Pharmacia, a new type of resin called Sephadex™, for *Separation Pharmacia Dextran*[608]. Made of small beads of a few tens of µm in diameter, it consists of fine dextran crosslinking.

These new developments have led to a new chromatography method, of these, among the easiest but also the most essential one is size-exclusion, also called gel-filtration (GF). It allows to separate biomolecules according to their mass, or more precisely, to their hydrodynamic volume or Stoke radius. This mass to Stoke radius relationship is constant in the case of globular proteins only. Unlike the methods described above, the protein does not bind to the resin instead the resin acts like a molecular sieve, allowing the proteins to travel through it by diffusing through the matrix pores. Depending on the size and the cross-linking degree, these pore size is important and thus offers different separation ranges (or fractionation ranges). This allows to purify small peptides as well as, big proteins or large macromolecular complexes. The resins can be categorised into three quality levels: macro-fractionation resins, preparative resins and analytical resins.

Elution in size-exclusion chromatography is called as isocratic, since the mobile phase, or eluent, is not modified over time by addition of a competitor, increasing salt concentration or pH gradient like previously reported. Molecules with a size greater than the resin pore diameter are excluded from the resin, (justifying the term “size-exclusion”) and travel faster than the molecules entering the beads. They are eluted first, at void volume V_0. Among the excluded proteins, aggregates are misfolded or unfolded protein accumulations, which are incompatible with functional or structural studies. Molecules with very small size, i.e., having a diameter smaller than the resin pores, penetrate through the resin, and thus are eluted last, at total volume V_t. These molecules correspond generally to small peptides, salts and other organic compounds. Finally, intermediate-size molecules enter only partially the pores of the resin and are eluted in the descending order of their hydrodynamic volume (*figure 46*).

An alternative role of gel-filtration is desalting and buffer-exchange. A protein in buffer A can be injected in a GF column, equilibrated beforehand in buffer B. The protein will thus be eluted in the buffer B, whereas the rest of the compounds of buffer A (salt, etc.) are eluted at total volume V_t.

Among the gel-filtration resins, a few are mentioned below.

- Sephadex, a cross-linked dextran resin, with large beads of 50 to 300 µm in diameter. It is a low-resolution macro-fractionation resin, ideal for intermediate purification steps or desalting.
- Sephacryl, a copolymer of dextran and bisacrylamide, with beads about 25 to 75 µm in diameter. This heterogeneity makes it a low-resolution resin, but still able to fractionate a large range of molecular masses.
- Sepharose, a cross-linked agarose resin (from 2 to 6 %), with beads of 45 to 200 µm in diameter. Here again, the macro-fractionation occurs on a large range of molecular masses. This resin also has the advantage of an increased stability in denaturing conditions (8 M urea or 7 M guanidinium chloride).

- Superose, a highly cross-linked agarose resin (from 6 to 12 %), with small beads of 10 to 15 µm in diameter. This resin is the preparative counterpart of the Sepharose resin. It allows a qualitative fractionation over a large range of molecular masses.
- Superdex, a copolymer of dextran and agarose, with fine beads of 7 to 13 µm in diameter. It is a high resolution analytical resin, able to fractionate a mix of protein on a small range of molecular masses.

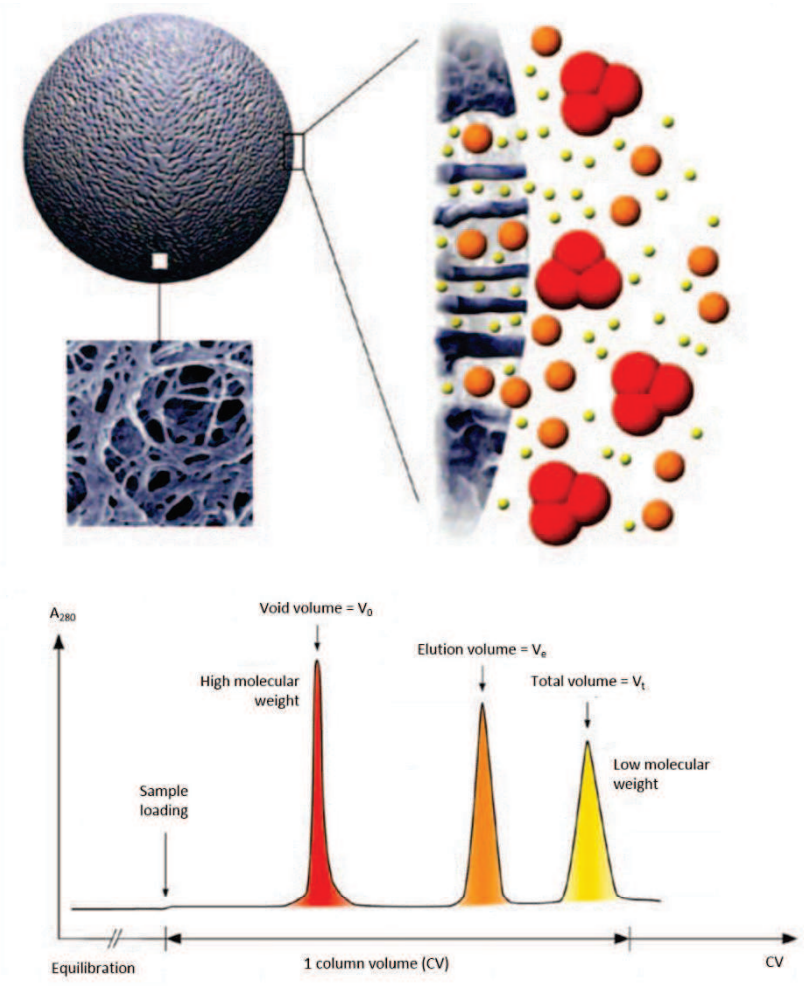


FIGURE 46

Size-exclusion chromatography – Gel-Filtration (GF)

Gel-filtration allows an isocratic elution of molecules, separated according to their hydrodynamic volume. After sample loading, the molecules diffuse into the pores of the resin, causing a more or less important retardation of their migration. High molecular weight molecules (above the pore size) are eluted first, in the void volume (V_0). Intermediate-size molecules are then eluted according to their decreasing hydrodynamic volume. Finally, the smallest molecules are eluted last, in the total volume (V_t).

Modified from Gel filtration handbook, 18-1022-18, GE Healthcare

2.3 Purification under denaturing conditions

Of all the expression systems, *E. coli* remains the one that gives the best production yields. However, as outlined above, absence of chaperones and post-translational modifications often leads to improper folding of eukaryotic proteins, which are then eliminated by the bacteria. Modifying the culture conditions can sometimes help improve the solubility of these proteins. Among the possible options are: induction at low IPTG concentration, growth at low temperature, between 15 and 18 °C (or even less with strains like BL21(DE3)ArcticExpress), or even the addition of sorbitol, arginine or trehalose in the culture medium. These techniques are, however, not always efficient, and the protein "waste" accumulate in inclusion bodies, non-lipidic membrane bound bacterial vesicles.

These inclusion bodies contain in great majority the protein of interest, not degraded instead in the form of aggregates. Despite their empirical and random nature, several approaches were imagined to recover these proteins, quasi-pure, and "renaturate" them, i.e., induce their refolding *in vitro* to restore their biological functions.

A standard procedure consists first of isolating the inclusion bodies. To do so, the bacterial pellet is resuspended in a wash buffer to dissolve and homogenize it. It contains 100 mM of NaCl, 20 mM of Tris-HCl pH 7.5 and 5 mM of BME. This suspension is then sonicated for 15 to 20 minutes, at 60 % amplitude, alternating 8 seconds of sonication and 2 seconds of break. This is followed by a centrifugation at 50000 Xg for 20 minutes at 4°C. The pellet is then washed three times with the previously described buffer, the second wash containing additionally 1 % of Triton X-100 to solubilize membranes. Each washing step is followed by a centrifugation at 50000 Xg for 20 minutes at 4°C. The pellet finally recovered contains the inclusion bodies, which can be further treated or stored at -80°C for future use.

The next step consist of dissolving the inclusion bodies in a strong denaturing agent (8 M urea or 7 M guanidinium chloride), or in ionic detergents like sarkosyl. These compounds abolish all non-covalent interactions between proteins to solubilize them. In parallel, a reducing agent like BME can be added to prevent intra- and intermolecular disulphide bond formation. The folding of the solubilized protein is then achieved by gradual decrease of the denaturing agents, by dialysis or gel-filtration. Addition of aggregation inhibitors or stabilisers can sometimes help improving the refolding yield like L-arginine and L-proline, cyclodextrins, cyclic oligosaccharide polymers; and stabilisers include glycerol, sucrose, trehalose and polyethylene glycol (PEG). The efficiency of refolding can finally be verified by biophysical characterisation or bioassay.

If successful, this method has the great advantage of reaching high yield of several tens of mg of protein from one litre of bacterial culture. Its implementation is however highly time-consuming and its efficacy is very low.

3. Biophysical characterisation methods

A great number of biophysical methods can be used to quantitatively or qualitatively characterise a proteinaceous sample.

3.1 Protein gel electrophoresis

This technique, also called PAGE (*PolyAcrylamide Gel Electrophoresis*) allows the analysis of a mixture of proteins in solution. As previously explained, proteins are amphoteric molecules. They can thus be separated in an electric field, depending on their physical properties, such as their charge, but also their mass, in a polyacrylamide gel which plays the role of molecular sieve.

Although the first electrophoresis experiments using sucrose gels date back to the 1930's (Arne Tiselius, 1930), it was not until 1959 that Ornstein and Davis developed polyacrylamide gels, still widely used nowadays in the lab. Polyacrylamide is a copolymer of cross-linked acrylamide and bisacrylamide which forms a mesh of regular porosity. The acrylamide concentration in the gel affects the final pore size and rigidity of the gel, so that the separation resolution can be optimised in accordance with the properties of the protein of interest.

The system comprises a gel sandwiched between two glass plates, itself placed between two independent chambers, each of which contains a migration buffer. The only electrical path between the two chambers passes through the gel, and the sample migrates therethrough from one electrode to the other. The separation depends on the migration speed of each molecules such that

$$v_p = u_p . E$$

where E is the intensity of the electric field and u_p, the electrophoretic mobility given by the equation

$$u_p = \frac{z}{6\pi\eta r}$$

with z, the net charge of the molecule, r, its radius, and η, the viscosity of the medium. Thus, small and highly charged molecules will have an increased migration speed, whereas large and poorly charged molecules run slower (*figure 47*).

Three electrophoresis techniques can be described:

- Native continuous electrophoresis: the technique developed in 1959 by Ornstein and Davis. The term "continuous" refers to the continuity of the pH throughout the gel, the sample and the migration solution. A large range of pH can be used, according to the physicochemical properties of the protein studied. The proteins do not, however, penetrate the gel at the same time: the sample is loaded in a "well" formed at one extremity of the gel, and the proteins that are at the bottom of the well, the closest to the gel, penetrate first. This leads to a low resolution. This type of system is thus only

rarely used for protein electrophoresis, but is entirely suitable for nucleic acid studies (see chapter: gel shift assay).

- Native discontinuous electrophoresis: in 1964, Ornstein and Davis improved their electrophoresis technique and developed disc electrophoresis, initially to study blood plasma. They imagined thus a two-phase gel, with an upper part at low acrylamide concentration (5 %) and pH = 6.8 (the "stacking gel"), and a lower part at higher acrylamide concentration (between 10 and 20 %) and pH = 8.8 (the "resolving gel"). Furthermore, this technique utilizes two electrolytes, a "fast" leading electrolyte and a "slow" terminating one. This allows the proteins to be concentrated in a reduced area and a better resolution is achieved. Typically, the gel contains the fast electrolyte Cl^- (in the form of Tris-HCl), and the migration buffer contains glycinate, a slow electrolyte (in the form of glycine).

  Two phenomenon can then be observed. The proteins together with the glycine penetrate the stacking gel and run from the cathode to the anode. Glycine, with its pI of 6.5, is mostly neutral at pH 6.8, which explains its very slow migration speed. Plus, the low abundance of ions in this part of the gel causes a local increase of the resistance, together with an increase of the voltage to keep a regular magnitude throughout the gel. Proteins thus migrate quickly, until they reach the Cl^- front, before entering the resolving gel. This phenomenon is referred as isotachophoresis. This first step thus serves to concentrate the sample into a thin band, and to eliminate all aggregates without disturbing the protein front when entering the resolving gel.

  Electrophoresis continues with an increase of the viscosity in the resolving gel, causing an increased friction force, and thus, a slowdown of the protein migration. Glycine, fully ionized into glycinate at pH 8.8, crosses the protein front because of its reduced size. A new glycinate/Cl- front appears downstream and proteins are free to migrate at constant voltage in accordance to their electrophoretic mobility (*figure 47)*.

  To study acidic proteins, an alanine-acetate or potassium-histidine couple, more acidic, can be used; another possibility is to add a dye to the sample, like Coomassie blue (or G-250), negatively charged. It binds to the hydrophobic regions of the proteins without denaturing them, and allows thus migration of all proteins, even acidic.

- Denaturing discontinuous electrophoresis: the native methods described above do not allow the molecular weight of the separated proteins to be determined. To overcome this issue, Laemmli had the idea, in 1970, to introduce a detergent, sodium dodecylsulphate (SDS), a strong anionic surfactant[609]. This property allows it to fully denature proteins (cancelling the structural effect on migration) and to cover them with anionic charges by direct linear and constant interaction with the SDS (cancelling the net charge effect on migration). Using a discontinuous system, proteins can then migrate along the gel depending only on their apparent mass, since the mass/charge ratio remains constant. This method, commonly called SDS-PAGE, has become the most popular form of electrophoresis in the lab.

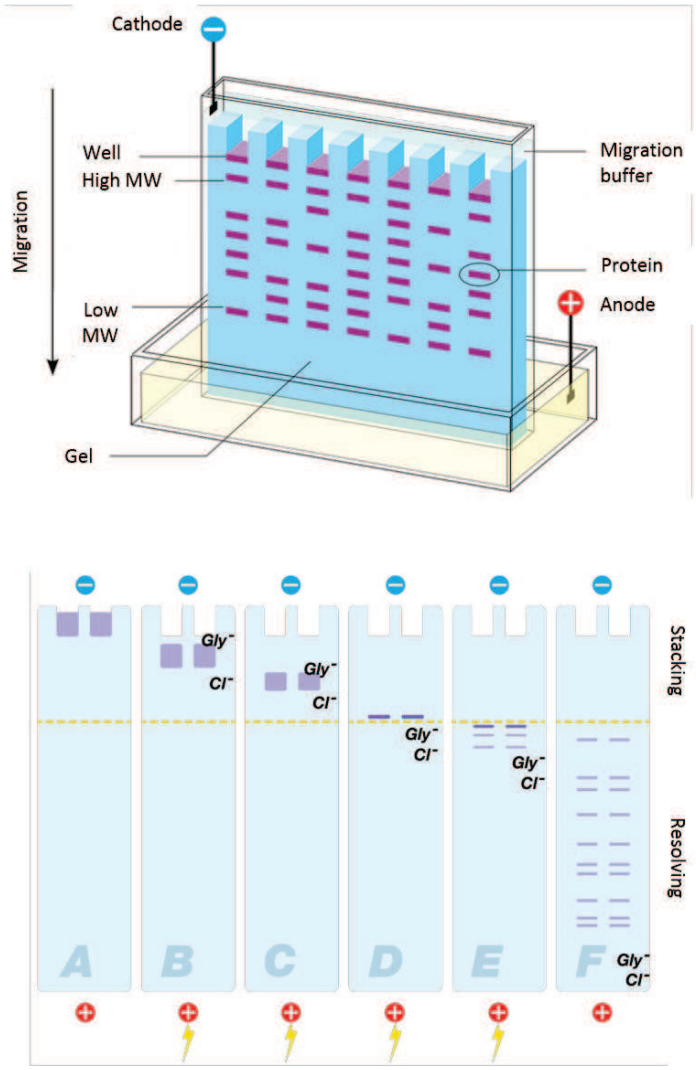


FIGURE 47

Polyacrylamide gel electrophoresis (PAGE)

Electrophoresis allows to separate charged molecules, in a gel placed in an electric field. Samples are loaded inside wells at one end of the gel, and subjected to an electric current. Proteins, which are negatively charged in electrophoresis conditions, migrate from the cathode to the anode, according to their size and net charge.

In order to get a better resolution, a first step, called “stacking”, uses the migration properties of two electrolytes (generally, glycine Gly^- and chloride ion Cl^-) to sandwich the sample and concentrate it into a narrow zone, avoiding thus diffusion effects before the “resolving” step.

Modified from A guide to polyacrylamide gel electrophoresis and detection, Bio-Rad

---

Although commercial gels are available, it is not unusual to make, or “cast”, gels oneself in the lab. Samples to be analysed are supplemented with a loading buffer, including glycerol and bromophenol blue for ease of loading, as well as BME and SDS in the case of denaturing gels. In the latter case, samples are boiled to achieve complete denaturation before loading. After migration, the gel is stained with a staining buffer, then destained to highlight the blue bands which correspond to proteins. The following table summarises the buffers and gel composition.

	Native gels		Denaturing gels
	Protein pI < 7	**Protein pI > 7**	
Migration buffer	350 mM β-Alanine 150 mM acetic acid pH 4.3	50 mM Tris-Base 380 mM Glycine pH 8.9	25 mM Tris-Base 250 mM Glycine 0.5 % SDS
"Stacking" gel	3 % Acryl/Bisacrylamide 39:1 250 mM KOH-Acetate pH 6.8 0.15 % APS 0.2 % TEMED	3 % Acryl/Bisacrylamide 39:1 125 mM Tris-HCl pH 6,8 0.15 % APS 0.2 % TEMED	5 % Acryl/Bisacrylamide 39:1 125 mM Tris-HCl pH 6.8 0.1 % SDS 0.15 % APS 0.2 % TEMED
"Resolving" gel	10-20 % Acryl/Bisacryl 39:1 400 mM KOH-Acetate pH 4.3 0.15 % APS 0.2 % TEMED	10-20 % Acryl/Bisacryl 39:1 350 mM Tris-HCl pH 8.8 0.15 % APS 0.2 % TEMED	10-20 % Acryl/Bisacryl 39:1 375 mM Tris-HCl pH 8.8 0,1 % SDS 0,15 % APS 0,2 % TEMED
Sample loading buffer	*Loading buffer 5x* 35 % Glycerol 250 mM KOH-Acetate pH 6.8 Bromophenol blue	*Loading buffer 5x* 35 % Glycerol 125 mM Tris-HCl pH 6.8 Bromophenol blue	*Laemmli buffer 3x* 40 % Glycerol 200 mM Tris-HCl pH 6.8 8% SDS 2 M BME Bromophenol blue
Staining buffer	45 % Ethanol 10 % acetic acid 3 mM Coomassie blue (G-250)		
Destaining buffer	30 % Ethanol 10 % acetic acid		

3.2 Gel shift assay

This technique is also known as EMSA (ElectroMobility Shift Assay). It is a native continuous electrophoresis technique used to highlight DNA-protein interactions. It is based on the simple idea that a nucleic acid will have a different electrophoretic mobility if it's bound to a protein, compared with a negative control. Generally, a protein-bound DNA migrates slower because of its higher mass. This results in a shift on the gel. Even though the EMSA technique was initially developed to quantify DNA-protein interactions[610,611], it is used today rather as a qualitative method.

In order to achieve a sufficient resolution to see the difference of migration between the DNA-protein complex and the DNA alone, a polyacrylamide gel is generally used. Agarose gels find their application in the study of very large macromolecular complexes.

However, despite its numerous advantages, this techniques also has limitations, in particular when it comes to the stability of the DNA-protein complexes during electrophoresis. This stability can be inherent to the complex itself if it has a high dissociation constant or if the equilibrium between the bound and unbound forms is not reached. Besides, native electrophoresis, although generally achieved at 4°C, generates a major release of heat which can destabilise the interactions within the complex. Typically, migration thus occurs at low voltage (≤ 150 V, compared with 250 V for a denaturing gel) and low intensity (≤ 15 mA), for a maximal total power around 2 W.

Tris-Glycine polyacrylamide gels like those described in the previous chapter can be used for EMSA, but it is more common to use lower ionic strength solutions, to avoid as much as possible destabilisation of the complex. TBE (89 mM Tris-HCl pH 8.3, 89 mM boric acid, 2 mM EDTA) is thus a solution of choice.

Visualization of the gel can be done either by autoradiography using a radioactively labelled DNA, by fluorescence using a fluorescent labelled DNA, or by more classical ways such as toluidine blue staining, or for higher sensitivity, ethidium bromide.

3.3 Protein dosage

Several methods are available to dose protein, i.e., to determine their concentration. Most of them are based on spectrophotometric measurements, to reveal spectral or reactional properties of some amino acids. Biuret, Lowry, Kejdahl and Bradford among others, contributed to the development of these methods. In the lab, two different methods are commonly used: the Bradford protein assay and the UV absorbance at 280 nm.

3.3.1 Bradford protein assay

The Bradford assay is a colorimetric dosage using Coomassie blue (G-250) as reagent. This dye, in its cationic form, has an absorption maxima at 465 nm (absorption in the blue which gives its red-brown colour). When in contact with proteins, the dye interacts with arginine residues (and to a lesser extent, lysine, histidine, tryptophan, tyrosine and phenylalanine) and undergoes a structural change. Its anionic form has its absorption maxima shifted towards 595 nm (absorption in the yellow, which gives its blue colour). The intensity of absorbance is proportional to the quantity of dye bound to proteins, and by extension, to the concentration of the sample. However, as pointed out by Marion Bradford herself in her 1976 article, both 465 and 595 nm spectrum overlap when their intensity is too high. As a result, the absorption is only proportional to the concentration of the protein in a narrow range, from 0 to 2 mg/mL. Serial dilution are thus often required to achieve a relevant estimation of the protein concentration.

However, this method also has a great advantage compared with all other dosage methods, in being only slightly sensitive to interference caused by various agents in the sample (buffer, chaotropic agents, nucleic acids, etc.), with the exception of detergents and polyphenols.

The measurement is performed at 595 nm by diluting the 5X Bradford reagent (Protein Assay Dye Reagent, BioRad) in 800 µL of water. Between 1 and 50 µL of protein are then added, depending on the concentration. The latter can finally be determined using a pre-established calibration curve.

3.3.2 UV absorbance

This non-colorimetric method is based on the strong absorbency of aromatic residues, in particular, tryptophan, at 280 nm (Layne, 1957). The absorbance is measured by a spectrophotometer and is correlated to the protein concentration using the Beer-Lambert equation

$$A_{280nm} = -\log_{10}\frac{I}{I_0} = \varepsilon_{280nm}.\ell.c$$

with A, the absorbance at 280 nm, $\frac{I}{I_0}$, the transmittance of the sample, ε, the molar absorption coefficient at 280 nm in $L.mol^{-1}.cm^{-1}$ (easily calculable from the primary sequence of the protein, using tools like ProtParam (Expasy)), ℓ, the length of the sample to travel across in cm and c, the concentration in mol/L.

In the lab, we use a Nanodrop spectrophotometer (Thermo Scientific), which allows us to measure the absorption spectrum of a 1 µL drop of sample, in a range from 190 to 840 nm. The advantage of this machine lies in the fact that only 1 µL of sample is necessary to measure the concentration, and the wide spectral range allows measurement of a variety of sample types (proteins at 280 nm, but also peptides at 205 nm, nucleic acids at 260 nm, etc.)

The same technique is performed in chromatography on Äkta purification workstations: the sample, when exiting the column, passes through a 280 nm spectrophotometer. This allows the follow-up of the purification and to detect the protein fractions.

3.4 Mass spectrometry

Mass spectrometry is a physical technique used to detect and identify molecules of interest. Its particularity lies in the separation in a gas phase of the different components of the sample, based on their motion in an electric field. Prior to this, the sample must be ionized, and the charged molecules are analysed according to their mass/charge relationship (m/z).

The first mass spectrometry study dates back to 1912. Joseph John Thompson, who had been the first to experimentally demonstrate the existence of electrons 15 years earlier, was investigating the composition of a mix of positive ions, called “anodic ions”, today known as anions. He could measure the deflexion of a neon beam, ionized by an electric discharge, and subject to a magnetic field. The photographic plate he used as a detector was then irradiated in two distinct points, corresponding to two different neon entities, of different mass but identical charge: isotopes 20 and 22. The technique was then upgraded and the first mass spectrometer was developed by Thompson’s student, Francis William Aston, in 1919.

Typically, a mass spectrometer is composed of an ionisation source in which the sample is injected, followed by one or more analysers which separate the ionic entities according to their m/z ratio; finally, a detector counts the ions, and transforms this value into an electric signal with an intensity proportional to their number. A computer processing system allows interpretation of the results. According to the molecules analysed and the type of information sought, several ionisation sources (Electrospray, MALDI, etc.) and analysers (TOF, quadrupole, ion trap, etc.) exist.

I will describe only the MALDI-TOF spectrometer, which was the system mainly used for mass spectrometry studies during this work (*figure 48*). This technique allows the characterisation of the proteins extracted from an SDS-PAGE gel. In order to facilitate their identification, the protein bands are cut out of the gel, dehydrated in acetonitrile to remove all traces of Coomassie blue, and specifically cleaved using trypsin. The MALDI-TOF analysis permits us to deduce the primary amino acid sequence of the ionized peptides, for which each m/z value correspond to a defined amino acid sequence. By overlaying the determined sequences, it is then possible to get the full protein sequence.

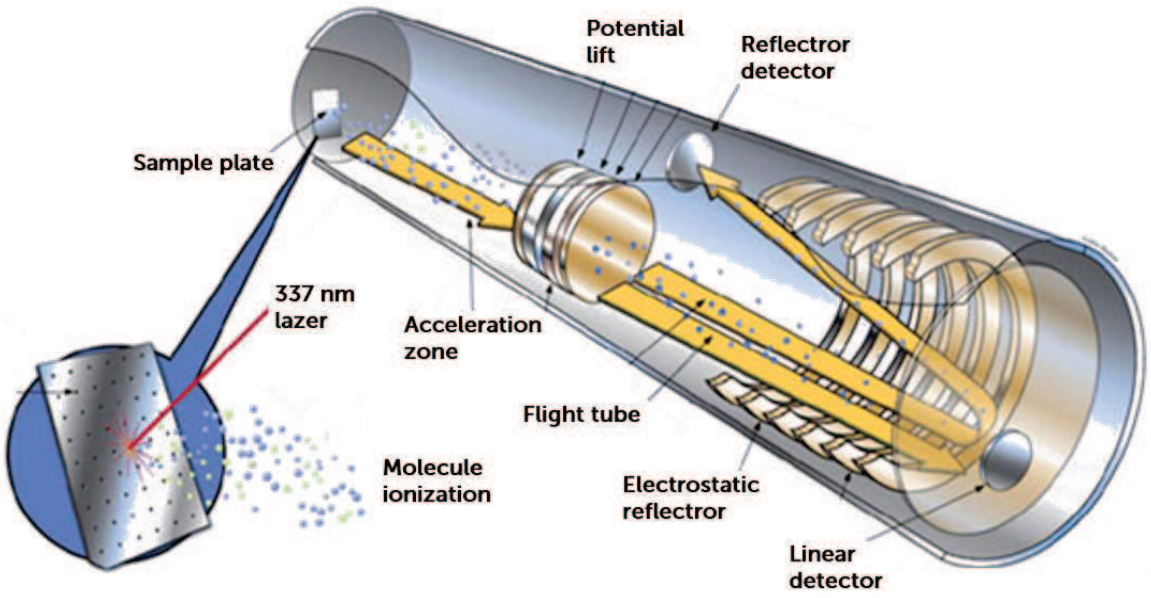


FIGURE 48

Mass spectrometry – MALDI-TOF

A MALDI-TOF spectrometer consists in an ionization source (a 337 nm laser, which ionizes particles embedded within an aromatic matrix) and a time-of-flight analyser. After ionization, particles, generally charged +1, are accelerated, and their velocity, as a function of their mass/charge ratio, is measured by a detector. The spectrometer can be equipped with a linear detector for a direct measurement, or, for a higher resolution, with a reflector detector, allowing lengthening of the flight zone without physically modifying the size of the spectrometer.

Modified from The Scientist, MALDI-TOF/TOF Mass Spectrometer, Jeffrey Perkel, April 12, 2004

3.4.1 Matrix assisted laser desorption ionisation (MALDI)

This soft ionization method, called MALDI utilizes a nitrogen laser beam to indirectly ionize the sample. The latter is mixed with a matrix, made up of a mix of aromatic acids (dihydrobenzoic, sinapic or cinnamic are among the most widely used), and crystallised on a metal cup. The laser beam at 337 nm is focused on the matrix/sample co-crystal and first ionizes the matrix, which then transfers then part of its charge to the peptides in the sample. This technique was imagined after the observation made in 1985 by Hillenkamp and Karas, that alanine was more easily ionisable if mixed with tryptophan, an aromatic amino acid. It was then shown that aromatic molecules had the capacity to absorb laser energy and to transfer it to non-absorbing molecules.

This technique has the advantage of protecting the sample from the disruptive energy of a direct laser beam, and to facilitate its vaporization and its quasi-molecular ionization (by loss or addition of a single proton).

3.4.2 The Time-Of-Flight analyser (TOF)

The TOF analyser allows the measurement of the time required for vaporized ions to travel a given distance to the detector. After ionization, the particles are accelerated in the acceleration area by a magnetic field of known strength, which furnishes a similar kinetic energy to all ions with the same charge. They travel through a field-free flight tube, in which their velocity depends only on the kinetic force acquired earlier, and thus, on their mass/charge ratio. Ions with a small m/z ratio arrive at the detector first.

The potential energy of a moving particle is given by the relationship

$$E_p = z.V$$

where z is the charge of the particle and V, its electrostatic potential. During acceleration of a particle in the TOF analyser, an electric potential difference or accelerating voltage, referred to as V_0, is observed. The potential energy E_p is then converted into kinetic energy E_k, because of the motion of the particles, according to the following relationship

$$E_p = z.V_0 = \frac{1}{2}m.v^2 = E_k$$

This simple equality contains the m/z ratio that we can directly deduce from the velocity v, in other words, from the time, t, an ion requires to travel a set distance, d, such as $v = \frac{d}{t}$, whence

$$z.V_0 = \frac{1}{2}m.\left(\frac{d}{t}\right)^2$$

$$t^2 = \frac{d^2}{2V_0}\frac{m}{z}$$

$$\frac{m}{z} = 2V_0\frac{t^2}{d^2}$$

The mass spectrometry studies were carried out by the proteomic common service of the IGBMC (Illkirch) and the platform “Protéomique-Strasbourg Esplanade” in the IBMC (Strasbourg).

3.5 Thermofluor®

In vivo, the stability of a protein depends on various parameters: the presence of protein or nucleic partners, concentration, post-translational modifications, physiological conditions, etc. *In vitro,* however, the issue is to determine simple conditions in which the protein can fold properly, to the detriment of *in vivo* parameters. To this end, Thermofluor®, also called Thermal Shift Assay or Differential Scanning Fluorimetry allows us to determine the thermal stability of a protein sample

(*figure 49*). This method was designed in 1997 by Pantoliano, initially to study the ligand-induced stabilisation of a protein.

Folding of a protein is a thermodynamic process, spontaneous when $\Delta G < 0$ or $\Delta S_{universe} = 0$, with G, the Gibbs free energy and S, the entropy. This is only valid in ideal conditions. In practice, there is a constant equilibrium between the native and the denatured form of proteins, which can be illustrated by the equation

$$\Delta G_d = G_d - G_n$$

with ΔG_d, the difference of free energy, G_d, the free energy of the denatured state and G_n, the free energy of the native state of the protein. The higher and positive ΔG_d is, the more stable the protein is. Gibbs free energy is composed of two terms, H, the enthalpy, and S, the entropy, depending on the temperature T of the system, such that

$$G = H - TS$$

The stability of the protein is thus inversely proportional to the temperature. The derivative of this equation can be written

$$\Delta G_d(T) = \Delta H_d(T) - T\Delta S_d(T)$$

Thus, when the temperature increases, entropy increases and Gibbs free energy decreases. In concrete terms, when the temperature increases, ΔG_d decreases and becomes null at equilibrium, when $T = T_m$ for which [native protein] = [denatured protein]. Proteins with higher T_m will thus be the more stable ones.

The principle of Thermofluor® is based on the detection of fluorescence emitted by SYPRO Orange, a fluorophore able to bind to hydrophobic regions of proteins. When SYPRO Orange is added to the protein sample, it is exposed to an aqueous environment and fluorescence emission is low. But as the temperature increases, the proteins undergo melting and reveal their hydrophobic patches. SYPRO Orange can then bind to these regions and become highly fluorescent (excitation: 473 nm; emission: 570 nm). Finally, at high temperature, the sample is fully denatured. It aggregates and precipitates, shutting down the fluorescence.

As part of that work, I used Thermofluor® to optimise buffer and pH conditions during purification. A condition screen was set up to seek the most favourable to the protein stability. Some biochemistry and structural biology experiments require a long incubation period, during which the protein has to remain in its native state. Plus, it has been shown that the use of conditions in which the protein is the most stable can facilitate crystallisation and improve the quality of the crystals[612,613].

Thermofluor® experiments were carried out in the lab, on a real-time-PCR thermocycler (MiniOpticon Real-Time PCR Detection System, BioRad). A standard protocol consists in mixing 20 µL of the solution to screen, 7 µL of SYPRO orange and 5 to 10 µg of protein, for a total volume of 30 µL. Up to 48 samples can be tested simultaneously. A gradient of temperature is achieved in the thermocycler, from 4 to 95°C, and measurements of the fluorescence are made every 0.5°C. The

fluorescence curve as a function of time gives an indirect follow-up of the sample denaturation, and the Tm can be determined to evaluate the protein stability in the condition tested.

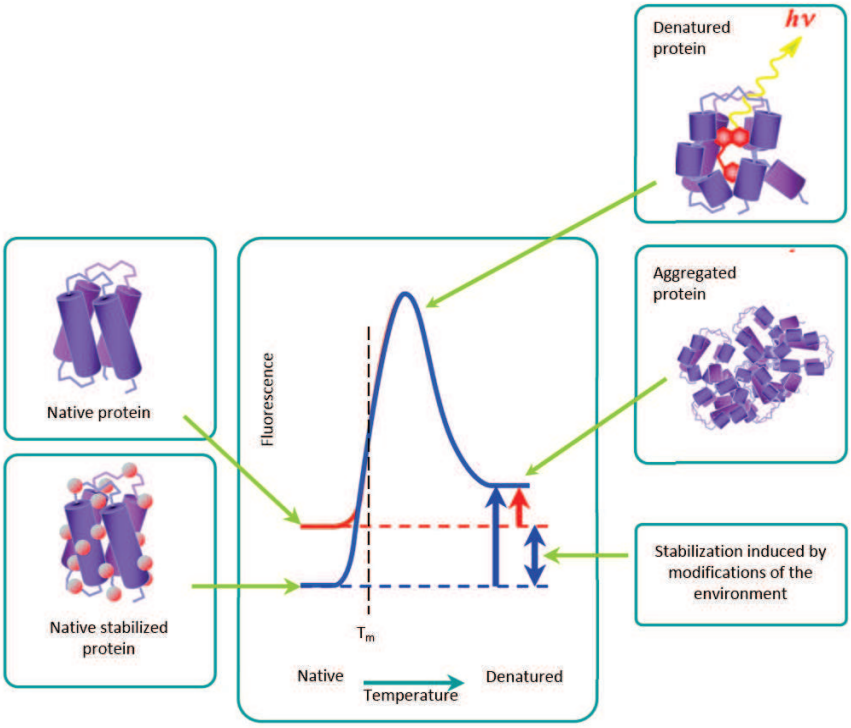


FIGURE 49

Differential scanning fluorimetry or Thermofluor®

Thermofluor® is based on the fluorescence emission of a molecule, Sypro Orange, when it interacts with hydrophobic patches of proteins. In native state, a protein does not exhibit hydrophobic patches and no fluorescence can be detected. However, as it undergoes denaturation, its unfolding allows binding of the fluorochrome and thus, fluorescence emission. Using a temperature gradient, it is thus possible to follow denaturation of a protein sample, and to test different conditions in order to define the protein stability.

Modified from http://www.beta-sheet.org/ (accessed September 06, 2014)

3.6 Dynamic light scattering

Dynamic light scattering or DLS, also known as Quasi-Elastic Light Scattering (QUELS) or Photon Correlation Spectroscopy (PCS), is one of the most popular techniques used to determine the size of particles. It consists in shining a monochromatic light beam, such as a laser, onto a solution containing particles in suspension. Elastic scattering, or Rayleigh scattering, is then observed when particles are hit by the laser. This scattering is correlated to the size of the particles, and by extension, to their rate of diffusion. A detector, placed at a given θ angle (generally 90°), measures the fluctuations of intensity of the light scattered over time (*figure 50*). These fluctuations can be quantified by a second-order correlation function such that

$$(g)^2(\tau) = \frac{\langle I(t).I(t+\tau)\rangle}{\langle I(t)\rangle^2}$$

where I is the intensity of the scattered light at time t, and τ, the delay time. In the case of a monodisperse sample, that is to say in which all particles are homogeneous in size and shape, the autocorrelation function is simply a single declining exponential defined by

$$(g)^2(\tau) = B + \beta e^{-2\Gamma\tau}$$

with B, the baseline of the correlation function at infinite delay (when $\tau \to \infty$), β, the correlation function amplitude at zero delay, and Γ, the decay rate. Using a nonlinear least squares fitting algorithm, it is possible to determine Γ from the experimental correlation function. It is finally possible to convert this value to the diffusion constant D, via the relationship

$$D = \frac{\Gamma}{q^2}$$

where q is the magnitude of the scattering vector given by

$$q = \frac{4\pi\eta}{\lambda}\sin\frac{\theta}{2}$$

with η, the viscosity of the sample and λ, the wavelength of the incident beam.

This experiment is based on two hypothesis: the first one is that particles in solution undergo Brownian motion. This effect is obtained when particles in solution collide which provokes their random motion. It is assumed that between two collisions, particles move in a straight line and at constant speed.

The second one is that particles are small and spherical (smaller than the laser wavelength). In that case, it is possible to apply the Stoke-Einstein relation, which interprets the Stoke radius or hydrodynamic radius of particles as a function of their diffusion D

$$r = \frac{\mathcal{R}}{6\pi\mathcal{N}_A}.\frac{T}{\eta D} = \frac{k.T}{6\pi\eta D}$$

where r is the radius of the particle, R is the ideal gas constant, $\mathcal{N}_A$ Avogadro's constant, T, the temperature, and k, the Boltzmann constant. In this relation, one can mention that the scattering is inversely proportional to the size of the particles (r) and the viscosity of the sample (η). The larger the particle or the more viscous the sample, the slower the motion.

These measurements allow us to determine the polydispersity of a proteinaceous sample, that is to say, the standard deviation of the size distribution of a protein mixture. For protein crystallisation, a low polydispersity is required, and it is assumed that a polydispersity index higher than 20 % reduces the probability of nucleation and crystal growth.

In the lab, we use a DynaPro DLS spectrometer (Wyatt) to analyse samples concentrated to 1 mg/mL in 50 µL disposable plastic cuvettes or in 3 µL quartz cuvettes. It includes a red laser diode (λ = 675 nm) as a light source, and a monomode optic fibre (Single mode fibre, SMF) to receive the light scattered perpendicular to the source (θ = 90°). Prior to measurement, the sample must be ultracentrifuged to remove all traces of aggregate or dust.

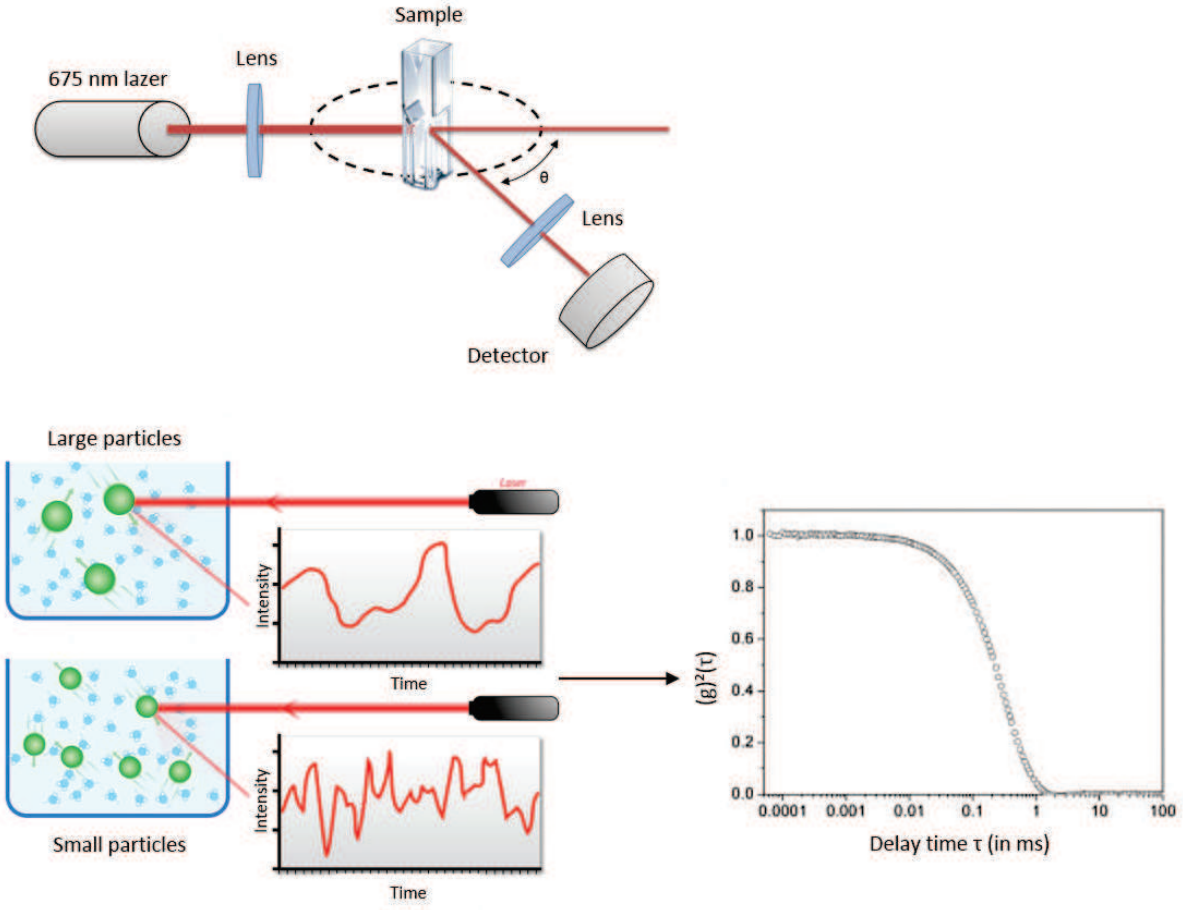


FIGURE 50

Dynamic light scattering or DLS

Top: the DLS spectrometer principle. It consists in a monochromatic laser source (λ = 675 nm), directed towards the sample. A detector, set at a defined angle θ (generally, 90°) measures the variations of elastic light scattering intensity.

Below: typical experimental curve. One can note that small particles produce important diffusion. This is illustrated by a rapid variation of the intensity over time. From this experimental curve I=f(t), a two-order correlation function allows quantification of these variations, and thus, determination of the diffusion constant and averaged size of the particles.

Modified from "DLS" by Mike Jones, Wikimedia Commons

3.7 Analytical ultracentrifugation

In the 1920s, even though the importance of biological macromolecules like proteins or nucleic acids had already been proven, their existence was still controversial. The assumption was that proteins were an assembly of tiny molecules, of undefined mass, aggregating together intermittently and reversibly, to fulfil a biological role. Theodor Svedberg then imagined a technique, in 1923, to sediment particles through an artificial force produced by a centrifuge, and to follow this sedimentation over time via an optical device. This first ultracentrifuge, built in Uppsala (Sweden), offered an acceleration of 5000 Xg. These first experiments of Svedberg confirmed the existence of proteins as intrinsically stable macromolecules.

When particles in suspension are subject to a gravitational field, three forces are exerted upon the particles:

- The first force is the sedimentation, or gravitational force, F_s, which is proportional to the molecule mass and the acceleration. In a centrifuge, acceleration is determined by the distance r between the sample and the rotor axis, and by the angular velocity or rotational speed ω, in rad/s, such that

$$F_s = m\omega^2 r = \frac{M}{\mathcal{N}_A}\omega^2 r$$

  with m, the mass of the sample, M, the molecular mass and $\mathcal{N}_A$, Avogadro's constant.

- The second force is buoyancy F_b, resulting from the Archimedes' principle, which is equal to the mass m_0 of volume displaced by the particle, such that

$$F_b = -(m_0)\omega^2 r = -(m\bar{v}\rho)\omega^2 r = -\left(\frac{M}{\mathcal{N}_A}\bar{v}\rho\right)\omega^2 r$$

  where ρ is the density of the solvent, in g/mL, and $\bar{v}$, the volume that each gram of particle occupies in the solution, also named the specific partial volume. This partial volume is fairly constant, varying by around 1%, for each type of biological macromolecule: 0.73 mL/g for proteins; 0.58 mL/g for DNA; 0.53 mL/g for RNA.

- Finally, assuming that the particle of interest has a higher density than the solvent, it will sediment. As it travels in the more or less viscous solvent, it is subject to a friction force F_f, proportional to its sedimentation speed u, such that

$$F_f = -fu$$

  with f, the friction coefficient, which depends on the shape and the size of the particle. Large and elongated particles undergo more friction force than small and spherical particles.

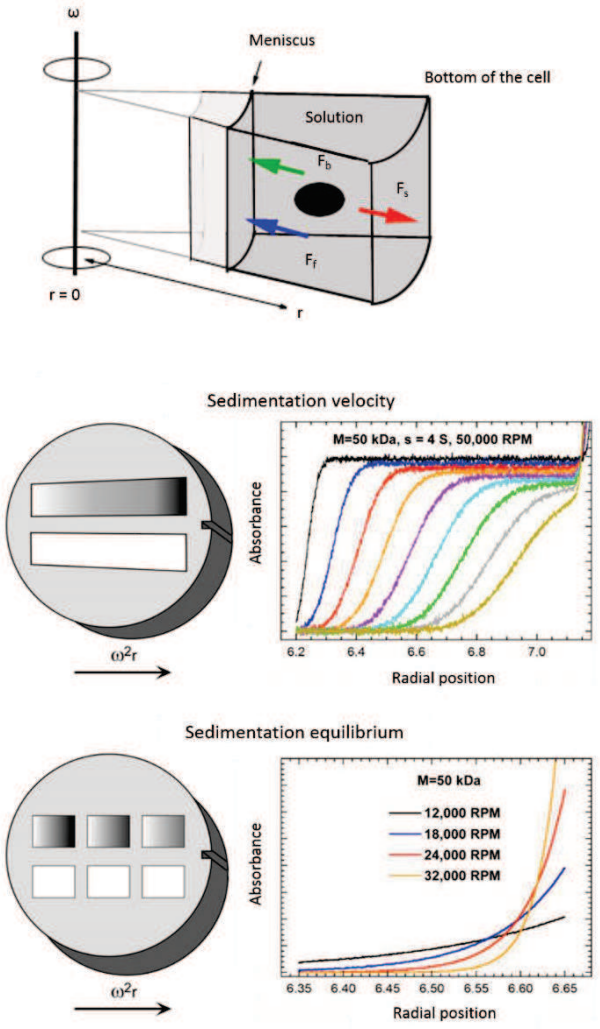


FIGURE 51

Analytical ultracentrifugation or AUC

During centrifugation, three forces are applied on particles in solution: sedimentation force F_s, buoyancy force F_b and friction force F_f. The optically transparent cell is crossed by a light source, allowing a direct measurement of absorbance over time.

When studying sedimentation velocity, the cell is subject to a very high centrifuge force. Absorbance is measured at several moment t over the total length of the cell, allowing a follow-up of the protein front migration.

When studying sedimentation equilibrium, the cell is subjected to a low centrifuge force. Absorbance is measured only once, after 16 hours of continuous centrifugation, when the sedimentation/diffusion equilibrium is reached. Different concentration gradients can be analysed in the different cell chambers.

Modified from James L. Cole et al., Analytical ultracentrifugation: sedimentation velocity and sedimentation equilibrium, Methods Cell Biol. 2008 ; 84: 143–179.

Rapidly, these three forces equilibrate such that $F_s + F_b + F_f = 0$, whence

$$\frac{M}{\mathcal{N}_A}\omega^2 r - \left(\frac{M}{\mathcal{N}_A}\bar{v}\rho\right)\omega^2 r - fu = 0$$

By rearranging the terms related to the particle of interest, and those related to the experimental conditions, one can write

$$\frac{M(1-\bar{v}\rho)}{\mathcal{N}_A.f} = \frac{u}{\omega^2 r} \equiv s$$

This new term "s", standing for svedberg, is called the sedimentation coefficient, expressed in Svedberg (S) such that 1 S = 10^{-13} seconds.

In the lab, we use an Optima ProteomeLab XL-I ultracentrifuge (Beckman Coulter). It is able to reach 60000 rpm (revolutions per minute), generating a gravitational force of about 250000 Xg on the sample. The specially designed rotors for analytical experiments include an optically transparent quartz or sapphire chamber, which can contain around 100 µL of sample. The optical detection device is synchronized according to the rotation of the sample in order to acquire the data when the sample passes in front of the light source. Typically, it is composed of a UV/visible spectrophotometer, equipped with a monochromator and an interferometer. The wavelength of the monochromator can be adjusted in accordance with the sample being analysed: 280 or 230 nm for proteins; 260 nm for nucleic acids; in the visible range for chromophores, etc.

Several applications are possible in analytical ultracentrifugation, but I will only describe hereafter the two main ones: the sedimentation velocity and the sedimentation equilibrium (*figure 51*).

- The study of sedimentation velocity is based on the difference of migration of a mixture of particles in a high centrifugal field (generally, ω ≥ 60000 rpm). The motion over time of the sedimentation boundaries, specific to each molecular species, is measured by detection of their absorbance. The experimental data obtained are a set of curves that plot the evolution of the OD along the sedimentation axis (from the rotor axis to the bottom of the sample chamber) at several time points.
  At the beginning of the experiment, the sample concentration is uniform. Then three areas quickly are formed: the first one at constant protein concentration (called the protein front, with a high and stable OD); an area at gradually decreasing concentration (the intermediate boundary, with an increasing OD); and an area of pure solvent (the supernatant area, with an OD ≈ 0). The sedimentation velocity is then determined by the distance r travelled by the intermediate boundary over time. This distance is experimentally illustrated by the inflexion point of each curve $OD = f(r)$. The graphical representation of $\ln r = f(t)$ allows the linear equation to be written as $\ln r = S\omega^2 t + cst$ and to determine S. This value can be compared to the theoretical value obtained by the Svedberg relation (but which requires previous knowledge of the Stoke radius of the particle of interest), and to define the oligomerisation state of the sample.

- The second application of analytical ultracentrifugation is the study of the sedimentation equilibrium. This measurement occurs at reduced speed (generally, ω ≤ 10000 rpm) and requires several hours to reach the needed equilibrium. Typically, in a weak centrifugal field, particles sediment gradually to the bottom of the chamber, but are subject to an opposite force, called diffusion. After several hours of centrifugation, both sedimentation and diffusion forces reach an equilibrium, so that the concentration of the particle along the r axis of sedimentation becomes a perfect exponential, time-invariant. The measurement of the concentration at several spots along the r axis allows us to directly deduce the molar mass of the particle, via the relation

$$M = \frac{2RT}{(1-\bar{\nu}\rho)} \times \frac{d(\ln c)}{dr^2}$$

where c is the concentration of the particle at a given r distance.

3.8 Isothermal titration calorimetry

ITC is a technique that allows to measure the thermodynamic characteristics of the interaction between two molecules. This interaction is a process that can be either exothermic (generating heat) or endothermic (absorbing heat). ITC measures this difference of heat at the time of the interaction and permits to experimentally deduce the change of enthalpy ΔH, the constant of association K_a and the stœchiometry n (*figure 52*). It is thus possible to calculate from the these values the free energy ΔG and the entropy ΔS, via the relationship

$$\Delta G = -RT\log(K_a) = \Delta H - T\Delta S$$

with R, the ideal gas constant and T, the temperature in °K.

In the lab, we use the MicroCal ITC200 (Malvern), which consists in an adiabatic chamber, i.e., without any possible heat transfer between the inside and the outside, and in which are two cells, one for the reference buffer solution and one for the sample. The first molecule, marked A, is placed in the 300 µL sample cell, while the reference buffer is placed in the reference cell. The second molecule, marked B, is then injected little by little (1-2 µL for each injection) in the sample cell using an automatic syringe. For each injection, the machine measures the required energy to keep the sample cell and the reference cell at the same temperature, by application of a weak electric current.

After about twenty injections of the B molecule (with a concentration around 10-fold higher than A), saturation is, theoretically, reached. The curve obtained plots the energy in µcal/s delivered by the machine over time. The profile is composed of a set of maxima or minima (whether the reaction is endo- or exothermic), corresponding to each injection of B. The time and slope for the baseline return after each injection peak gives an information on the kinetic, i.e., the speed of reaction. Furthermore, the integral of each of these peaks corresponds to the interaction-associated heat (Q), in kcal/mole. This value decreases steadily as the system reaches saturation, the molecule A becoming then unavailable for further interactions. The heat which is then generated or absorbed

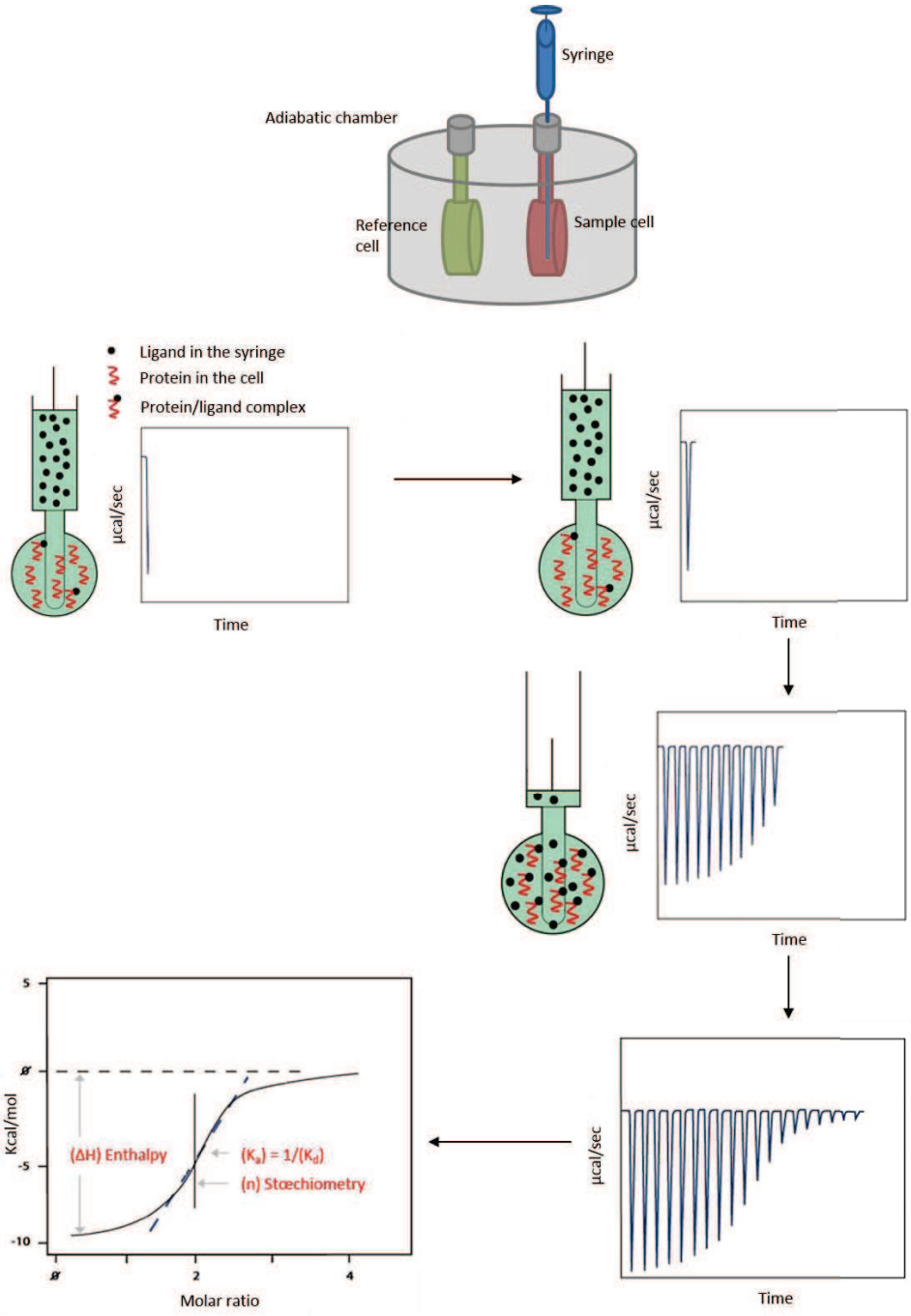


FIGURE 52

Isothermal titration calorimetry or ITC

Top: schematic representation of a microcalorimetry device.

Below: example of ITC measurements, in the case of an exothermic interaction. The µcal/sec curve decreases over time, indicating that the saturation has been reached. This raw experimental curve can be integrated and the Kcal/mol curve over the molar ratio allows to determine the interaction K_d, stœchiometry and differential enthalpy.

Modified from http://www.huck.psu.edu/facilities/calorimetry-up/guides/itc (accessed June 06, 2014)

hence corresponds only to the thermodynamic variations caused by the mix of the molecules, without any interaction.

The amount of heat Q can be defined as follows

$$Q_j = v\Delta H[AB]_f$$

where v is the volume of the sample cell and $[AB]_f$, the final concentration of the complex formed during injection j. The titration curve can then be expressed by Q over the molar ratio [B]/[A]. This new curve can finally be analysed to determine the constant of association K_a at the inflexion point of the curve, as well as the stœchiometry. The constant of dissociation K_d, which is generally used to characterise the binding affinity of two molecules, corresponds to $1/K_a$. Finally, ΔH corresponds to the amount of heat generated or absorbed during the first injection.

4. Structural biology methods

During my PhD, two major and complementary techniques of structural biology have been employed.

X-ray crystallography is a method allowing to determine the 3-D structure of a biological macromolecule at atomic scale (from 1 to 3-4 Å). This technique is probably the most resolutive, but its implementation requires to grow crystals of the molecule of interest. This step, called crystallisation, consists in bringing the molecule from a liquid state (in solution) to a solid state (the crystal), perfectly organized. However, this phase transition can be accompanied by more or less important modifications of the 3-D structure of the molecule. Dynamic information are, for example, lost, and the single 3-D structure determined might, in some cases, not accurately reflect the biological truth. This is especially the case for inter-domains conformations, which are often flexible.

A second technique was thus employed: single-particle cryo-EM. This transmission microscopy technique allows determining a 3-D map of the electronic density of a macromolecule. Unlike crystallography, this technique provides the advantage of working on samples frozen in a thin layer of vitreous ice, i.e., fixed in their native hydrated state, without having to grow crystals. However, it offers only a medium resolution range (although with current means, it comes close to the atomic resolution in some cases) and is only suitable for large macromolecules and complexes (> 100-150 kilo Daltons).

4.1 Structural study by X-ray crystallography

4.1.1 History

The history of X-ray crystallography begins in 1895. At that time, Wilhelm Röntgen decided to take over the work of Julius Plücker, initiated 40 years earlier in Bonn, Germany. This German physicist had discovered cathode rays, an unknown substance that travels in a straight line from a cathode to an anode, when placed in a tube filled with “rarefied air” (i.e., under vacuum) when an electric current is applied. These rays produce a green luminous glow, namely phosphorescence,

when they reach the wall of the tube opposite to the cathode, and can furthermore be deviated with a magnet. Without knowing it, Julius Plücker had revealed the existence of electrons, of which Joseph Thompson would measure the charge in 1897.

Röntgen took over the already advanced studies on cathode rays, and could incidentally observe a luminous glow reflecting on a barium platino-cyanide screen placed 2 m away from the glass tube, each time a discharge was applied. Philipp Lenard had previously demonstrated that cathode rays were rapidly dispersed in the thickness of glass and in air. Röntgen therefore assumed that it should be something else, the unknown X, a matter that travels in a straight line from the spot where the cathode rays hit the glass tube wall. He named this matter "X-rays", rays that are absorbed by solid substances but much slower than cathode rays. He demonstrated the properties of these X-rays, and first took a picture of brass weights enclosed in a wooden box; and later, the very first radiography of his wife's hand, on the 22nd of December, 1895.

Quickly, X-rays became popular, not only in radiology, a new medical speciality, but also in everyday life: in funfairs, where everyone wanted to have a picture of their skeleton, in orthopaedic shops, etc. It was not until 1925 that the first recommendations regarding the use of ionizing radiations were published.

In parallel to their medical development, X-rays were rapidly put to good use for crystal studies. Originally, crystallographic studies were only descriptive. The ancient Greeks thought that a crystal came from the freezing of a material, whence the name crystal, from the ancient Greek κρύσταλλος/krýstallos, the ice. In 1781, René-Just Haüy defined mineralogical species (crystals) as "*une collection de corps dont les molécules intégrantes sont semblables par leurs formes et composés des mêmes principes unis entre eux dans le même rapport*", that is to say, a set of bodies of identical geometry, and made of the same constituents in identical proportions. Auguste Bravais endorsed this idea in 1848 and described the 14 Bravais lattices, illustrating the 14 different types of 3-D crystal arrangements possible (*figure 53*). He made the following postulate: "*Étant donné un point P, quelconque dans un cristal, il existe dans le milieu, une infinité discrète, illimitée dans les trois directions de l'espace, de points autour desquels l'arrangement de la matière est la même qu'autour du point P*", i.e., "given any point P of the system, there is in the medium, a discrete, unlimited in the three directions of space, number of points around which the arrangement of the matter is identical, with the same orientation".

In 1912, Max von Laue demonstrated thus that copper sulphate crystals could diffract X-rays, and that an interference pattern could be photographed on a silver film. He was however unable to interpret this pattern. One year later, William Lawrence Bragg, then 23, suggested a simple relationship allowing the description of the crystal structure according to its interference pattern. He took, for this, the opposing idea to his father's, William Henry Bragg, who supported the idea that X-rays were not electromagnetic waves but rather particles. According to W. L. Bragg, X-rays, if waves, must be deviated from their trajectory (diffusion) when they hit an electron (the diffuser). If we consider that in a crystal, all the diffusers are regularly arranged, the resulting diffusion should also be regular. Figure 54 illustrate this theory. Considering two waves (rays 1 and 2), spaced by a distance equivalent to λ, and each reflecting from two different planes (planes 1 and 2), we can define the path difference δ via the relation $\delta=BD+DC$. This path difference corresponds to the difference of the distance travelled by ray 2 compared to ray 1. And since $BD=DC=d\sin\theta$, the relation can be rephrased $\delta=2d\sin\theta$. However, interferences are only constructive when the reflections are in phase, which is only true when δ is a multiple of λ. Bragg thus formulated his law as follow:

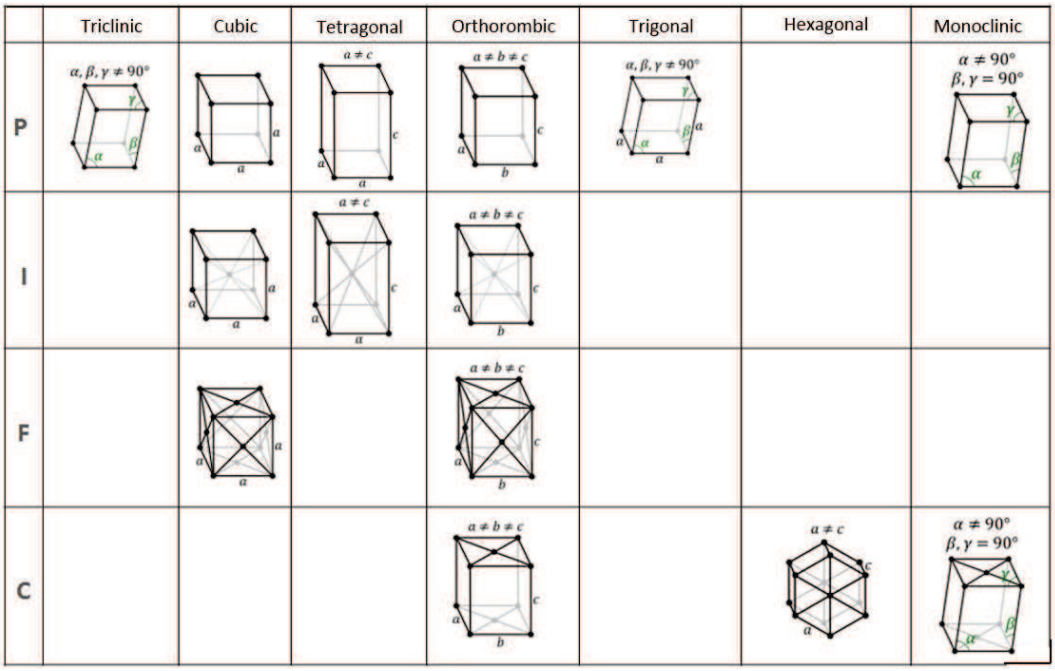


FIGURE 53

The 14 Bravais lattices

In 1848, Bravais described seven geometric networks allowing a regular and periodic distribution of "nodes" in a crystal, i.e., atomic or molecular patterns. To each of these seven crystal family can be added four different modes:

- Primitive (P), where the nodes are on the cell corners only;
- Volume-centered (I), where the nodes are on the cell corners and at the centre of the cell;
- Face-centered (F), where the nodes are on the cell corners and at the centre of each face of the cell;
- Base-centered (C), where the nodes are on the cell corners and at the centre of each of one pair of the cell faces;

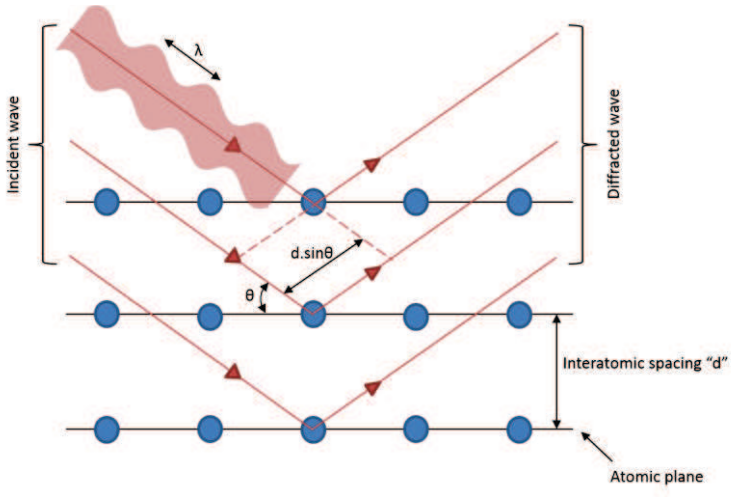


FIGURE 54

Bragg's law

Demonstration of Bragg's law ($2d \sin\theta = n\lambda$). Considering the interaction of two parallel rays with two atoms, located on the same straight-line perpendicular to the surface. The path difference δ travelled by the "deeper" ray is 2dsinθ. Interferences are constructive only if this path difference introduces a phase shift which is a multiple of 2π, i.e., if the path difference is a multiple of λ.

$$2d \sin\theta = n\lambda$$

with d, the space between two crystallographic planes; θ, Bragg's angle or half-scattering angle; n, the order of diffraction; and λ, the X-ray wavelength. This law permitted him and his father to solve the very first X-ray structure of table salt or sodium chloride. Numerous other structures followed, for example diamond, calcium fluoride, calcite, pyrite, spinel, etc. It was not until 1953 that the structure of the DNA was solved[614] and 1958 for the first X-ray structure of a protein, that of sperm whale myoglobin[615].

4.1.2 Principle

A crystal is an ordered and periodic arrangement (also called a lattice) of molecules in the three directions of space. It can be described by its unit cell of vectors $\vec{a}, \overrightarrow{b}, \vec{c}$, characterised by the length a, b and c of the cell edges and the angles α, β and γ between them. These six parameters allow us to define the 14 different crystal systems (the Bravais lattices) and 230 space groups, corresponding to the different symmetries that can be found in the crystal lattice.

The translation of the primitive cell via the relation $\vec{t} = m\vec{a} + n\vec{b} + p\vec{c}$, with m, n and p, whole numbers, allows us to rebuild the crystal. The volume of a cell corresponds to

$$V = abc\sqrt{1 - \cos^2\alpha - \cos^2\beta - \cos^2\gamma + 2\cos\alpha\cos\beta\cos\gamma}$$
$$= \vec{a} \cdot \left(\overrightarrow{b} \wedge \vec{c}\right) = \vec{b} \cdot \left(\overrightarrow{c} \wedge \vec{a}\right) = \vec{c} \cdot \left(\overrightarrow{a} \wedge \vec{b}\right)$$

However, proteins are not tightly stacked within the crystal and a non-negligible proportion of the latter is occupied by the solvent. This proportion can be calculated from the Matthews coefficient (Vm)[616]

$$V_m = \frac{Cell\ volume\ (\text{Å}^3)}{Cell\ content\ mass\ (Dalton)}$$

This value ranges typically between 1.66 and 4 $\text{Å}^3.\text{Da}^{-1}$ as observed for solved X-ray structures in the PDB. Out of this value, it is possible to deduce the percentage of solvent content in the crystal (generally, between 30 and 75 %) via the relation

$$\%S = 1 - \frac{1{,}23}{V_m}$$

Within the crystal lattice, an infinity of parallel planes (nodal planes or atomic planes), regularly spaced, and intercepting three nodes of the lattice, can be described. These planes are considered to be responsible for the crystal diffraction, and are defined in space by a three-value notation called the Miller index, and marked (hkl). Taking the edges of the unit cell as reference vectors, the vector coordinates of each atomic plane constitute the Miller index. Plus, the space between two adjacent planes, also called interplanar spacing, is written d_{hkl}. It is thus possible to rewrite Bragg's law in accordance with this new nomenclature, such that $2d_{hkl} \sin\theta = n\lambda$.

The principle of diffraction can be interpreted as a geometric construct suggested by Paul Ewald in 1921, known as the Ewald's sphere (*figure 55*). It allows us to graphically determine the nodes of the reciprocal lattice (by opposition to the direct lattice), that is to say, the set of atomic planes which give rise to diffraction. A set of atomic planes (hkl) in the direct lattice (the "real" lattice of the crystal) produces hkl reflections in the diffraction pattern. This pattern of the direct lattice is, itself, a lattice, called the reciprocal lattice, and whose dimensions are inversely proportional to the dimensions of the direct lattice. It is defined by the vectors $\overrightarrow{a*}, \overrightarrow{b*}, \overrightarrow{c*}$, characterised by the length a*, b* and c* and the angles α*, β* and γ*, such that a*=1/a, b*=1/b and c*=1/c. The Ewald's sphere, centred on the crystal M, has a radius of 1/λ. When a plane monochromatic wave, defined by its wave vector $\overrightarrow{s_0}$, interacts with the crystal, it is scattered in all directions of space. An hkl reflection can then be observed when a wave, scattered at an angle of 2θ, with a wave vector $\vec{s}$, coincides with the node P of the reciprocal lattice located on the Ewald's sphere. By rotating the crystal within the X-ray beam, the Ewald's sphere remains fixed, whereas the reciprocal lattice moves together with the direct lattice, producing new reflections at each intersection with the sphere.

A unit cell with volume V is characterised at any point (x, y, z, with vector $\vec{r}$) by an electron density ρ($\vec{r}$). For a given scattering direction $\vec{s}$, the amplitude and the phase of the wave can be illustrated by a structure factor F, over the scattering vector $\vec{S} = \vec{s} - \overrightarrow{s_0}$ so that

$$F(\vec{S}) = \int_V \rho(\vec{r}) e^{i2\pi(\vec{S}|\vec{r})} d^3r$$

In the case of a crystal, each hkl reflection, corresponding to all the nodes of the reciprocal lattice located on the Ewald's sphere, can be characterised by a structure factor marked F_{hkl}. The full set of structure factors allows the calculation of an electron density function, reflecting the electron distribution in the crystal lattice. The maxima of density are then interpreted in terms of atomic positions. This electron density, at a given point with coordinates (x,y,z) in the lattice, is defined by the relation

$$\rho(x,y,z) = \frac{1}{V} \sum_{h=-\infty}^{+\infty} \sum_{k=-\infty}^{+\infty} \sum_{l=-\infty}^{+\infty} F_{hkl}\, e^{-2i\pi(hx+ky+lz)}$$

with F_{hkl}, the structure factor for a given hkl reflection, defined by an amplitude and a phase such that

$$F_{hkl} = |F_{hkl}| e^{i\phi_{hkl}}$$

where $|F_{hkl}|$ is the modulus and ϕ_{hkl} the phase of the structure factor for a given hkl reflection, whence

$$\rho(x,y,z) = \frac{1}{V} \sum_{h=-\infty}^{+\infty} \sum_{k=-\infty}^{+\infty} \sum_{l=-\infty}^{+\infty} |F_{hkl}| e^{i\phi_{hkl}}\, e^{-2i\pi(hx+ky+lz)}$$

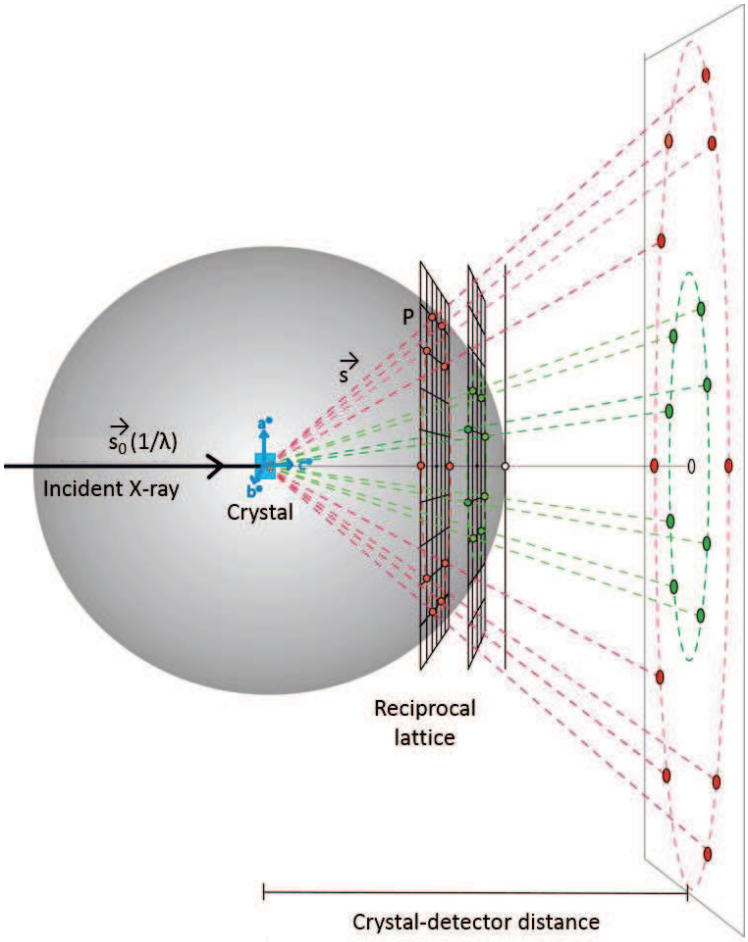


FIGURE 55

Ewald's sphere

When a plane monochromatic ray, defined by its wave vector $\overrightarrow{s_0}$, interacts with a crystal, it is scattered in all directions of space. For a given direction, characterized by its wave vector $\vec{s}$ and angle 2θ with the incident ray, an hkl reflection can be observed if a node P in the reciprocal lattice is located on the Ewald's sphere, centred on the crystal and of radius 1/λ.

The structure factor F_{hkl} is directly linked to the position and the nature of the atoms found in the lattice, via the relation

$$F_{hkl} = \sum_{j=1}^{N} q_j . f_j . w_j^{hkl} . e^{2i\pi(hx_j+ky_j+lz_j)}$$

with q_j, the occupancy factor of the j^{th} atom with relative coordinates (x_j,y_j,z_j), f_j, the scattering factor or form factor of atom j, and w_j, the atomic Debye-Waller factor of atom j. The scattering factor translates the interaction of the electronic cloud of an atom with the incident X-rays. In concrete terms, it is the Fourier transform of the electron density surrounding the atom. The occupancy and Debye-Waller factors refer, for their part, to the static disorder and motion of each atom.

Experimental diffraction data provide an intensity value for each hkl reflection. This intensity is proportional to the square of the modulus of the structure factor, in the case of a low mosaicity crystal, such that

$$I_{hkl} = k.(|F_{hkl}|)^2 = k.F_{hkl}.F^*_{hkl}$$

with F^*_{hkl}, the conjugate complex of F_{hkl}. Finally, the phase information can be seen as the depth through the unit cells. However, recording equipment only allows the measurement of the total intensity throughout time, i.e., the amplitude of each reflection, and the phase information ϕ_{hkl} is lost. Determination of this phase thus presents a major challenge in crystallography, and techniques have been developed to solve this problem:

- Multiple isomorphous replacement (MIR) consists in soaking the crystals in a solution containing heavy atoms (such as mercury, gold, lead or platinum) and to compare the diffraction patterns of the native crystal and the heavy atom derivative, to reveal the position of these heavy atoms in the unit cell.
- Single or multiple wavelength anomalous dispersion (SAD or MAD) is based on the inelastic scattering observed with heavy atoms whose absorption edges are close to the X-ray energy (from sulphur and above; selenomethionines are generally incorporated directly in the protein in place of methionines). The excited electrons then create anomalous diffraction; Friedel's law on centrosymmetry is no longer true, and the intensity of hkl and -h-k-l reflections are no longer equal.
- Molecular replacement is based on the usage of the phases of a model with high sequence identity (above 30 % in general) towards the protein of interest, to solve its structure factors. During my PhD work, this option has been favoured, since partial atomic structures were already available in the PDB.

Once the phases are known, it is possible to calculate the electron density map ρ(x,y,z). This map allows, in accordance with the diffraction limits and the accuracy of the phases, the determination of the atomic structure of the molecule.

4.1.3 Biological macromolecules crystallisation

For any given molecule, four areas corresponding to the different "phases" of the sample can be presented as a phase diagram (*figure 56*):

- the solubility zone, in which the molecule is undersaturated. It remains stable in solution.
- the clear or metastable zone, where the molecule is slightly supersaturated. Hydrodynamic constraints are such that they favour precipitation or microcrystal formation, but the kinetics are too slow to allow crystal growth.
- the labile or nucleation zone, in which the molecule is supersaturated. In this zone, thermodynamic and kinetic constraints are ideal to favour crystal growth.
- the precipitation zone, where the molecule is highly supersaturated. As its name implies, the conditions in this zone cause amorphous precipitate formation.

The solubility and the metastable zones are separated by the saturation curve, the "ideal" equilibrium condition in which a crystal added to the solution will neither grow nor dissolve.

The aim of crystallisation is to ensure the transition between the soluble state of a molecule to a solid state in the formed of ordered crystals. Several parameters can be altered to get optimal crystallisation conditions, illustrating the empirical nature of this step. Indeed, each variation of a given parameter causes a change in the phase diagram, more or less conducive to crystal growth. Among these parameters, one can mention:

- the concentration of the molecule of interest
- the purity of the sample (absence of contaminants, of micro-heterogeneity, etc.)
- the nature and pH of the buffer solution
- the nature and concentration of the additives (salts, etc.)
- the nature and concentration of the crystallisation agents
    - Polymers, such as PEG (polyethylene glycol), act by solvent exclusion in the vicinity of the protein;
    - Organic solvents (MPD, isopropanol, dioxane, etc.), which decrease the dielectric constant of the environment;
    - Non-chaotropic salts, which modify the ionic strength and increase hydrophobic interactions by solvent exclusion;
- the temperature
- convection forces
- vibrations, etc.

Several crystallisation techniques are used in the lab: sitting and hanging drop vapour diffusion; counterdiffusion in capillaries; etc. I will further describe sitting drop vapour diffusion only, which was widely employed during my work (*figure 57*).

In a closed chamber, an equilibrium is established by vapour diffusion between a reservoir containing a large volume (50 µL) of crystallisation solution (a mix of salts, buffer, crystallising agents, etc.), and a drop (from 200 to 800 nL in general), containing a mixture of the protein solution and the crystallisation solution. The volatile species in the drop (mainly water) thus diffuse in this closed chamber until the same vapour pressure is reached in both the reservoir and the drop. The volume of the latter will thus reduce, leading to a gradual concentration of both the protein and the crystallisation agent. If the conditions are favourable, the protein will enter the nucleation zone and organize themselves to form crystals.

As stated above, crystallisation of a molecule is an empirical process, which does not satisfy any general rule. In order to find the right crystallisation condition, a large number of parameters must be tested. In the lab, the “structural biology and genomics” platform provides about twenty different commercial screens, making it possible to test over 2000 different conditions. These screening conditions have been developed from statistical data collected in the PDB.

Screening is performed by the crystallisation robots Mosquito (TTP Labtech) or Cartesian Honeybee 8+1, setting sitting drops of 100 to 500 nL. The plates used are generally MRC 2 or MRC 3 (Swissci), and are stored at 4°C or 20°C in automated imaging systems (Rock Imager - Formulatrix). On a regular basis, the system browses the plates and takes pictures of the drops, in visible and/or UV light. This allows a regular follow-up of the images, and, if applicable, of the crystal growth.

To facilitate partially unfolded protein crystallisation, *in situ* mild proteolysis techniques were used. This technique is based on the idea that a native domain of a protein will crystallise more easily than a full-length protein exhibiting unfolded regions. The proteases used will generally cleave

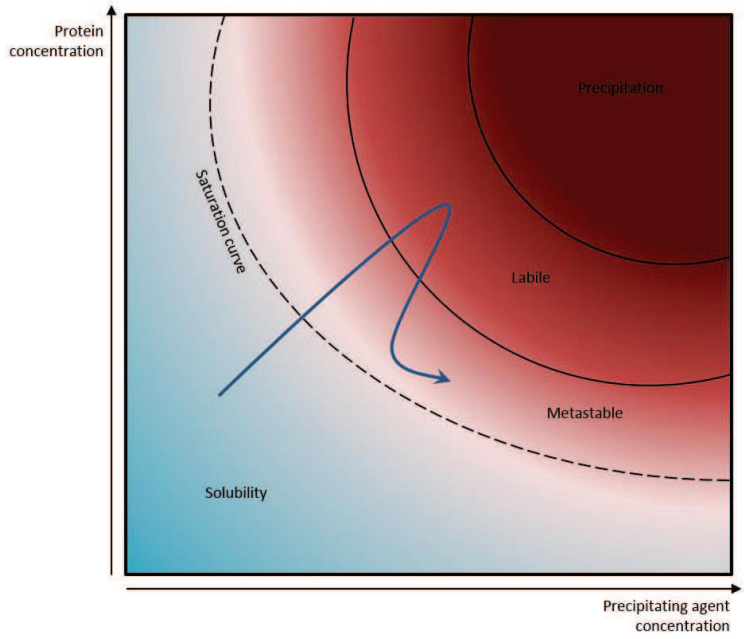


FIGURE 56

Phase diagram

Simplified representation of a macromolecule 2-dimensionnal phase diagram. Solubility of a protein depends on its concentration as well as the concentration of the precipitating agent. The blue arrow shows the ideal progression of a molecule through this phase diagram in order to obtain crystals.

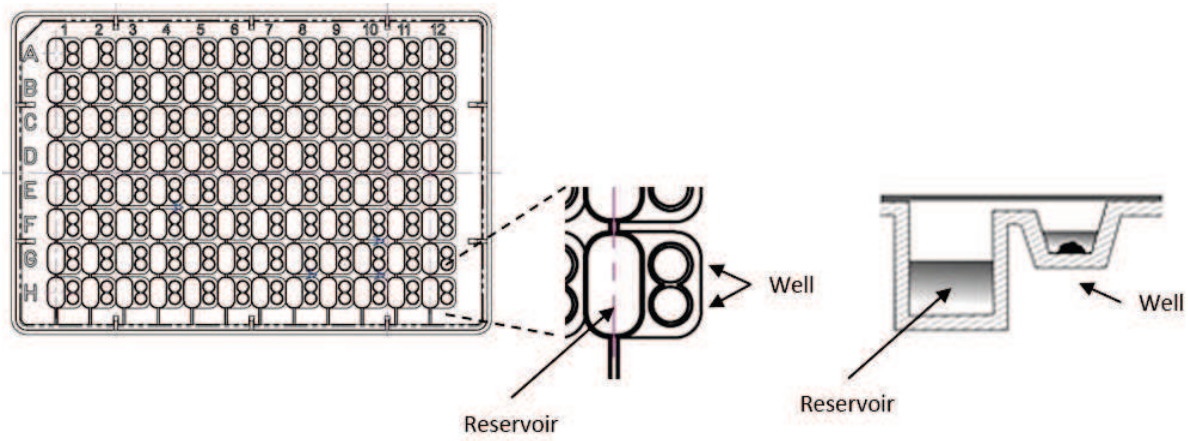


FIGURE 57

Sitting-drop vapour diffusion principle

When screening crystallization conditions, 96-well plates are used (MRC 2 or 3, Swissci; CrystalQuick 96, Greiner; CrystalEX, Corning). The figure here shows a MRC 2 plate (Swissci). Each of the 96 positions consists of a 50 µL-reservoir for the crystallization solution, and two wells for 200 nL to 1 µL-drops. Vapour diffusion between the drop and the reservoir allows equilibration, after hermetic closure of the plate.

the unfolded regions at the border of a domain. The choice of proteases is wide, (a dozen in the Proti-Ace kit, Hampton) and each of these shows a different proteolysis spectrum. First, mild proteolysis trials can be carried out and checked by SDS-PAGE to ascertain the efficiency of each protease. Finally, a 1/1000 to 1/10000 ratio (w/w) of the chosen protease are sufficient to induce proteolysis. It can be added either directly in the drop, or upstream, in the protein sample.

4.1.4 Diffraction data collection

According to Bragg's law ($2d_{hkl}sin\theta = n\lambda$), the conditions required for crystal diffraction depend on two experimental parameters, λ, the X-ray beam wavelength, and θ, Bragg's angle between the incident beam and the hkl planes. Two collection methods are thus possible:

- under monochromatic light, where λ remains constant and θ varies by rotating the crystal around an axis perpendicular to the X-ray beam. This method, called the oscillation method, is the most widely used in crystallography.
- under polychromatic light, where θ remains constant and λ varies. This method is also called the Laue method, since it was used by Max von Laue to record his first diffraction spectrum in 1912, varying λ from 0.45 to 0.95 Å.

The X-ray sources used during this work are of two types. First, the in-house source is made up of a MicroMax-007 HF rotating anode (Rigaku), equipped with OSMIC Confocal Varimax HF optics. The generated beam is thus monochromatic, and corresponds to the copper K_α emission line, with λ = 1.54 Å. The second type of source used was synchrotron light sources. Synchrotrons are large cyclic electron or positon accelerators, generating electromagnetic waves: called synchrotron radiation. This radiation has the advantage of being exceptionally bright compared to in-house sources (around 10^{19} photons/sec/mm²/mR², against 3.39 x 10^9 for our in-house source), it has low divergence, and has an emission spectrum extending from infrared to X-rays. Through focalization and monochromation to the desired λ wavelength, this radiation can be used in bio-crystallography for crystal diffraction. Diffraction data were collected on the PX II and PX III beamlines at the Swiss Light Source (Villigen, Switzerland) with beam sizes of 50 µm x 10 µm and 80 µm x 45 µm respectively; and the I04-1 line at the Diamond Light Source (Oxfordshire, England), on a fully automated "in plate" data collection system, with a beam size of 60 µm x 50 µm.

Detector devices are generally screens made up with multiple CCD detectors, which transform the photon flux into an electrical signal by the photoelectric effect. In the lab, we use a Saturn 944 detector (94 x 94 mm) (Rigaku); in the synchrotron sources used the bio-crystallography lines are equipped with new generation Pilatus detectors (respectively, 6M and 2M), offering a very high dynamic range, a large size, low noise and a read-out time in the order of milliseconds, which allows continuous data recording.

Diffraction data collection is generally achieved under cryoprotectant conditions. The main interest of this method is that it extends the lifetime of the crystals, by limiting propagation of damage caused by radiation (in particular, free radical formation due to water radiolysis). It requires however, freezing of the crystals in liquid nitrogen, a difficult operation since formation of crystalline ice can crack or even smash the crystal. Cryoprotective solutions can thus be used to favour amorphous ice formation while preserving the diffracting power of the crystals[617,618]. Several

techniques to protect crystals exist, and once again, their result remains empirical. For example, the crystal can be directly plunged into a drop of cryoprotective solution and immediately frozen afterwards, or after some time (from a few seconds to minutes). Alternatively, protection can be achieved by progressive soaking in increasing concentrations of cryoprotectant. The choice of the cryoprotective solution requires in general several trials to determine its nature and optimal concentration. Commonly used cryoprotectants are polymers, such as PEG, alcohols (glycerol, 2-methyl-2,4-pentanediol – MPD), salts (ammonium sulphate, sodium chloride, lithium sulphate) or organic compounds (ethylene glycol, dioxane). It is nevertheless important to carry out a diffraction test at room temperature, to ensure the good quality of the crystal prior to cryoprotection, which can in some cases totally annihilate the diffraction power.

4.1.5 Diffraction data processing

After collection of a full data set, generally achieved by crystal rotation in 0.1 to 1° increment, the processing includes four main steps: indexing, integration, scaling and merging/truncation of the data.

- Indexing provides information about the crystal orientation and the cell parameters, from the diffraction spots coupled with the oscillation angle. Auto-indexing programs seek the three vectors connecting all the reflections, in order to predict the full diffraction pattern and to measure its intensity.
- Integration consists in allocating, for each hkl reflection, its coordinates (hkl) and intensity I_{hkl} by simple summation of the pixel counts in the spot region.
- Scaling attempts to normalize the intensity of reflections assumed to be similar after the crystal symmetry. Experimentally, crystal radiation damage, variable incident beam intensity, or the detector calibration, produce reflections with different intensities when they are supposed to be identical. Scaling software minimises the differences between an individual reflection and the weighted mean of all the symmetry-related equivalents of the given reflection.
- Merging of the symmetry-related reflections into unique observations allows truncation, to calculate the structure factors modulus. By Fourier transform, it is then possible to deduce the electron density function.

There are numerous software, often compiled in packs, to process diffraction data. In the lab, we use mainly XDS, HKL2000 and mosflm.

During processing, several statistical criteria are calculated and give an indication of the quality of the data set and the model refinement. Among these criteria, one can mention:

- The resolution: it ascertains the fineness of the details in the electron density map.
- The R_{sym} factor (also called R_{merge} when datasets from different crystals are merged together)[619]: it depicts the discordance between measured intensities of equivalent reflexions. The lower this value, the better the consistency of the dataset. It illustrates the relation

$$R_{sym} = \frac{\sum_{hkl} \sum_{i=1}^{n} \left| I_{i_{hkl}} - \langle I_{h'k'l'} \rangle \right|}{\sum_{hkl} \sum_{i=1}^{n} I_{i_{hkl}}}$$

where $\langle I_{h'k'l'} \rangle$ is the averaged intensity of the single reflexion h'k'l' and I_{hkl} is the measured intensity of the reflexion hkl.

- The R_{meas} factor[620]: this factor, sometimes also called R_{rim} for redundancy-independent merging R factor (R_{merge}), is a robust variant of R_{sym}, which takes into account the contribution of each reflection normalised by $\sqrt{\frac{n}{n-1}}$ where n is the multiplicity.

$$R_{meas} = \frac{\sum_{hkl} \sqrt{\frac{n}{n-1}} \sum_{i=1}^{n} \left| I_{i_{hkl}} - \langle I_{h'k'l'} \rangle \right|}{\sum_{hkl} \sum_{i=1}^{n} I_{i_{hkl}}}$$

- The $R_{p.i.m.}$ factor[621]: the precision indicating merging R factor (R_{merge}) describes the precision of averaged intensities measurements

$$R_{p.i.m.} = \frac{\sum_{hkl} \sqrt{\frac{1}{n-1}} \sum_{i=1}^{n} \left| I_{i_{hkl}} - \langle I_{h'k'l'} \rangle \right|}{\sum_{hkl} \sum_{i=1}^{n} I_{i_{hkl}}}$$

- R and R_{free} factors: they allow to evaluate and validate the quality of the final model, by measuring the fit between observed and calculated structure factors. The R factors uses all the structure factors available, whereas the R_{free} factor only uses a randomly chosen set of reflexions T (between 5 and 10 %) which are excluded from the dataset used to refine the model[622]. These factors have to be the smallest possible, in the range from 0 to 1, 1 pointing out a bad and low resolution model.

$$R = \frac{\sum_{hkl} \left| |F_{obs}| - |F_{calc}| \right|}{\sum_{hkl} |F_{obs}|}$$

$$R_{free} = \frac{\sum_{(hkl) \in T} \left| |F_{obs}| - |F_{calc}| \right|}{\sum_{(hkl) \in T} |F_{obs}|}$$

- The signal/noise ratio <I/σ(I)>: the bigger this ratio is, the better the accuracy of the measurement of the structure.
- The mean half-set correlation (CC1/2)[623]: corresponds to the Pearson's correlation coefficient of one randomly chosen half of the observation to the other half. This value is near 1 at low resolution (> 2 Å) and drops to near 0.1 at high resolution (≈ 1.4 Å).
- The completeness: it corresponds to the ratio between the number of single reflexions experimentally measured and the theoretical number of reflexions needed to solve the data. It shall be the highest possible, ideally above 90 %.
- The redundancy/multiplicity: it illustrates the number of times the intensity of a single reflexion is measured. The bigger this value, the better the measured averaged intensity for a single given reflexion.

4.2 Structural study by cryo-electron microscopy

4.2.1 History

Light microscopes were developed in the early 16th century, probably by Zacharias Janssen or Galileo. It uses the visible light, focused with glass lenses and refracted when passing through a sample, to get a magnified picture of the said sample. However, the resolution limit of a microscope, i.e., its capacity to distinguish between two adjacent details, is limited by the diffraction and the wave nature of light. This principle was highlighted during the 19th century by Airy, Rayleigh and Abbe. It stipulates that, because of diffraction, the magnified image of a point is not a point but a spot, called the Airy disk. Two distinct points of the sample give thus rise to two spots on the image (with the condition that the point spread function is optimal), which can overlap and thus not be distinguishable anymore. According to Abbe, the resolution limit "d" of a microscope depends on the light wavelength λ, so that

$$d = \frac{0{,}6\lambda}{n \sin \alpha}$$

with n, the refractive index of the medium between the objective and the cover glass and α, the angular aperture. Thus, selecting the shortest wavelength in the visible spectrum (blue light, λ = 400 nm), a large aperture of 70°, and a refractive index of 1.5 (using an oil-immersion lens) the maximal resolution limit of a microscope would be around 170 nm.

In order to enhance the resolution limit of microscopes, the idea of using other electromagnetic waves sources, with shorter wavelength, emerged. In 1897, Joseph Thompson discovered the electron, and numerous works in the early 20th century showed that it was possible to align and focus an electron beam using magnetic fields. In 1924, Louis de Broglie stated the wave nature of electrons and surmised their wavelength as function of their relative mass "m" and relative velocity "v" acquired by acceleration under a differential electric potential "V", so that for V = 100 kV, the electron wavelength is 0.0037 nm, i.e., 10000 times less than blue light. The Abbe law goes on to state that the maximal resolution limit of an electron microscope for V = 100 kV could be 0.002 nm. All these studies gave birth to the very first electron microscope in 1933, designed by Knoll and Ruska. The German manufacturer Siemens marketed then the first microscope in 1938 with a resolution limit of 10 nm, and improved it to reach 1 nm in 1945. The theoretical limit is, however, far from being achieved, even today. This is in particular due to spherical aberration which prevents from developing very large aperture objectives (although Cs corrector have been developed since 1997[624] and equip now the new generation of microscopes).

4.2.2 Principle

The operating principle of an electron microscope is similar to that of a light microscope. The electromagnetic wave, here, electrons, is delivered by an electron gun. In the lab, we use Schottky-type guns. It is made up of a thin sharp-pointed tungsten cathode (the electron source itself), from which electrons are extracted with a soft differential potential (from 2 to 7 kV) and speeded up in a

high voltage field (between 80 and 300 kV) between the said cathode and a thin holey metal plate anode.

The electron beam is then focused onto the sample by the use of condenser lenses. The quality of these lenses, and in particular the weight of the spherical aberrations, determines the quality of the images obtained and the maximal resolution limit of the microscope.

When the electron beam reaches the sample, it passes through it and interacts with charged matter. Three different behaviours can then be observed:

- electrons can travel through the matter without meeting any atom. They are thus not deviated from their incident trajectory and don't loss any energy.
- electrons can pass through in the immediate neighbourhood of an atomic nucleus. They then undergo an elastic scattering and are deflected from their incident trajectory at an atom's electric potential-related angle. They however don't loss any energy.
- electrons can travel through the electronic cloud of an atom. They then undergo inelastic scattering, and are deflected from their trajectory with transfer of a part of their energy to the electrons of the atom.

Transmitted electrons pass one last time through projection lenses for further magnification, and are finally collected by the acquisition system. In the lab, we currently use Falcon and Falcon II CCD and CMOS cameras (FEI), with a resolution of 4096 x 4096 pixels.

4.2.3 Sample preparation: vitrification

Water represents between 65 and 90% of the cell mass. This preponderance suggests thus a major role, and encourages retaining biological samples hydrated during structural studies. The removal of water has indeed been widely studied and the hydration shell has been shown to be necessary to maintain the structure and biological activity of macromolecules.

In our environment, at atmospheric pressure, water can be found in three different states (*figure 58*): below 0°C, in a solid state (ice), between 0°C and 100°C, in a liquid state, and above 100°C, in a gaseous state (vapour). In its solid state, at atmospheric pressure and according to the cooling speed, different types of ice can form. Hexagonal ice (or I_h) is the most common one. It keeps the tetrahedral organization of water molecules due to hydrogen bonds formed in the continuity of the H-O covalent bonds, with cell parameters a = 4.52 Å and c = 7.39 Å. Another ice-type, called cubic ice (or I_c), can form at atmospheric pressure for temperatures below -70°C. This ice is composed of small crystals randomly organized, with cell parameters a = 6.35 Å. These two types of ice however have the disadvantage to be crystalline, i.e. both destructive for the sample and electron-dense.

In the early 1980's, Brüggeller and Mayer[625], followed by Dubochet and McDowall[626], got the idea to vitrify biological sample, that is to say, to fix them in non-crystalline ice also called amorphous ice (or I_v). This vitreous state is obtained when the cooling occurs very quickly below -135°C. The thermal motion decreases thus drastically, so that hydrogen bonds don't have time to rearrange to form crystalline ice (*figure 59*). The required cooling speed is in the order of 10^6°K/s at atmospheric pressure. Hence, liquid nitrogen can't be used as a cryogenic agent because of its Leidenfrost effect: its melting point (-210°C) and boiling point (-196°C) being very close, an insulating vapour layer forms around a "hot" object plunged in liquid nitrogen, and prevents its vitrification. Liquid ethane on the other hand has melting (-190°C) and boiling (-89°C) temperature long way from

each other. It is thus a good cryogenic agent if its temperature is maintained close to its melting point, by nitrogen-cooling.

Frozen-hydrated sample preparation is carried out by putting a 2 µL drop of sample on a copper microscopy grid, covered with a thin holey film of carbon (*figure 60*). This grid is previously made hydrophilic by glow discharge. When in contact with the grid, the sample penetrates the holes in the carbon film, and the excess is removed with a paper filter. Sample meniscus thus form in the holes, which contains the molecules of interest in suspension, randomly spread. The grid is finally plunged into liquid ethane, then stored in a nitrogen-cooled grid-holder, before examination under the microscope. In the lab, grid freezing is a semi-automatic process using the Vitrobot (FEI).

4.2.4 Biological sample observation by electron microscopy

The final image obtained is composed of grey levels, corresponding to the variations of electron scattering through the sample. The deflection angle of these electrons depends on the number Z of protons in the encountered atoms. Hence, biological samples, mainly composed of atoms with low atomic number ($Z_{hydrogen}$ = 1; Z_{carbon} = 6; $Z_{nitrogen}$ = 7; Z_{oxygen} = 8) have a weak contrast, compared to heavy atoms. The method employed is this case is the bright field mode. An aperture diaphragm is placed downstream the sample in the focal image plane of the objective lens, and called contrast diaphragm. It allows to eliminate the highly scattered electrons (θ > α), which

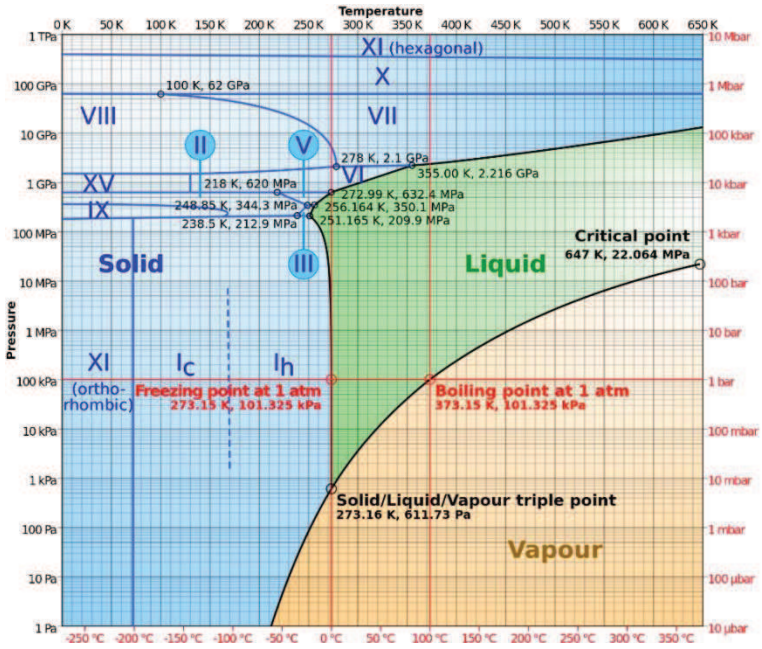


FIGURE 58

Water phase diagram

This 2-dimensionnal phase diagram shows the three states of water (vapour, liquid, solid), as a function of the temperature and the pressure. The Roman numerals indicate the different type of crystalline ice. At atmospheric pressure, it is mainly hexagonal ice (I_h) and to a lesser extent, cubic ice (I_c). Vitreous ice is not shown on this diagram since it is not crystalline. It could however be placed at the border between I_c and XI ice.

By Cmglee, via Wikimedia Commons

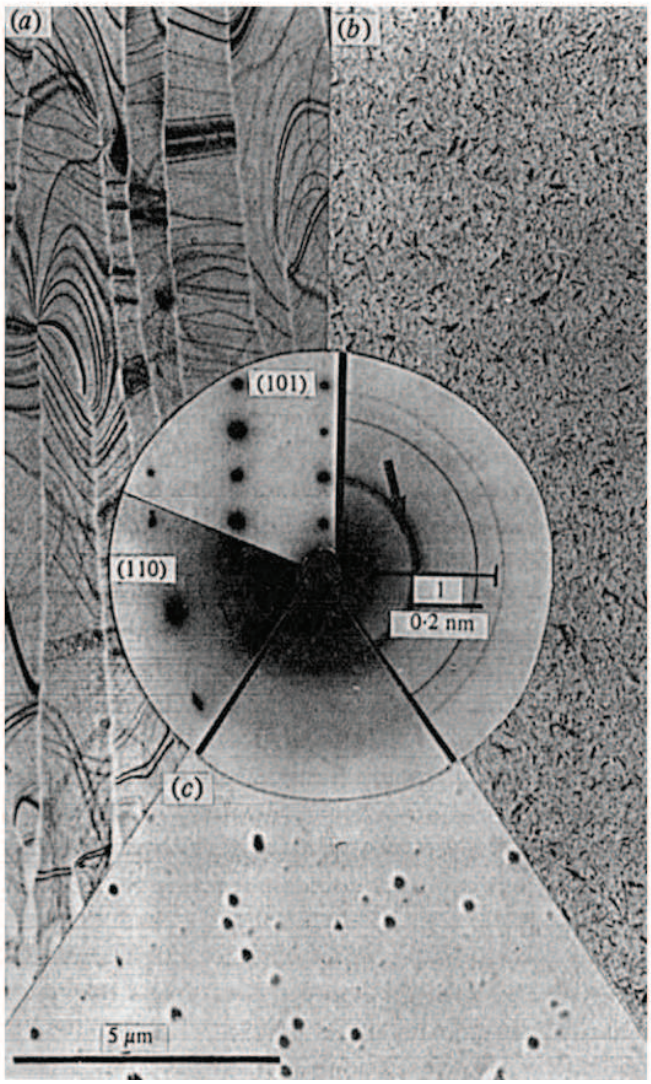
FIGURE 59

The different states of ice

Typical electron micrographs on three different types of ice.

a) Hexagonal ice (I_h) shows the (101) and (110) planes of the reciprocal lattice.
b) Cubic ice (I_c) shows the (111) plane.
c) Vitreous ice does not show any reflexion, illustrating its amorphous and non-crystalline nature.

From Dubochet et al., Cryo-electron microscopy of vitrified specimens. Q Rev Biophys. 1988 May;21(2):129-228.



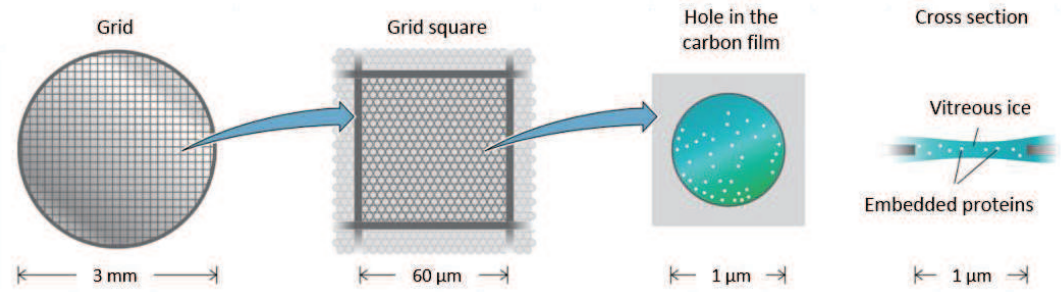


FIGURE 60

Microscopy grid

3 mm-grids are typically used in cryo-EM. They are composed of copper, covered with a thin holey carbon film. These 1 µm-holes are filled with sample which forms a meniscus. After fast freezing in liquid ethane, particles are embedded in their hydrated state and random orientations.

encountered high atomic number atoms. This technique is used to increase the contrast of biological samples. But this contrast, called amplitude contrast, accounts for only 7 to 10 % of the image[627]. The remaining 90 to 93 % comes from the phase contrast. It arises on the wave nature of electrons, leading to constructive interferences of the sample-delayed electrons (in the form of spherical waves) with non-delayed electrons (in the form of plane waves). This interference remains however weak in the case of biomolecules. It can thus be accentuated, in practise, by sample under-focusing.

Information transmission is gathered as a function, called contrast transfer function or CTF. The objective lens defects, in particular, induce modifications of the information which have to be corrected. Among these defects, on can mention spherical aberration (Cs), due to the fact that lens is more convergent for highly deflected electrons; and chromatic aberration (Cc), due to the fact that all electrons don't have the exact same wavelength and thus don't converge at the same point. The Fourier transform of the image (FT_{image}) depends thus on the Fourier transform of the object itself (FT_{object}) and the CTF of the microscope H(S), so that

$$FT_{image} = FT_{object} \times H(S)$$

The CTF itself is defined by the relationship

$$H(S) = A\sin[-\gamma(S)] - (1 - A)\cos[-\gamma(S)]$$

where A is the proportion of signal arising from the phase contrast (between 90 and 93 % as mentioned earlier), and γ, the phase delay introduced by the spherical aberration of the objective lens and by defocus, so that

$$\gamma(S) = \frac{2\pi}{\lambda} \times \left(\frac{Cs(\lambda S)^4}{4} - \frac{\Delta Z(\lambda S)^2}{2}\right)$$

with ΔZ, the defocus, Cs, the objective lens spherical aberration factor, S, the spatial frequency and λ, the electron wavelength.

The CTF is thus an oscillatory function, depending both on the lens defects and also on the sample defocus ΔZ. When $\sin[-\gamma(S)]$ tends toward 0, the amplitude of the transmitted wave decreases and inverts. Varying defocus allows to vary the CTF and thus, to increase or decreases intensity areas in the sample image. From different micrographs obtained at different ΔZ values, the CTF can be corrected to restore the real contrast of the sample as well as the signal which corresponds to the untransmitted frequencies when $\sin[-\gamma(S)] = 0$.

Finally, biological macromolecules are extremely sensitive to radiations produced by electrons, leading to free radical formation which destroys the sample, by bubbling effect[628]. Biological sample should thus not be subject to doses above a few tens of $e^-/Å^2$. Several acquisition modes are available on the microscope, to first explore the grid at low magnification and minimal radiation, to find areas of interest, before the acquisition itself, at higher magnification and electron dose.

4.2.5 Data processing

(For a complete and detailed review, see van Heel et al., 2000[629])

In single-particle electron cryo-microscopy, the data collected are a set of micrographs, each exhibiting a multitude of copies of the same macromolecule. These molecules, fixed in vitreous ice, are trapped in different orientations, and each physical particle is represented by a two-dimensional projection (*figure 61*). Interpretation of these images is semi-automated, by use of several software. In the lab, we mainly use EMAN2, RELION or IMAGIC.

Single isolated particles are picked (or boxed out) on the micrographs and referred hereafter to as "molecular images". A few thousands to tens of thousands of particles are required to achieve a first 3-D reconstruction at low resolution, and up to several hundreds of thousands for a medium/high resolution (3-8 Å). It is thus a time-consuming step, which can be achieved automatically by auto-boxing software. This option however does not ensure an optimal selection quality, and it is therefore necessary to cross-check by eye.

A band-pass filtering and normalisation of the raw molecular images is then achieved to reduce the influence of the high and/or low spatial frequencies. Algorithms use Gaussian blur overlays, in the reciprocal Fourier space, to delete given range of frequencies. Typically, high-pass filters delete low frequencies such as ramp or gradient effects, which might seriously disturb further alignment procedures; whereas low-pass filters delete very noisy high frequencies. However, the latter also contain fine details that need to be retrieved to get a high-resolution structure. High frequencies are thus only suppressed in preliminary analysis, and are reintroduced in later refinement procedures. After filtering, the unwanted background surrounding the particle is removed by imposing a soft-edged circular mask and the data within the mask are normalized to zero average density and given a normalised variance value of 100.

The CTF correction allows restoring, at least partly, the high resolution details, blurred by underfocus and lens aberrations. This is however compromised by the presence of noise as well as loss of information when H(S) = 0. Several images obtained at different defocus can thus be used to try to recover the lost information. Finally, techniques like π-phase flipping or Wiener filtering, to filter out or at least mitigate the noise, can be used[630].

Single molecules in solution are not held in a given orientation and have thus six degrees of freedom: the translational vectors x, y and z; and the Euler angles α, β and γ. Electron microscopy gives a projection along the z axis, leaving five parameters to be determined. Out of these five, three can be removed by alignment techniques. They correspond to the three in-plane parameters x, y and α. Alignment of a set of molecular images is an iterative process. Similar molecular images in similar rotational orientations can be classified by multivariate statistical analysis, and, within a given class, the corresponding pixels can be added together to increase the signal/noise ratio. These class averages can then be used as reference molecular images for further classification iteration rounds, and, after a few iteration rounds, good class averages with high signal/noise ratio can be obtained.

Each class average shows a density distribution through the object. Cryo-EM takes advantage of the fact that a vitreous ice-fixed object adopts a random angular distribution. This is however only theoretical, since in practice, a preferential distribution is generally observed, depending on the shape of the molecule, surface tension, etc. To perform 3-D reconstruction, relative β and γ Euler angles for each class have first to be estimated. This can be achieved by

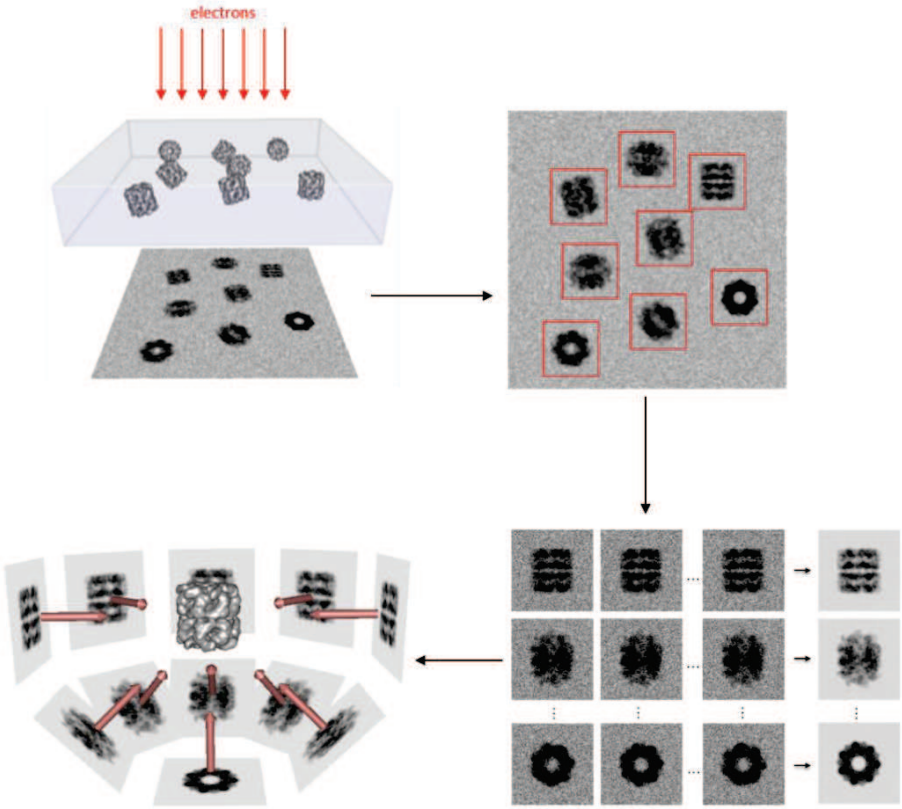


FIGURE 61

Reconstruction principle in cryo-electron microscopy

The sample is subjected to an electron beam, resulting in 2-D images of 3-D objects. These images represent the object in random orientations. For a first reconstruction, around ten thousand images are picked (or boxed, represented with red squares). For a finer structure, several hundreds of thousands of image can be required. Once selected, the images are classified according to their relative orientation, then averaged within same classes. After correction steps (band-pass filtering, CTF correction, etc.), relative Euler angles can be determined and the 3-D reconstruction results in a density map.

From Grigore Pintilie, Segmentation and registration of molecular components in 3-dimensional density maps from cryo-electron microscopy

---

angular reconstruction techniques, based on the common-lines projection theorem, stating that any pair of 2-D projections of a 3-D object has at least one common line in their respective 1-D projections[631]. A set of 1-D projections of a 2-D molecular image in different orientations is called a sinogram. Via sinogram correlation functions, it is possible to compare sinograms line-by-line (1-D projection by 1-D projection) to find pairs of shared lines, and to assign Euler relative angles with respect to a “reference” class average.

Finally, 3-D reconstruction can be achieved by backprojection of the different class averages with respect to their relative Euler angles. The 3-D reconstruction volume (or density map) obtained can be further refined by an iterative process. A few reprojections of the reconstituted 3-D object over different Euler angles can be used as reference molecular images for new classification rounds and/or to refine Euler relative angles of class averages.

To finish, once the 3-D map refined, it can be interpreted by atomic structure fitting for example, if an X-ray structure is available, or if the resolution is good enough, directly by atomic modelling.

5. *In vitro* recombinant nucleosome production

Recombinant nucleosome reconstitution follows the protocol established in 1999 by Karolin Luger[45,587] and optimised in 2004 by Pamela Dyer[632]. This reconstitution details into two majors axis: on one hand, production and purification of nucleosomal DNA, from 145 to 147 base-pairs; and on the other hand, production and purification of the isolated four histones H2A, H2B, H3 and H4. Finally, reconstitution consists in mixing the different partners to get recombinant nucleosomes. In the lab, this work is carried out by Dr. Kareem Mohideen Abdul. The following lines were inspired from the traineeship report of our Master student, Sinthuja Peiris.

5.1 Nucleosomal DNA

In the frame of this work, three different nucleosomal DNA have been used: α32, Widom145 and Widom147. The first one is a natural DNA, whose structure comes from a human α-satellite. It is found in the centromere of all chromosomes, and consists in 171 bp repeats. There, it plays an important role by recruiting in particular CENP-B, but is not essential for the proper function of the centromere. Only a limited part of this α-satellite sequence is used in our study. It is thus 32 copies of 84 bp fragments which are cloned in a pUC19 vector, provided by Timothy J. Richmond's lab (ETH, Zurich). By digestion-ligation, described later, the final product is a 147-bp palindromic nucleosomal DNA.

The Widom145 and Widom147 DNA correspond to the artificial and non-palindromic sequences described by Peggy Lowary and Jonathan Widom in 1998[633], and known as 601 DNA. This DNA is part of a DNA base including several billions of single DNA fragments, each tested for their affinity towards the histone octamer as well as for their positioning on the octamer. Here again, these sequences are cloned in pUC19 vectors, provided by Richmond's lab.

5.1.1 Cloning

The cloning strategy must be adapted to the DNA sequence: in the case of a non-palindromic DNA, the entire sequence can be cloned in a vector and extracted by a simple digestion with restriction enzymes. However, because of the recombinase activity of bacteria, a palindromic sequence won't be efficiently amplified. Thus, in that case, two half sequences, separated by a restriction site, can be cloned and further assembled by ligation to get the final desired product.

Besides, for a better yield, several repeats of the DNA fragment can be inserted in the plasmid, generally a pUC plasmid. The DNA sequence to clone is then flanked by restriction sites as illustrated on figure 62, where A is a unique restriction site (e.g., KpnI); B and B' are non-identical but with compatible sticky ends (e.g., BamHI and BglII); C is the restriction sites allowing to extract the DNA fragment out of the plasmid. A blunt-end enzyme is thus required, like EcoRV; D is the

restriction site used to ligate two half-fragments. This site has thus to produce sticky ends with high ligation efficiency. EcoRI can be used to produce a perfectly palindromic fragment of 146 bp; and HinfI is preferred to produce a 147 bp fragment.

Digestion of the plasmid by enzymes A and B linearizes the vector, which can incorporate a DNA insert generated with enzymes A and B'. Thus, the restriction site B is destroyed at the B-B' junction. This allows a new digestion with A and B to insert a new fragment generated by A and B'. And with each new restriction-ligation cycle, the quantity of insert can thus be doubled.

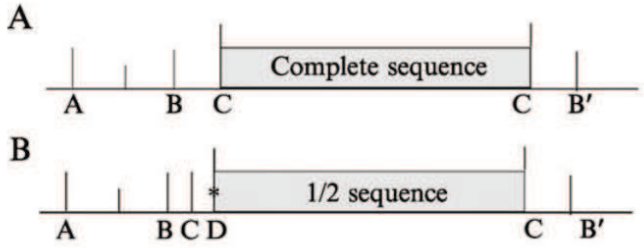


FIGURE 62

Cloning of multiple DNA fragments

a) Cloning of an entire DNA fragment (Widom145 and Widom147 for example).

b) Cloning of half a DNA fragment (palindromic α32 for example).

A: KpnI; B: BamHI; B': BglII; C: EcoRV; D: EcoRI or HinfI

5.1.2 Production

Chemocompetent *E. coli* HB101 cells were transformed with 100 ng of pUC19 plasmid, and incubated for 30 minutes on ice. This period allows bacterial membrane stabilisation as well as DNA neutralization with the help of calcium ions found in the bacterial stock. A heat shock at 42°C during 45 seconds is then applied, modifying the membrane fluidity and facilitating plasmid integration. The cells are finally put back on ice for a few minutes to stop the heat shock.

Addition of 200 µL of LB in the bacterial preparation and incubation at 37°C for one hour permits the bacteria to slowly enter growth phase and to transcribe their ampicillin resistance gene, bla. This gene encodes the TEM-1 ß-lactamase, a broad-range penicillin inhibitor. Lastly, the cell preparation is plated on LB-Agar and ampicillin (100 µg/mL) and left at 37°C for 18 hours.

Transformed bacteria, in presence of ampicillin, form distinct colonies on agar plates. One single colony is selected to inoculate a 5 mL LB miniculture, supplemented with ampicillin. After 3 hours at 37°C, this miniculture is used to inoculate a larger volume of 250 mL of LB plus ampicillin. Finally, when the OD_{600nm} reaches 0.6, these 250 mL of cultures are used to inoculate a Techfors-S (Infors HT) bioreactor of 20 L of TB plus ampicillin.

Cell growth is maintained at 37°C during 18 hours. The oxygen level (pO2) supplying the culture is a good indicator of the cell growth: during exponential growth phase, bacteria use oxygen and the pO2 decreases. It reaches a minimal plateau when the culture is stationary at maximum level, and increases back during cell decline. The cell culture is then ready to be harvested, by soft centrifugation at 6000 Xg for 5 to 10 minutes.

5.1.3 Purification

5.1.3.1 Plasmid extraction

For 100 g of bacterial cells, 290 mL of alkaline lysis solution “1” are added. It contains 50 mM of glucose to maintain osmotic pressure and avoid cells burst; 10 mM of EDTA to chelate divalent cations (Ca^{2+} and Mg^{2+}), which are essential to the DNase activity and the membrane integrity; 25 mM of Tris-HCl pH 8.0; and RNAse A to degrade bacterial RNA.

After homogenization, 270 mL of alkaline lysis solution “2” are added. It contains 1 % of SDS, a detergent to solubilize bacterial membrane; and 200 mM of NaOH, to denature genomic and plasmidic DNA, by hydrogen bond breaking. This step required 3h of incubation on ice.

After homogenization, 530 mL of cold alkaline lysis solution “3” are added and incubated for 20 minutes on ice with regular mixing. This solution contains 11.5 % of acetic acid and 3 M of potassium acetate, to neutralize alkalinity. Hydrogen bonds between complementary bases can thus form again. This is a selective step: plasmidic DNA, which is small, can easily renaturate while genomic DNA which is bigger remains as single-stranded. Plus, this step required soft mixing, to avoid breaking genomic DNA into small pieces which could then renaturate and contaminate the plasmidic DNA preparation.

By centrifugation at 12000 Xg and 4°C, it is possible to easily separate the double-stranded plasmidic DNA from the insoluble single-stranded genomic DNA, cellular debris and SDS. For each litre of supernatant, 520 mL of isopropanol are added to precipitate plasmidic DNA, insoluble in alcohols. The mix is incubated under stirring for 30 minutes, at room temperature, prior to centrifugation at 15000 Xg for 30 minutes, at 20°C. The resulting pellet contains the plasmidic DNA which is resuspended in TE 10/50 buffer (10 mM of Tris-HCl pH 8.0, 50 mM of EDTA).

A v/v equivalent of phenol is added to the DNA suspension, and centrifuged at 27000 Xg for 20 minutes, at 20°C. This step is repeated several times to eliminate all the protein contaminants found in the organic phase. Then, a v/v equivalent of CIA (24:1 Chloroform: Isoamyl Alcohol) is added to the aqueous phase and centrifuged at 15000 Xg for 15 minutes, at 20°C, to eliminate all traces of phenol and precipitate remaining proteins and lipids in the organic phase. This step is generally repeated several times in order to achieve a high degree of purity.

Finally, 10 % of PEG 6000 and 500 mM NaCl are added to the aqueous phase. This precipitation step permits to remove all traces of 300 bp nucleic acids and more. The mix is incubated on ice for 30 minutes, and centrifuged at 27000 Xg for 20 minutes at 4°C. The pellet, which contains the plasmidic DNA, is dissolved in a TE 10/0.1 buffer (10 mM of Tris-HCl pH 8.0, 0.1 mM of EDTA).

In order to remove the PEG traces, two CIA extractions are carried out and the pure plasmid is finally precipitated in 70 % ethanol and 110 mM NaCl. A one-hour incubation at -80°C is achieved and the mix is centrifuged at 27000 Xg for 30 minutes, at 4°C.

The final pellet is dissolved in TE 10/0.1 buffer and stored at -20°C.

5.1.3.2 EcoRV digestion

The purified plasmid is digested with EcoRV in the recommended commercial buffer. The reaction mix is incubated for 24 hours at 37°C. This step allows release of the target DNA from the plasmid.

5.1.3.3 PEG extraction

PEG extraction is used to remove the linearized plasmid (around 2.6 kb), emptied of its insert. This extraction is carried out in 40 % of PEG 600 and 500 mM of NaCl. The mix is incubated on ice for 1 hour, and then centrifuged at 27000 Xg for 20 minutes. Supernatant contains the DNA inserts which are precipitated with ethanol and dissolved in TE 10/0.1 buffer.

For the Widom145 and Widom147 DNA, the purification ends with this step. The α32 DNA requires, for its part, further steps: indeed, the 84 bp fragments must undergo digestion and ligation in order to reconstruct the 147-bp palindromic DNA.

5.1.3.4 Dephosphorylation

Dephosphorylation of the 84-bp α32 DNA is a prerequisite of the following steps. This dephosphorylation consists in removing all 5'-phosphate ends by means of the CIAP enzyme, in the recommended commercial buffer. The reaction mix is incubated for 24 hours at 37°C. Finally, the dephosphorylated DNA fragments are purified with a solution of phenol-CIA 1:1 (v/v). The fragments contained in the aqueous phase are then precipitated with ethanol as previously described and dissolved in TE 10/0.1.

5.1.3.5 HinfII digestion

This digestion step with HinfI generates 5'-sticky ends, to ensure ligation of two fragments. This digestion of the 84-bp fragments produces 72(+3)-bp and 9(+3)-bp fragments. The reaction mix is incubated for 24 hours at 37°C.

5.1.3.6 Purification

In order to isolate the 72(+3)-bp fragment, an anion-exchange chromatography is carried out, on a MonoQ 5/50 GL HR column (GE Healthcare). It is made of a divinylbenzene and polystyrene matrix, positively charged with a quaternary ammonium moiety and associated with counter-ions. DNA being a negatively charged polymer, each additional base-pair reduces its net charge by 2. An increasing ionic strength gradient using NaCl allows eluting the DNA. Small fragments like the 9(+3)-bp one are eluted first, around 450 mM of NaCl; whereas bigger ones like the 72(+3)-bp fragment are eluted around 680 mM.

5.1.3.7 Ligation

To get palindromic α32 DNA fragments of 147 or 145 bp, the 72(+3)-bp DNA are incubated with the T4 ligase, at room temperature for 24 hours.

5.1.3.8 Purification

Finally, the ligation product is again loaded on a MonoQ 5/50 GL HR column. As previously described, the 72(+3)-bp unligated fragments are eluted first, followed by the ligated 145 or 147-bp DNA. The corresponding fractions are pooled and precipitated with ethanol as previously explained. The pellet is finally dissolved in TE 10/0.1, and stored at -20°C.

5.2 Recombinant histones

The production and purification methods for the four H2A, H2B, H3 and H4 histones are roughly identical, with one exception in the composition of a buffer for histone H3, which shall be specified when appropriate.

5.2.1 Recombinant histones production

Chemocompetent *E.coli* BL21(DE3)pLysS cells were transformed with 100 ng of pET plasmid containing the gene of interest, and incubated for 30 minutes on ice. A heat shock at 42°C during 45 seconds is then applied and the cells are finally put back on ice for a few minutes to stop the heat shock.

Addition of 200 µL of LB in the bacterial preparation and incubation at 37°C for one hour permits the bacteria to slowly enter growth phase and to transcribe their ampicillin and chloramphenicol resistance genes. The latter encodes the MdfA transporter, a chloramphenicol-specific efflux pump. Lastly, the cell preparation is plated on LB-Agar with ampicillin (100 µg/mL) and chloramphenicol (25 µg/mL) and left at 37°C for 18 hours.

Transformed bacteria, in presence of ampicillin and chloramphenicol, form distinct colonies on agar plates. One single colony is selected to inoculate a 5 mL LB miniculture, supplemented with ampicillin and chloramphenicol. After 3 hours at 37°C, this miniculture is used to inoculate a larger volume of 250 mL of LB plus ampicillin and chloramphenicol. Finally, when the OD_{600nm} reaches 0.6, these 250 mL of cultures are used to inoculate the bioreactor of 20 L of 2xYT plus ampicillin and chloramphenicol.

Expression is induced with 0.4 mM final of IPTG when the culture reaches OD_{600nm} = 0.6 and that pO2 level is around 60 %. Cell growth is then maintained at 37°C for 3 hours, and then harvested by soft centrifugation at 6000 Xg for 5 to 10 minutes.

5.2.2 Inclusion bodies isolation

The pellet is resuspended in 50 mL of wash buffer to dissolve and homogenize it. It contains 50 mM of Tris-HCl pH 7.5, 100 mM of NaCl and 5 mM of BME.

The suspension is then sonicated for 15 minutes, alternating 10 seconds of sonication and 5 seconds of break, at 40 % amplitude. This step allows lysing the cell membranes, in order to release the inclusion bodies containing all the insoluble or misfolded proteins, including histones. This sonication is followed by a centrifugation at 27000 Xg for 30 minutes at 20°C.

The pellet is then washed three times with the previously described washing buffer, the second wash containing additionally 1 % of Triton X-100, a detergent to solubilize membranes. Each washing step is followed by a centrifugation at 27000 Xg for 30 minutes at 20°C. In the case of histone H3, the washing solution contains 10 mM of BME.

Finally, the pellet is dissolved in an unfolding buffer, containing 7 M of guanidinium-HCl, a denaturing agent able to dissolve inclusion bodies, 20 mM of sodium acetate pH 5.2 and 5 mM of DTT. Sequential centrifugations at 50000 Xg for 30 minutes at 20°C are carried out, until complete clarification of the supernatant, free of cell debris. This supernatant, mainly containing histones, is lastly filtered through 5 µm and 0.2 µm filters and stored at 4°C.

5.2.3 Isolated histones purification

The previously filtered solution is loaded onto a Sephacryl S-200 HR 26/60 size exclusion column (GE Healthcare). The elution buffer contains 7 M of deionized urea to unfold proteins, 1 M of NaCl, 20 mM of sodium acetate pH 5.2, 5 mM of BME and 1 mM of EDTA. The corresponding fractions are pooled and dialysed (SnakeSkin Dialyse Tubing 7000 MW, Thermo Scientific) against ultra-pure water and 10 mM of BME. This dialysis is carried out at 4°C for 24 hours, with renewal of the dialysis solution after 3 hours and 9 hours. Isolated histones, soluble in water, can this way be cleared of any protein contaminants, mostly insoluble. After centrifugation at 4000 Xg for 30 minutes at 4°C, the supernatant is frozen in liquid nitrogen and lyophilized for 24 hours under vacuum.

In order to complete the purification process, the histone powder is dissolved in 7 M of deionized urea, 20 mM of sodium acetate pH 5.0, 1 mM of EDTA and 10 mM of BME. This suspension is loaded onto a Resource S XK 20 mL cation-exchange column (GE Healthcare). It is made of a divinylbenzene and polystyrene matrix, negatively charged with a sulphonate moiety and associated with counter-ions. An increasing ionic strength gradient using NaCl allows elution of the bound proteins according to their net charge. Histone-containing fractions are then pooled and dialysed against ultra-pure water and 10 mM of BME, as previously described. After centrifugation at 4000 Xg for 30 minutes at 4°C, the supernatant is frozen in liquid nitrogen and lyophilized for 24 hours under vacuum. Pure histones can finally be stored at -80°C.

5.3 Nucleosomal particles reconstitution

5.3.1 Histone octamer reconstitution

Each pure and lyophilized histone is dissolved in the unfolding buffer previously described. After one hour of incubation at room temperature, the concentration of each of the four histones is measured, prior to mixing them together in a 1.5:1.5:1:1 ratio, with excess of H2A and H2B. This mix is then dialysed at 4°C for 24 hour with renewal of the solution after 3 hours and 9 hours, against a refolding buffer containing 50 mM of Tris-HCl pH 7.5, 2 M of NaCl and 10 mM of DTT. This step, by

slowly decreasing the urea concentration until complete removal, allows the histones to adopt their native conformation and associate together to form the histone octamer. The latter is nevertheless very sensitive and has to be stored at 4°C.

5.3.2 Histone octamer purification

The histone octamer previously recovered after dialysis is concentrated on a 10 kilo Daltons membrane (Amicon Ultra, Millipore). Histones having a tendency to stick to the concentration membrane, it is thus essential to resuspend regularly the sample to prevent important material loss.

The concentrated histone octamer is finally loaded onto a Sephacryl S-300 HR 26/60 size exclusion column (GE Healthcare), to separate the H2A-H2B dimer in excess. The refolding buffer is used for elution. Octamer-containing fractions are pooled and stored at 4°C in the short term or -20°C in 50 % of glycerol in the long term.

5.3.3 Nucleosome reconstitution

The purified histone octamer and the 147-bp DNA are mixed in equimolar ratio, in a buffer containing 20 mM of Tris-HCl pH 7.5, 1 M of KCl, 5 mM of DTT and 1 mM of EDTA. By continuous flow dialysis at room temperature for 10 hours, the salt concentration is gradually decreased until complete removal, allowing the octamer and the DNA to slowly associate.

The majority of reconstituted nucleosomes are off-centred, meaning that the central DNA base-pair is not symmetrically placed on the octamer, at SHL0. The sample is thus warmed up for 20 minutes at 42°C to centre the DNA. Finally, the sample is concentrated and stored at room temperature in the short term, or in the longer term in 20 mM of potassium cacodylate pH 6.0 and 1 mM of EDTA.

RESULTS

My results will be divided into three parts: by way of introduction, I shall describe the design of the expression vectors, required for the production of all NuRD subunits, as well as the expression tests carried out; the second part will focus exclusively on MBD3, my main topic of interest during my PhD; finally, the third part reflects on the study of RbAp46 and RbAp48 that I initiated.

RESULTS – PART I

EXPRESSION VECTOR DESIGN

1. Designing baculovirus vectors

When studying a protein *de novo*, the *E. coli* expression system is generally preferred, for the sake of convenience, yield, and cost. The baculovirus system is usually considered subsequently, in response to production issues in bacteria.

In the frame of the NuRD complex study, the expression vectors have been designed on a global and long-term basis: the baculovirus system, unlike the bacterial one, indeed allows co-infections with two (or more) recombinant genes, for *in vivo* subcomplex formation. Besides, the baculovirus system provides the advantage (but which can also be a disadvantage in some cases), unlike bacteria, that it allows post-translational modifications, disulphide bond formation, and chaperone-assisted folding, processes that are sometimes essential to ensure the stability of the protein.

To this end, all the subunits of the NuRD complex have been cloned in baculovirus vectors, fused to different tags to allow tandem-affinity purifications of subcomplexes (TAP-TAG). I designed these vectors in the lab of Dr. Ali Hamiche, with the help of Dr. Arnaud Depaux. The human genes of the different subunits come from cDNA libraries. For each of these genes, the protocol follows the same logic:

- Sub-cloning of the gene in a pFastBac1 plasmid with different tags on the N-terminal or C-terminal, and a cleavage site;
- Transformation of *E. coli* DH5α with the ligation product and plate-culture on LB-Agar (+ 100 µg/mL ampicillin);
- 3 mL-minicultures in LB from one single colony;
- Purification “miniprep” of the plasmid;
- Checking of the gene insertion by cleavage and agarose gel electrophoresis;
- 20 mL-maxiculture in LB from the selected miniculture;
- Purification “maxiprep” of the plasmid for storage;
- Sequencing;
- Transformation of *E. coli* DH10Bac with the recombinant plasmid and plate-culture on LB-Agar (+ 15 µg/mL tetracycline; 50 µg/mL kanamycin; 7 µg/mL gentamycin; 0.4 mM IPTG final; 625 µg/mL X-Gal);
- 3 mL-minicultures in LB from one single white colony;
- Purification “miniprep” of the bacmid;
- PCR with “M13 reverse” and “gene forward” primers and agarose gel electrophoresis;
- 3 mL-miniculture in LB from the selected miniculture;
- Purification “miniprep” of the bacmid;
- Storage in 70 % ethanol for the “Baculovirus” service.

In order to design vectors with different tags, I first had to modify the commercial pFastBac1 vector. Therefore, I ordered several DNA oligos based on the model “BamHI – ATG – Tag – (Cleavage site) – XhoI – NotI – HindIII”. After annealing of the sense and antisense oligos, these were mixed with a pFastBac1 vector that had previously been linearized with BamHI and HindIII restriction

enzymes and purified. Ligation was carried out overnight at 15°C using T4 DNA ligase in the provided commercial buffer, and the ligation product was then used to transform *E. coli* DH5α cells. After purification of the amplified plasmids, the correct insertion of the oligos was verified by sequencing (GATC-Biotech).

The genes of the different NuRD subunits were modified by PCR to append XhoI and NotI restriction sites on both sides. I could then ligate the genes of interest, digested twice with XhoI/NotI, in the destination pFastBac1 vector, itself linearized with the same enzymes. The mix consists approximately one vector for three inserts.

After miniculture and purification of the recombinant pFastBac1 vectors, insertion of the gene of interest was verified by digestion and agarose gel electrophoresis, as well as by sequencing. I could then proceed to transposition of the genes of interest into bacmids. To do so, I transformed DH10BAC cells with the recombinant pFastBac1 vectors, and cultivated them. Recombination was achieved automatically in these cells, made possible with a transposase expressed by these bacteria. After bacmid purification, the correct transposition of the gene of interest was checked by PCR. The right choice of primers is critical for this step. The bacmid indeed possesses two M13 sites, downstream and upstream from the cloning cassette. However, even if a PCR using "M13 Forward" and "M13 Reverse" primers would confirm the transposition of the gene (the PCR product then being larger than in absence of recombination), this choice would not allow the confirmation of the correct orientation of the gene. Indeed, Tn7 recombination is a non-directional process, and the gene can, half of the time, be transposed the wrong way. This is why I favoured a bacmid-specific primer ("M13 Reverse") and a gene-specific primer ("Gene Forward"). The principle of this PCR is illustrated on figure 63.

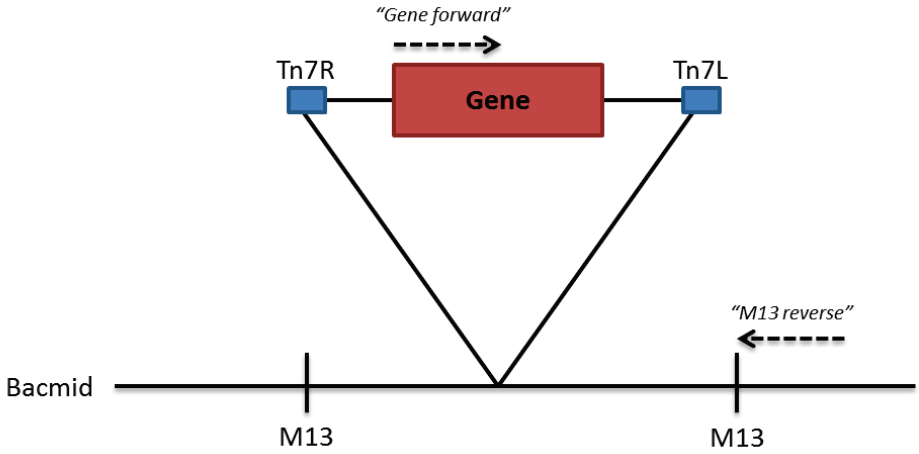


FIGURE 63

Recombination in bacmids

Although using M13 forward and M13 reverse primers would ensure the transposition of the gene, its correct orientation would still be unconfirmed. To overcome this issue, we made the choice of a "bacmid-specific" primer (M13 reverse) and a "gene-specific" primer (Gene forward).

For all the positive bacmids, I used a few µL of the previous miniprep to inoculate a fresh 3 mL-miniculture, then I purified the amplified bacmid and gave it to the "Baculovirus" service for baculoviral strain development.

In the lab, we have at our disposal plasmid and bacmid extraction kits (NucleoSpin® Plasmid and NucleoBond® BAC 100, Macherey-Nagel). These kits have the advantage of being easy-of-use, but the yield is rarely optimal, in particular in the case of low-copy plasmids or bacmids. To overcome this problem, I chose an alternative extraction technique, using phenol-chloroform. This technique consists of adding an equivalent volume of phenol-chloroform (1:1) to the cell lysate, that has been cleared of its genomic DNA by alkaline lysis. This mix allows protein denaturation and lipid dissolution. At neutral pH, DNA and RNA are both negatively charged and are separated in the aqueous phase. After RNAse treatment, plasmidic DNA can be precipitated in ethanol/sodium acetate.

The table below summarizes the baculovirus vector I designed, specifying for each the tag used and the presence or not of a cleavage site.

Protein	Tag	Cleavage site
CHD4	HA (N-ter)	-
HDAC1	Myc (C-ter)	-
HDAC2	Myc (C-ter)	-
MTA2	Flag (N-ter)	-
MTA2	6x-His (N-ter)	TEV
MBD3	Flag (N-ter)	-
MBD3	6x-His (N-ter)	-
MBD3	6x-His (N-ter)	TEV
RbAp46	6x-His (N-ter)	-
RbAp48	Myc (C-ter)	-
RbAp48	6x-His (N-ter)	-

2. Expression tests

The "Baculovirus" service of the IGBMC ensures the development of recombinant viral strains, by infecting a small volume of insect cells with the recombinant bacmid provided. After cultivation, the encapsidated baculovirus are released in the culture medium and can be harvested, to further infect larger volumes of insect cells.

As a first step, expression tests must be conducted to define the optimal infection and culture conditions. Several parameters can be adjusted: the cell line (Sf9, Sf21, Hi5), the quantity of

virus for the infection (between 0.1 and 10 pfu) as well as the cultivation time (from 48 to 72 hours). The growing temperature remains stable at 27°C. Expression of recombinant proteins being an equivocal process, I decided to test all these parameters.

For practical reasons, I started to test all the constructs which had a 6x-His-tag. I first tested the expression in Sf9 cells, varying the viral titration and the cultivation time. In the case of MBD3 N-His and N-His-TEV, no significant difference was observed between the cultivation conditions (*figure 64*). I thus opted for the simplest condition (1 pfu and 48 hours of culture). However, only very weak overexpression could be detected for RbAp46 and RbAp48. I thus carried out other tests, by changing the insect cell line. The Sf9, Sf21 and Hi5 lines were tested in parallel. High levels of overexpression were then observed for RbAp48 in Hi5 cells. RbAp46 also showed good levels of overexpression in Hi5 cells, although less than its paralogue RbAp48. In the case of MBD3, no difference in overexpression could be detected (*figure 65*)

All these encouraging results have been determinant in the choice and the further study of these proteins.

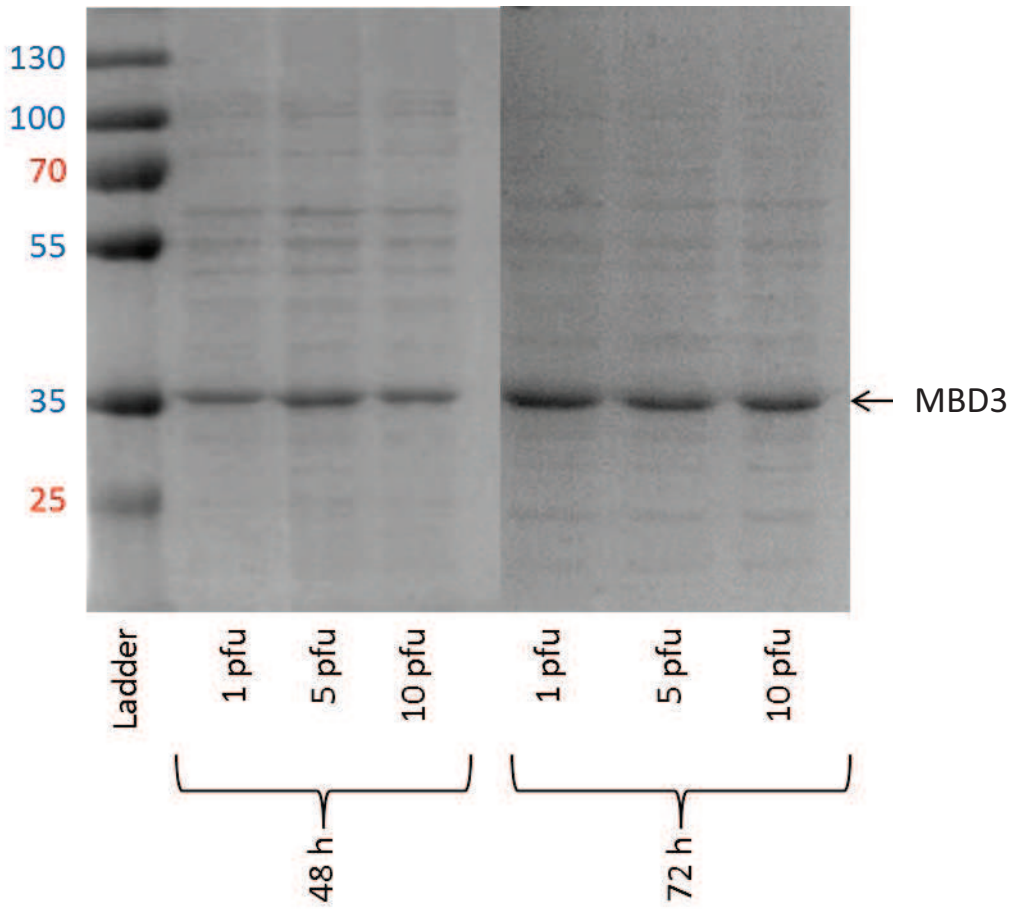


FIGURE 64

MBD3 N-His-TEV: expression tests

SDS-PAGE, Coomassie staining.

Expression tests carried out on the MBD3 N-His-TEV construct, in Sf9 cells. Three different virus concentrations were tested (1, 5 and 10 pfu), and two cultivation times (48 and 72 hours post-infection).

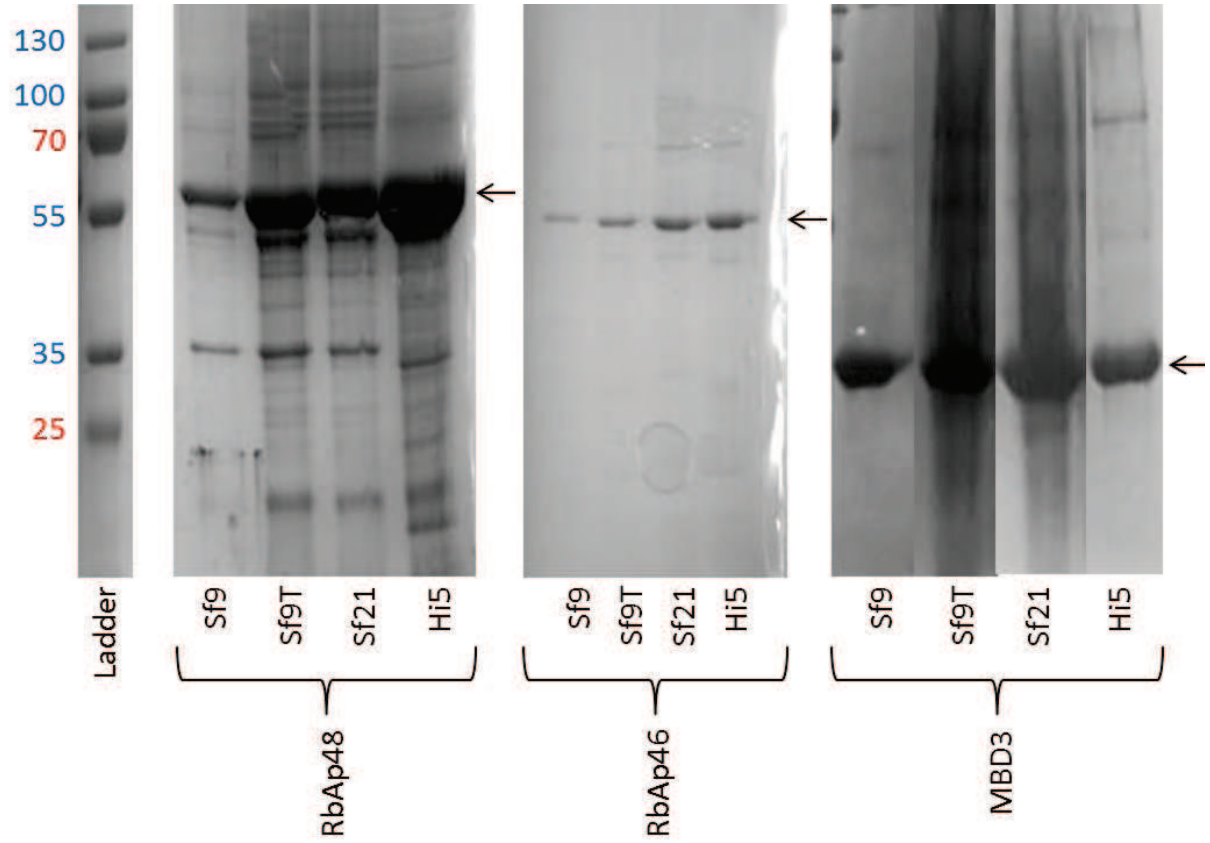


FIGURE 65

RbAp48 N-His, RbAp46 N-His and MBD3 N-His-TEV: expression tests

SDS-PAGE, Coomassie staining

Three different strains were tested: Sf9 (with two different cultivation conditions), Sf21 and Hi5. For RbAp48, a clear overexpression is seen in Hi5 cells. Same for RbAp46, although in lesser amounts. For MBD3, all three strains showed more or less the same overexpression.

In parallel, I carried out preliminary trials for co-infections. The literature mentions, among others, interactions of MTA2 with both MBD3 and CHD4. A stable subcomplex of NuRD could then associate these three subunits. Co-expression tests were done by varying the titration of each virus (1:1:1, 1:1:5, 1:5:1, 1:5:5, 5:1:1, 5:1:5 and 5:5:1). After 48 hours of culture in Sf9 cells, these were lysed and the total extract was analysed by SDS-PAGE. These tests have highlighted a good overexpression of the three proteins, especially in 1:1:5 and 5:1:5 (MTA2:MBD3:CHD4) ratios, as seen by SDS-PAGE (*figure 66*).

However, no co-purification trial was carried out to date on this subcomplex. It thus remains one avenue worth exploring in the future, within the frame of NuRD project.

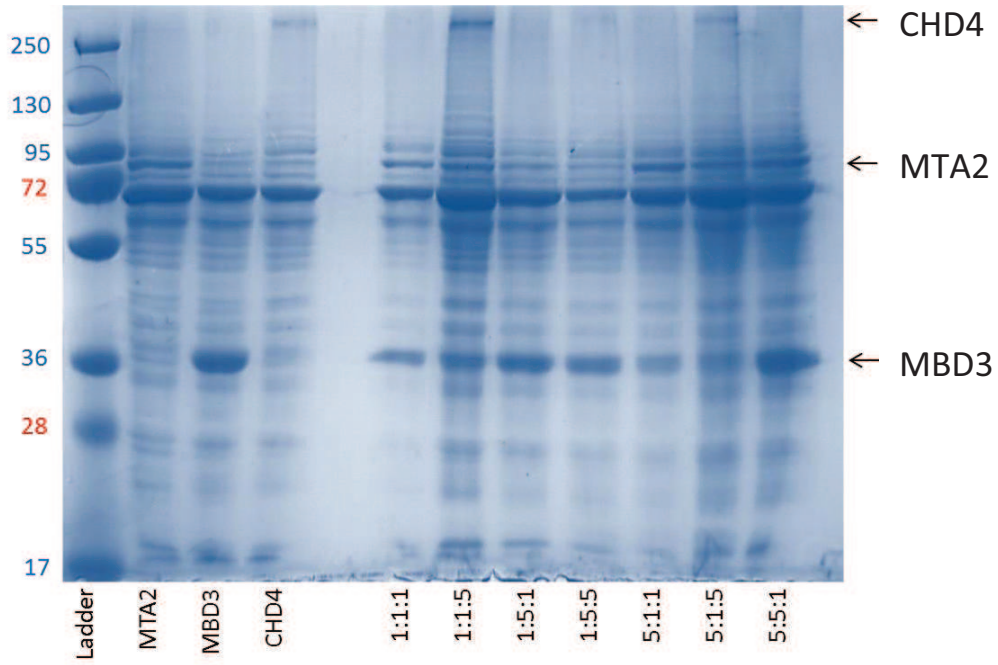


FIGURE 66

Co-expression tests

SDS-PAGE, Coomassie staining

A co-infection test has been carried out using the three constructs MTA2 N-Flag, MBD3 N-His and CHD4 N-HA. First three lines after the ladder are the references for each of these constructs. Two ratios allowed nice observation of the three proteins in the same sample: 1:1:5 and 5:1:5.

RESULTS – PART II

STUDY OF THE PROTEIN MBD3

1. The different MBD3 isoforms

1.1 Context

Human MBD3 is composed of 291 amino acids, including the initiator methionine. Its gene is located on the chromosome 19p13.3, in the locus 1576671-1592761. It is composed of 7 exons, the last one being non-coding. In addition, an alternative splicing site exists in the reading frame of exon 1, leading to the expression of a short isoform of MBD3, named MBD3Δ. This isoform lacks residues 5 to 36, i.e., the central part of the MBD domain, which is responsible for DNA-binding.

In the frame of this work, I have been interested in the structural aspect of DNA-binding by MBD3. The expression vectors having been designed and optimal cultivation conditions determined, I then had to develop an efficient purification protocol, to isolate MBD3 in a soluble and non-aggregated form. For this study, I chose to work with the N-His-TEV construct, which includes a TEV cleavage site directly after the 6x-His-tag. This allowed me to remove the tag after purification and recover MBD3 without the artificial sequence fused.

1.2 Purification protocol design

The lack of biochemical data forced me to undertake trials with a basic protocol for protein purification. For that purpose, I harvested the infected insect cells after 48 hours of cultivation with a soft centrifugation at 300 Xg, for 15 minutes at 4°C. The cell pellet was then washed with cold PBS/10 % glycerol (137 mM of NaCl, 2.7 mM of KCl, 10 mM of Na_2HPO_4, 1.76 mM of KH_2PO_4, 10 % of glycerol). After a second centrifugation, the pellet was resuspended in a lysis buffer (1 M of NaCl, 20 mM of Tris-HCl pH 7.5, 10 mM of imidazole, 10 % of glycerol, 0.01 % of NP-40, cOmplete EDTA-free) and ground using a Dounce homogenizer. Finally, the sample was sonicated for 3 minutes at 60 % amplitude (Labsonic M, Sartorius), then centrifuged to remove the cell debris at 5000 Xg for 10 minutes. The supernatant was allowed to bind to Ni-NTA resin for 2 hours at 4°C. This resin was then washed several times with a wash buffer (150 mM of NaCl, 20 mM of Tris-HCl pH 7.5, 10 mM of imidazole, 10 % of glycerol, 0.01 % of NP-40, cOmplete EDTA-free), and the protein was finally eluted with an elution buffer (150 mM of NaCl, 20 mM of Tris-HCl pH 7.5, 300 mM of imidazole, 10 % of glycerol, 0.01 % of NP-40, cOmplete EDTA-free). The quality of purification was checked by SDS-PAGE.

As seen on the SDS-gel (*figure 67*), this first purification trial was very conclusive. Although a large part of the protein remained in the pellet, the affinity purification allowed the recovery of MBD3 and to get rid of a major part of contaminants. Plus, MBD3 in fusion with its tag has a molecular weight of 34.5 kDa, which corresponded to its migration on the gel.

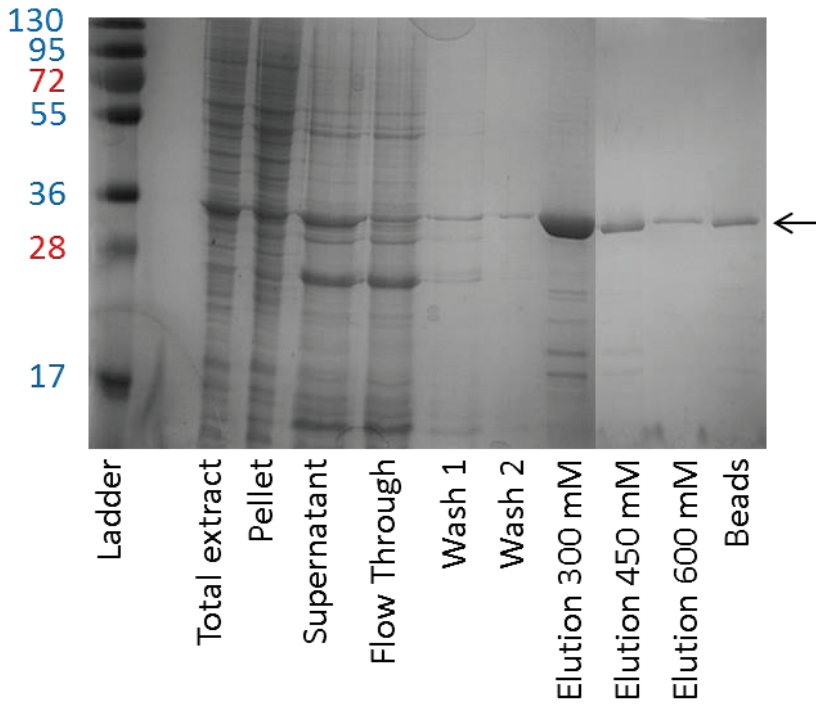


FIGURE 67

MBD3 N-His-TEV: first purification assay

SDS-PAGE, Coomassie staining

As seen on the gel, though a significant part of the protein remains in the pellet, a first affinity step could be carried out and MBD3 could be recovered, almost pure, using 300 mM of imidazole. Ni-NTA resin was further washed with increasing concentrations of imidazole (450 mM and 600 mM), and showed only little loss. Furthermore, MBD3 has a molecular weight around 34.5 kilo Daltons, which corresponds to its migration on SDS gel.

After this first success, the culture was scaled-up to produce enough material, in order to proceed and further purify MBD3. However, and although the previously described protocol had been reproduced to the letter, MBD3 was no longer soluble. I then tried to optimise the purification buffers and the lysis methods, and tested over thirty different conditions. Among others, I varied the salt type and concentration (from 150 mM to 1 M of NaCl or KCl), the detergent, replacing NP-40 by CHAPS, a zwitterionic detergent whose critical micellar concentration (CMC) is about ten times higher than that of non-ionic detergent (6 mM against 0.5 mM) or addition of a non-denaturing concentration of urea (between 1 and 2 M). Among the lysis methods, I tried Dounce homogenizer lysis and/or sonication. After several months of unsuccessful results, I finally managed to solubilize MBD3 using the following protocol: I harvested the cells by soft centrifugation at 300 Xg for 15 minutes at 4°C. The cell pellet was not washed with PBS/glycerol to avoid lysing the fragile insect cells. The pellet was then directly resuspended in the lysis buffer (500 mM of NaCl, 20 mM of Tris-HCl pH 7.5, 10 % of glycerol, 0.2 % of NP-40) and homogenized using a Dounce homogenizer. One

tab of cOmplete EDTA-free was then added, previously dissolved in lysis buffer, and the sample was sonicated for 1 minute at 60 % amplitude. The cell lysis was checked using a light microscope, after Trypan blue coloration, and the sample was centrifuged for 10 minutes at 5000 Xg. The supernatant was allowed to bind to Ni-NTA resin for one hour at 4°C. The resin was then washed five times with lysis buffer supplemented with 25 mM of imidazole and poured into a polypropylene gravity flow column (Poly-Prep®, Bio-Rad). Elution was done directly on the column, by addition of an elution buffer (500 mM of NaCl, 20 mM of Tris-HCl pH 7.5, 300 mM of imidazole, 10 % of glycerol, 0.2 % of NP-40). This protocol was applied in parallel to Sf9 and Sf21 cell lines, to check for differences of solubility. SDS-PAGE showed that a large fraction of the protein was insoluble, but that there was also a non-negligible quantity of soluble protein. Also, this assays led to the conclusion that Sf21 were more appropriate to express MBD3 (*figure 68*). This new protocol not being significantly different from the first one, it raised the question of culture quality over this period. Numerous parameters can indeed affect the quality, such as cell age and shape, virus quality, viral stock age, culture medium batch, possible contaminations, etc.

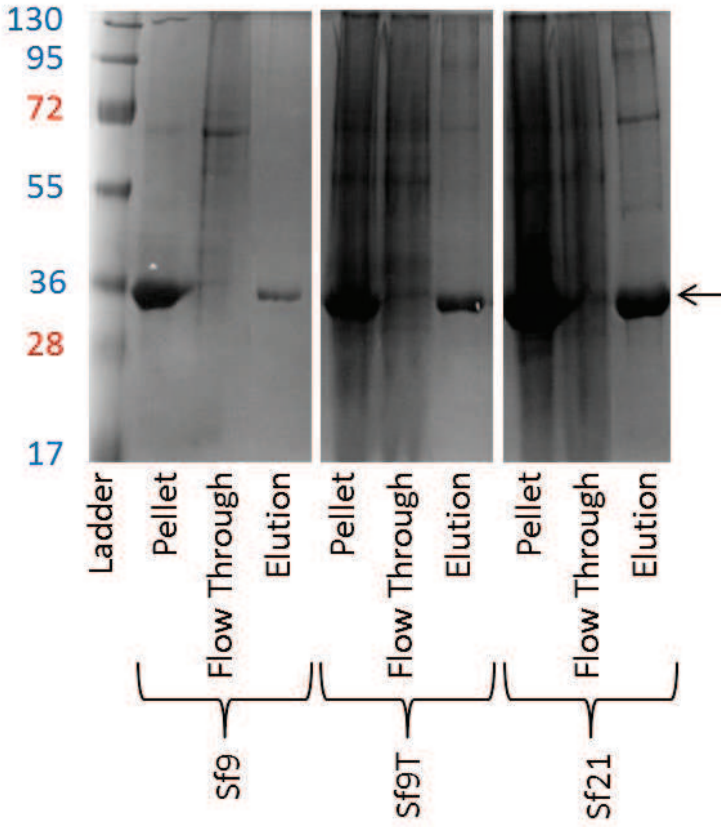


FIGURE 68

MBD3 N-His-TEV: solubility tests

SDS-PAGE, Coomassie staining

To check for solubility among different cell lines, MBD3 N-His-TEV has been expressed in Sf9 and Sf21 cell lines. After lysis, the clarified supernatant was allowed to bind to Ni-NTA resin for 2 hours, prior to elution with imidazole. The gel shows a clear difference of solubility, although MBD3 is found in significant amounts in the pellets of all cell lines. Nevertheless, the Sf21 cell line showed an improved solubility and has thus been chosen to carry on MBD3 expression.

New trials were carried out in order to automate and increase the quality of the purification. To this end, I chose to use an affinity chromatography column connected to an Äkta Purifier chromatography system. Two types of columns were tested: HisTrap FF 5 mL (Ge Healthcare), containing of Ni-NTA resin, and HiTrap IMAC FF 5 mL (Ge Healthcare) composed of cross-linked agarose beads 6 %, modified with chelating moieties at their surface. While HisTrap columns showed an optimal binding and elution with 130 mM imidazole, HiTrap IMAC columns loaded with cobalt chloride did not show any affinity for MBD3. Column-affinity purification is, nevertheless, difficult to achieve when salt or pH conditions are not optimal, or when the 6x-His-tag is not freely available, due to close secondary structure, for example. The loading flow rate of the sample onto the column is indeed rather fast, up to 5 mL/minute on a 5 mL column. To address the reason for such low affinity, and define whether it was due to the resin itself or to the protein's affinity towards cobalt, I carried out other tests in parallel with Ni-NTA and "Talon" resin (Clontech), using the batch method. Results showed a good affinity for nickel but no affinity for cobalt, even after 1 hour of binding (*figure 69*).

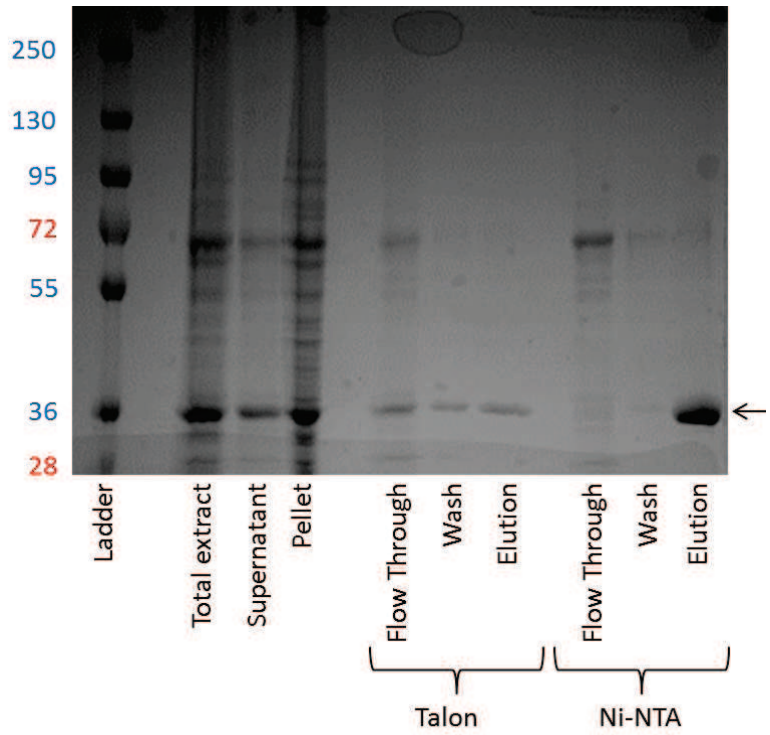


FIGURE 69

MBD3 N-His-TEV: Ni-NTA vs. Talon resin

SDS-PAGE, Coomassie staining

6xHis-tags show an improved affinity towards bivalent metals, such as nickel or cobalt. To assess which of the two is the most suitable for MBD3 purification, tests were carried out using Ni-NTA resin (Ni^{2+}) and Talon resin (Co^{2+}). Surprisingly, a clear difference could be observed between the two, and the choice fell on Ni-NTA resin.

After finally achieving a soluble fraction of MBD3, I could then carry on with a second step of purification, by loading the concentrated protein on a Superdex 200 10/300 GL size exclusion column (GE Healthcare). The elution buffer was composed of 150 mM of NaCl, 20 mM of Tris-HCl pH 7.5, 10 % of glycerol and 0.2 % of NP-40. However the chromatogram showed only a single peak, eluted in the void volume (V_0 = 9 mL): MBD3 was thus fully aggregated in this buffer.

After several unsuccessful trials, ending up with a majority of aggregates, I finally managed to get two different species from the size exclusion chromatography: at V_e = 9 mL, limited quantities of aggregates of MBD3; and at V_e = 13.5 mL, pure MBD3 as confirmed by SDS-PAGE (*figure 70*) - this elution volume corresponds to a globular protein of around 70 kilo Daltons. The elution buffer used was composed of 500 mM of NaCl, 20 mM of Tris-HCl pH 7.5, 5 % of glycerol, 1 mM of TCEP, 5 mM of EDTA and 0.01 % of Tween-20. These observations have led me to consider the dimerisation of MBD3. Indeed, literature mentions homo- and hetero-dimerisation of MBD2 and MBD3[16].

At the end of this purification process, I obtained around 0.5 mg/mL of pure MBD3 dimer in 1 mL, i.e., 500 µg of protein, from 200 mL of insect cells culture. The final yield is thus estimated to 2.5 mg/L of culture.

1.3 X-ray crystallography

In order to start the first structural studies on MBD3, I scaled-up the culture volume to 2 to 3 L in order to get 5 to 6 mg of pure protein. The protocol has been further optimised by replacing the Tween-20 by 4 mM of CHAPS, which has the advantage of being easily dialyzable in case of need.

The first crystallisation trials were done on the MBD3 dimer, at several concentrations (2 mg/mL, 5 mg/mL and 10 mg/mL). I used the following commercial screens: The PEGs, The Cations, JCSG+, The $AmSO_4$, and The Classics (Qiagen); and Wizard I+II (Emerald BioSystems). These tests were performed in MRC 2 96-well plates, with a reservoir volume of 50 µL of 1X crystallisation solution, and drops of 200 nL of protein + 200 nL of 1X crystallisation solution (0.5X final). After a couple of days at 20°C, I noted the first observations. In particular, no precipitation occurred at high salt concentration (> 2 M) and for a pH between 6.5 and 9. I hypothesized thus that salting-in could help the crystallisation of this protein. Indeed, the generally used method in vapour diffusion crystallisation is salting-out: the more the salt concentration increases in the drop, the less important the protein solvation effect is, leading to their precipitation or, more ideally, their crystallisation. During salting-in, the protein crystallises (or precipitates) when the salt concentration decreases, leading to a loss of stabilisation of the surface charges by ions in the drop.

I thus proceeded with new crystallisation trials of MBD3 at 5 mg/mL, using the same commercial screens. However, the crystallisation solution in the reservoir was diluted twice. In practical terms, I filled the reservoir with 12.5 µL of 1X solution, sat up 200 + 200 nL drops of protein and 1X crystallization solution (0.5X final), and finally diluted the reservoir to 0.25X, by adding 37.5 µL of water.

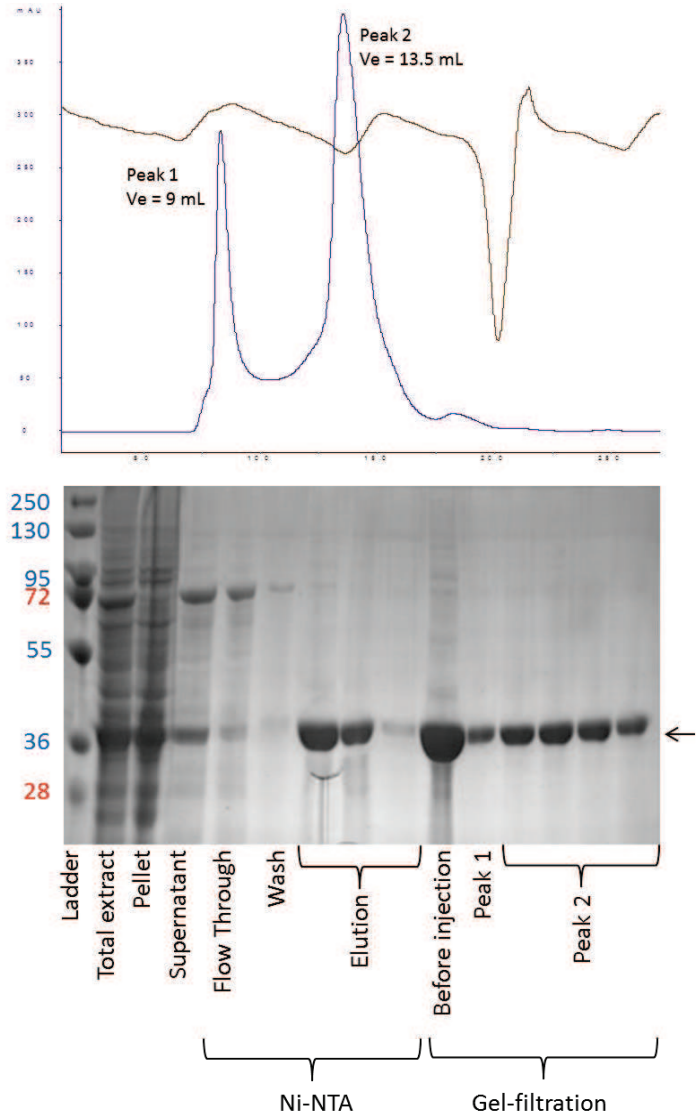


FIGURE 70

MBD3 N-His-TEV: purification

Top: gel-filtration profile (Superdex 200 10/300 GL)

Below: SDS-PAGE, Coomassie staining

After a first affinity step, MBD3 could be further purified using gel-filtration. The chromatogram shows two peaks, corresponding to aggregates (peak 1), and pure MBD3 (peak 2). Furthermore, considering the elution volume of peak 2, MBD3 is eluted as a 70 kilo Dalton species, which corresponds to a dimer.

In total, over 3500 crystallisation conditions were tested, among which only one gave crystals (*figure 71*). It was the condition F6 of the commercial screen Wizards I+II (condition #18 in the Wizards II screen), containing 200 mM of calcium acetate, 100 mM of Tris-HCl pH 7.0, 20 % w/v of PEG 3000, final pH 7.4. These crystals were obtained from the first salting-out experiments, with 5 mg/mL MBD3 in 500 mM of NaCl, 20 mM of Tris-HCl pH 7.5, 5 % of glycerol, 5 mM of EDTA and 1 mM of TCEP. They appeared at day 6, and they grew during three weeks before stabilisation. The majority of these crystals were spherulites, but a few cubic or tetragonal crystals were also

observed. These crystals were fished and mounted on loops, then frozen in liquid nitrogen after cryoprotection in 200 mM of calcium acetate, 100 mM of Tris-HCl pH 7.0, 250 mM of NaCl, 10 % of glycerol and 25% w/v of PEG 3000. We tested their diffraction on the 25th of March 2012 on the PXIII beamline (SLS Synchrotron, Villigen, Switzerland). Unfortunately, diffraction patterns were characteristic of salt crystals, most probably calcium acetate crystals.

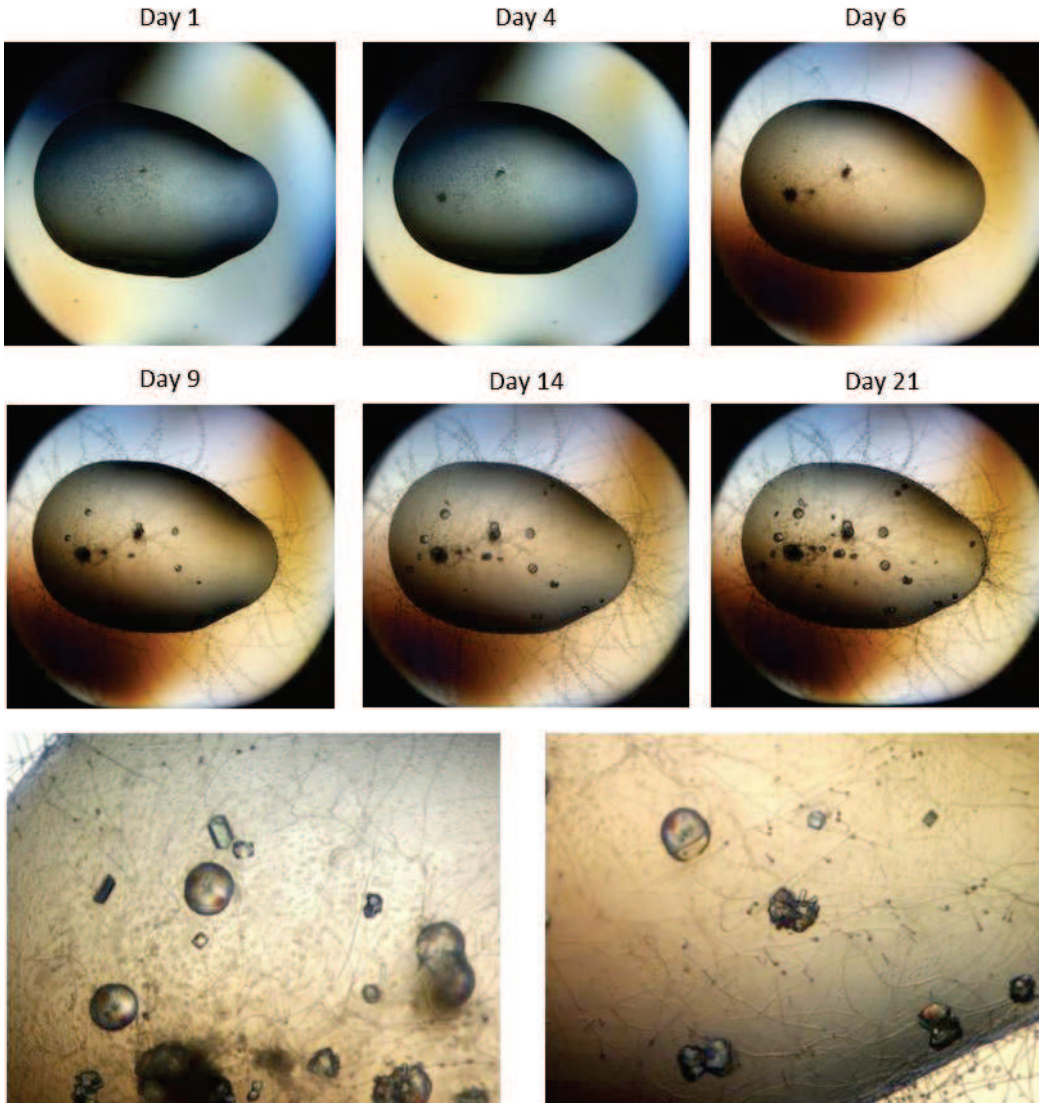


FIGURE 71

MBD3 N-His-TEV: crystallization

After a few days, crystals appeared and grew in a commercial screen condition: The Wizards I+II, condition F6 (200 mM calcium acetate, 100 mM Tris-HCl pH 7.0, 20% PEG 3k, final pH 7.4). These were harvested and frozen, but turned out to be salt when tested in the SLS (Villigen, Switzerland).

1.4 Mass spectrometry analysis

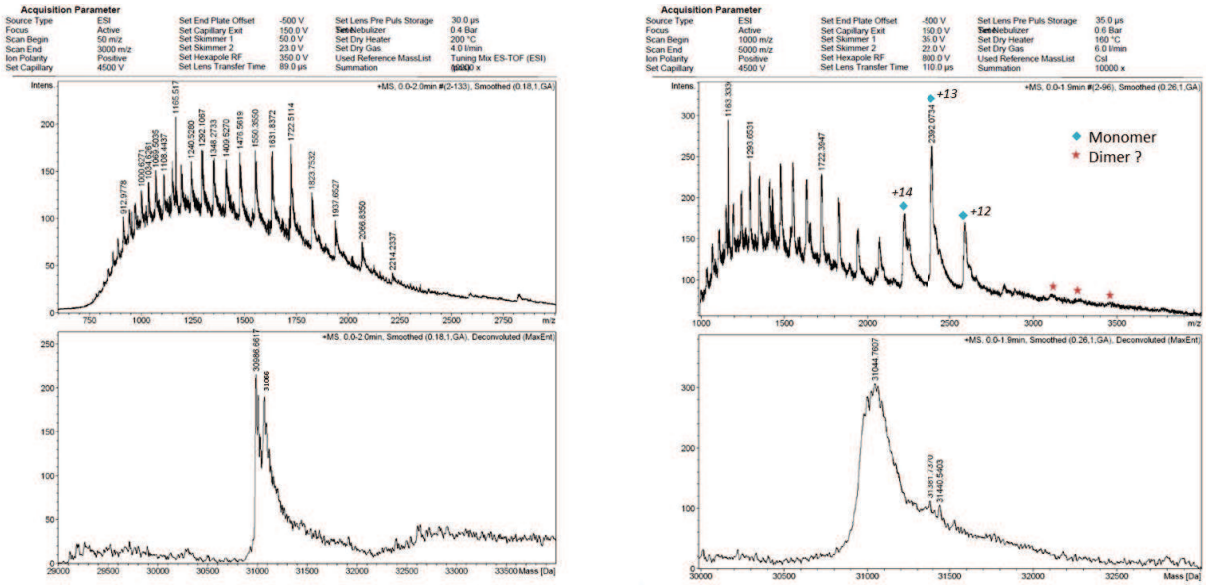
With the aim of confirming the dimerisation of MBD3, I chose to carry out mass spectrometry analysis in the proteomic service of the IGBMC. Three different techniques were used: a MALDI-TOF analysis of MBD3 on SDS-PAGE gel, to get a peptidic coverage and ensure protein integrity; a denaturing ESI-TOF analysis on MBD3 in solution, to determine its exact molecular weight; and finally, a native ESI-TOF analysis to highlight the existence of a dimer *in vitro*.

ESI-TOF requires particular working conditions, especially, the complete absence of detergent which results in high background noise, as well as a change of salt condition, to a low concentration of ammonium acetate (between 25 and 200 mM). The poor stability of MBD3 made these conditions hard to achieve. Tween-20, used during this purification, is hard, if not impossible, to remove by dialysis or gel-filtration. I thus had to redo the purification from the beginning without including any detergent in the solutions, which led to a slightly poorer yield (250 µL at 3.85 mg/mL from 1 L culture). Besides, when exchanging the salt from NaCl to 200 mM of ammonium acetate, the sample precipitated and I could recover only 0.22 mg/mL in 150 µL, i.e., 7 µM of sample, which is very low for efficient ESI-TOF measurements. Some results could nevertheless be obtained: the denaturing analysis showed the existence of two species of 30986 Daltons and 31066 Daltons, i.e., 80 Daltons difference between both (*figure 72*). This difference could result from a phosphorylation of the protein, all the more likely since there are three phosphorylatable serines in MBD3 (S56, S85 and S144). In native conditions, only a monomer of MBD3 could be observed, although the background noise suggested, but without any certainty, the presence of a dimer. Salt exchange seemed to greatly disturb the stability of the protein, which could thus have dissociated.

However, these results posed a major problem. MBD3, with its 6x-His-tag and TEV cleavage site, is around 34500 Daltons, meaning that 3500 Daltons were missing, maybe due to the proteolysis of one extremity.

MALDI-TOF analysis was carried out by cutting the corresponding band out of an SDS-PAGE gel, to extract the protein. After tryptic digestion, the 22 analysed peptides allowed us to create a coverage map. This map showed that MBD3 was not proteolysed at its C-terminal extremity, but that the 45 first amino acids at the N-terminal end were not covered (*figure 73*). However, successful affinity purification using nickel indicated that the N-terminal part was intact since the 6x-His-tag was still functional. Moreover, this analysis did not show any post-translational modification on the covered serines. All these results brought together led to the conclusion that I had not been working with the main isoform of MBD3, but rather with its short isoform MBD3Δ. Sequencing of the expression vector confirmed this hypothesis: the cDNA cloned corresponded to MBD3Δ.

Following this observation, and in spite of the time and work I invested on this protein, I decided to stop studying MBD3Δ and to undertake the design of new expression vectors, to produce the long isoform of MBD3. In the following, the term “MBD3” will only refer to the long isoform of the protein.



Denaturing condition
1% Formic acid in acetonitrile

Native condition
200 mM ammonium acetate

FIGURE 72

MBD3 N-His-TEV: ESI-TOF analysis

Mass-spectrometry analysis in denaturing conditions revealed the existence of two species of 30986 and 31066 Daltons. In native conditions, only a MBD3 monomer could be observed (blue spots), although the background noise suggests the presence of a higher molecular weight species (red stars).

	6xHis-tag	*TEV site*	
	MI HHHHHH	ENLYFQG LE	**ERK**RWECPA
11	LPQGWEREEV	PRRSGLSAGH	RDVFYY**SPSG**
41	**KKFR**SKPQLA	RYLGGSMDLS	TFDFR**TGKML**
71	**MSKMNKSRQR**	VRYDSSNQVK	GKPDLNTALP
101	VRQTASIFKQ	PVTKITNHPS	NK**VK**SDPQKA
131	VDQPRQLFWE	KKLSGLNAFD	IAEELVK**TMD**
161	**LPKGLQGVGP**	**GCTDETLLSA**	**IASALHTSTM**
191	**PITGQLSAAV**	**EK**NPGVWLNT	TQPLCKAFMV
221	TDEDIRKQEE	LVQQVRKRLE	EALMADMLAH
251	VEELARDGEA	PLDKACAEDD	DEEDEEEEEE
281	EPDPDPEMEH	V	

FIGURE 73

MBD3 N-His-TEV: MALDI-TOF analysis

The sequence of MBD3 N-His-TEV is represented.

In dark blue, the 6xHis tag and in light blue, the TEV cleavage site.

In bold fonts, the sequence of MBD3 and in red, the peptide coverage map defined by on-gel mass-spectrometry analysis. This coverage map lacks the N-terminal part of MBD3.

In grey, the 32 amino acids that are missing the short isoform of MBD3, MBD3Δ.

Underlined are two key residues: F34 is responsible for the loss of affinity towards methylated DNA; and R23 has been shown to be mutated into methionine in neurological diseases.

2. Studying MBD3 full-length

2.1 Designing new vectors

In the frame of this new study that I started on the long isoform of human MBD3, I decided, for convenience purposes, to work with the bacterial expression system. Mass spectrometry on MBD3Δ had indeed previously shown that no post-translational modifications were involved, making bacteria suitable for expression of this protein, not to mention rapid implementation and ease of cultivation. I chose to use a pET28b plasmid, modified to express a 6x-His-tag and 3C cleavage site in the N-terminal. This plasmid carries the KanR gene, encoding an aminoglycoside 3'-phosphotransferase which confers resistance to kanamycine for selection. The gene of interest can be subcloned in frame using XhoI and NotI restriction sites.

After unsuccessful searching in cDNA libraries for the human gene of the long isoform of MBD3, I decided to order four synthetic genes from GenScript:

- Human full-length MBD3, residues 2 to 291 (missing the ^{ini}Met) with XhoI and NotI sites;
- Human full-length MBD3, residues 2 to 291 (missing the ^{ini}Met), with the point mutation F34Y, with XhoI and NotI sites;
- Human MBD3ΔC, residues 2 to 267 (missing the ^{ini}Met) with XhoI and NotI sites;
- Human MBD3ΔC, residues 2 to 267 (missing the ^{ini}Met) with the point mutation F34Y, with XhoI and NotI sites;

The F34Y point mutation corresponds to the key residue involved in methylated DNA recognition. In mammals, the phenylalanine 34 is responsible for the loss of specificity towards methylation, whereas the tyrosine 34 in invertebrates allows high binding specificity to 5mC.

The MBD3ΔC deletion mutant lacks the C-terminal tail (residues 268 to 291). This tail is highly acidic (18 acidic residues out of 24) and is responsible for a significant drop of pI: MBD3 full-length has a pI around 5.2, while the ΔC mutant has a pI above 9.2.

These genes were codon-optimised for optimal production in both bacteria and insect cells, using online-tools. In this way, I could obtain a codon adaptive index (CAI) above 0.7 (1.0 being ideal) for bacterial expression and a GC content between 30 and 70 % (65%).

Finally, I also subcloned in anticipation these four genes in a pFastBac1 vector with a 6x-His-tag and TEV cleavage site, for future expression using the baculovirus system, the CAI being around 0.8 for expression in insect cells.

I first carried out expression tests on each of the clones I obtained with the four different synthetic genes, using *E. coli* BL21(DE3) cells, and cultivating cell in 2xLB and AI media. All of them showed a high overexpression of the protein, with a slight advantage for AI medium (*figure 74*).

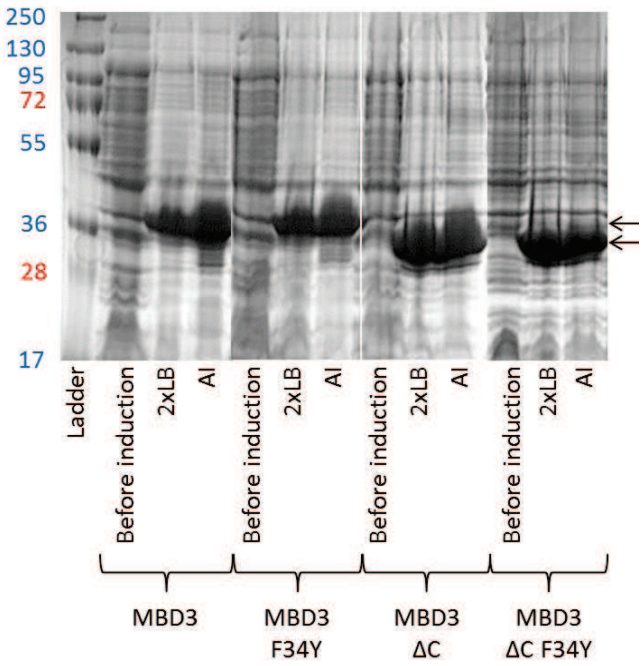


FIGURE 74

MBD3 constructs: expression tests

SDS-PAGE, Coomassie staining

Four constructs have been designed to express MBD3 in bacterial expression system: the full-length protein, namely MBD3, as well as three mutants (F34Y, ΔC that lacks the C-terminal acidic tail, and F34YΔC).

Two culture media have been tested in parallel: 2xLB and auto-inducible medium (AI), and showed high overexpression, with an advantage for the AI medium.

2.2 First purification assays

For these first purification assays, I chose to use the same buffer conditions as described previously for MBDΔ expressed in insect cells. After cultivation in AI medium at 37°C, the cells were harvested by centrifugation at 4000 Xg, and lysed in a lysis buffer (500 mM of NaCl, 20 mM of Tris-HCl pH 7.5, 5 % of glycerol, 4 mM of CHAPS, 1 mM of TCEP and 10 mM of imidazole). After Dounce homogenization, the sample was sonicated on ice for 15 minutes, at 50 % amplitude, alternating 1 second of sonication and 1 second of break. The sample was then centrifuged at 25000 Xg for 45 minutes. Though a lot of protein remained in the cell pellet, the supernatant was loaded on a HisTrap FF 5 mL column connected to an Äkta Purifier system, and a stepwise elution was done with 10, 20, 50, 200 and 500 mM of imidazole. MBD3 could thus be nicely separated from many contaminants as shown on the SDS gel, at 200 mM of imidazole (*figure 75*). In spite of various optimisations tested, further purification could not be achieved due to low yield after the first step of purification and precipitation problems. Among these optimisations, different culture conditions were tested with high or low concentration of IPTG (0.1 mM and 1 mM) and addition of glucose to the medium. Also, glucose and trehalose were added directly in the lysis buffer, for their effect on

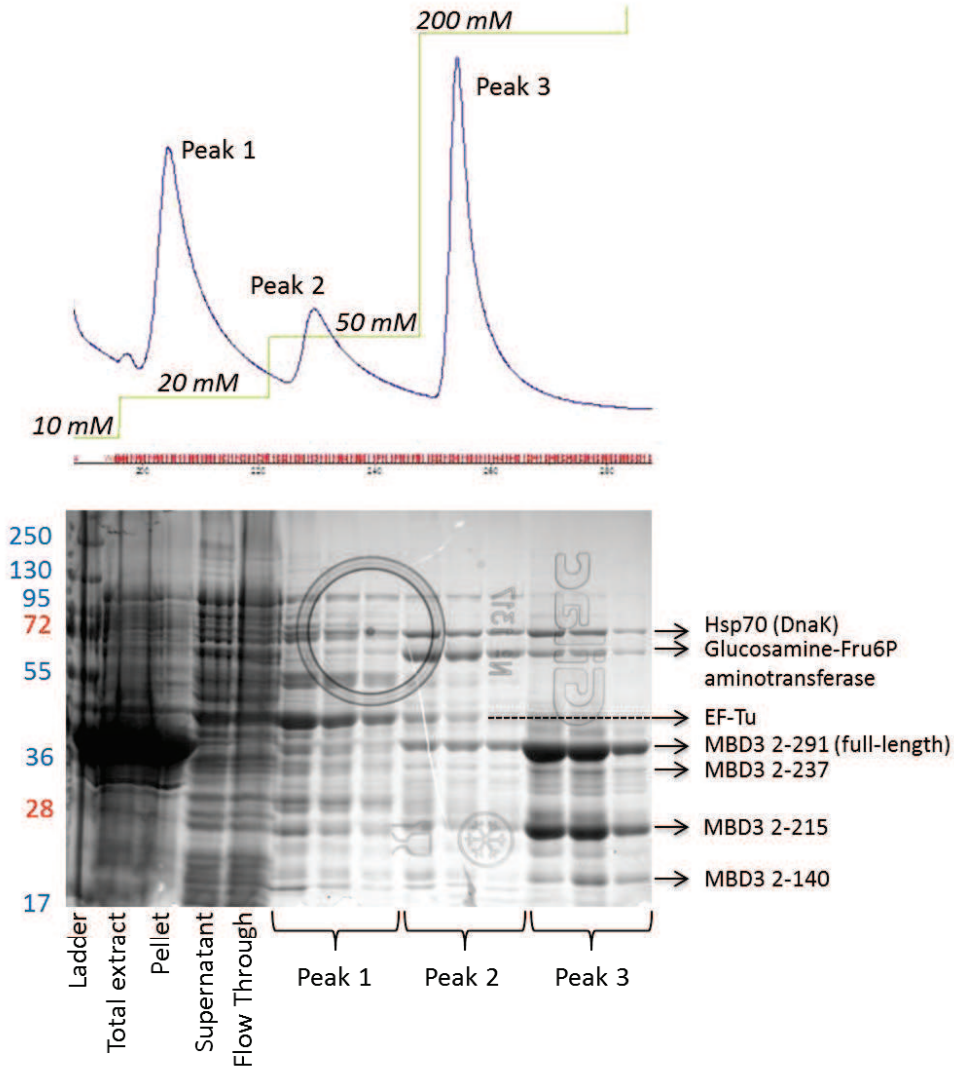


FIGURE 75

MBD3 N-His-3C: solubility test

Top: affinity chromatography profile (HisTrap FF 5 mL)

Below: SDS-PAGE, Coomassie staining

The SDS gel shows that, though a significant amount of protein remains in the cell pellet, MBD3 could be purified by affinity chromatography, using 200 mM of imidazole.

On-gel mass-spectrometry analysis revealed the presence of the full-length protein MBD3 (2-291), but also presence of degradation products (2-237, 2-215 and 2-140). Also, the heat shock protein DnaK is shown to be overexpressed and copurifies with MBD3. This suggests that MBD3 might not be correctly folded.

---

recombinant protein stability, but without success. Finally, the use of a hypotonic buffer with 10 mM salt or hypertonic buffer with 1 M salt did not help solubilizing the protein.

An SDS gel of the first purification step was analysed by MALDI-TOF, and showed important degradation of MBD3, whether it was the full length construct or MBD3ΔC. Also, this analysis

revealed expression of DnaK, the major bacterial Hsp70 (*figure 75*). This heat shock protein interacts with hydrophobic patches of newly synthesised proteins, preventing their aggregation before complete folding. Co-purification of this chaperone suggests thus that MBD3 might not be completely folded and still exhibits hydrophobic patches.

After several trials, MBD3 remained mostly insoluble after cell lysis and highly unstable during purification. I decided thus to pursue the purification of MBD3 in denaturing conditions.

2.3 Purification under denaturing conditions

In the case of MBD3, a major part of the protein was insoluble, found in the pellet after cell lysis. Furthermore, co-purification of DnaK together with MBD3 instability upon purification suggest that it is misfolded to some extent. All these factors led to believe that MBD3 could be found in inclusion bodies.

Protein purification in denaturing conditions starts with the isolation of inclusion bodies, followed by their lysis using a strong denaturing agent. Then, the released protein is purified using classical liquid chromatography methods, and finally refolded in vitro by gentle buffer exchange, often achieved by dialysis.

A first purification trial was achieved, by dissolving the inclusion bodies in a buffer containing 8 M of deionized urea, 20 mM of Tris-HCl pH 7.5, 5 mM of BME and 5 % of DMSO. The latter is an organic solvent with a strong dipole moment, that perturbs the structure of water molecules and thus, destabilising protein and nucleic acid structures. After ultracentrifugation and filtering of the supernatant, it was loaded on a HiTrap Q 5mL column (GE Healthcare) and eluted with a gradient of salt concentration. The recovered protein was nearly pure as shown on the SDS gel (*figure 76*). To polish further the protein sample, it was loaded onto a Sephacryl S-200 26/60 HR size-exclusion column (GE Healthcare), and eluted with a buffer containing 8 M of deionized urea, 100 mM of NaCl, 20 mM of Tris-HCl pH 7.5 and 5 mM of BME. However, MBD3 was eluted in the void volume and the spectrophotometer profile measured with the Nanodrop showed two peaks, at 230 nm corresponding to proteins, and 275 nm corresponding to DNA. Thus, in spite of the high concentration of urea, MBD3 was still contaminated with nucleic acids.

After several attempts to get rid of the DNA, by increasing the DMSO concentration in particular, I decided to change the denaturing agent to guanidinium chloride. The latter is charged while urea is neutral. This has an effect on the stability of a protein and competition for hydrogen bonds, making it a stronger denaturing agent than urea. However, because it is charged, guanidinium chloride is not compatible with ion-exchange chromatography and thus needs to be exchanged against urea. Inclusion bodies were therefore dissolved in a buffer containing 7M of guanidinium chloride, 20 mM of Tris-HCl pH 7.5, 5 mM BME and 10 % of DMSO. The sample was first loaded onto a Sephacryl S-200 26/60 HR size-exclusion column and eluted with 8 M of deionized urea, 20 mM of Tris-HCl pH 7.5 and 5 mM of BME, but again, MBD3 was eluted in the void volume

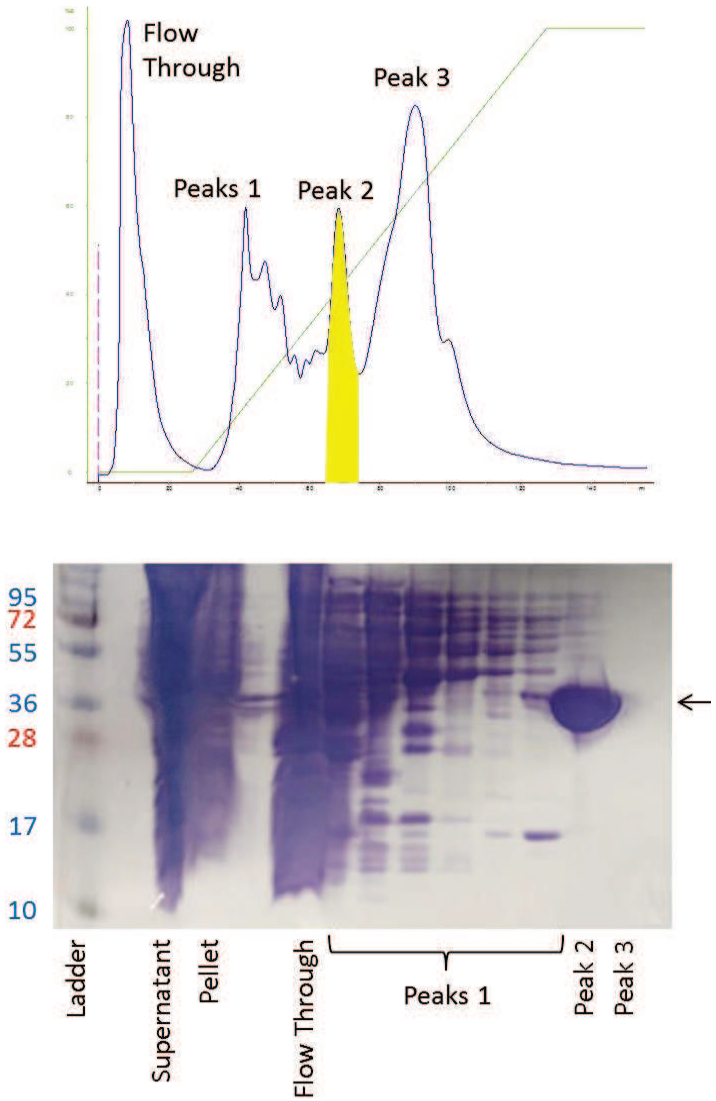


FIGURE 76

MBD3 N-His-3C: purification under denaturing conditions

Top: ion-exchange chromatography profile (HiTrapQ 5 mL)

Below: SDS-PAGE, Coomassie staining

After dissolving inclusion bodies in 8 M urea, the clarified supernatant was loaded onto a HiTrapQ column. An increasing gradient of salt concentration allowed to isolate several peaks, among which one contained the pure MBD3 protein (eluted around 400 mM of NaCl, shown in yellow).

---

together with DNA contamination. Nevertheless, I decided to pursue the purification with ion-exchange chromatography, using the same buffer and eluting with a salt gradient. MBD3 could then be recovered at high purity, but DNA contamination was still present. This sample was however used for refolding trials.

Finally, I managed to get rid of the DNA, by applying the same protocol as previously described and adding 1 M of NaCl in the buffer. The size-exclusion chromatography profile showed a peak corresponding to DNA in the void volume, followed by a shoulder corresponding to MBD3. The

latter was pooled and dialyzed against a low salt buffer, than loaded onto a HiTrap Q 5mL column. MBD3 could be eluted as previously described around 430 mM of NaCl. Finally, a polishing step using the same size-exclusion chromatography column using 8 M of deionized urea, 200 mM of NaCl, 20 mM of Tris-HCl pH 7.5 and 5 mM of BME could confirm the high purity of MBD3, cleared of any DNA contamination (*figure 77*).

Both samples, with and without DNA contamination, were lyophilized after water-dialysis and nitrogen-freezing. As expected, a lot of precipitation appeared during dialysis against water but an SDS-gel showed that most of the precipitate was composed of protein contaminants, while MBD3 was still soluble in water. This observation is also typically made in the case of histones, which remain soluble in water. Thus, MBD3 showed great promise for refolding trials.

These two samples yielded to 200 to 300 mg of pure protein each, from 3 L of bacterial culture. Such high amounts of protein could be used for refolding trials and allowed to test various refolding conditions. A standard protocol consisted of resuspending the protein powder in a buffer containing 8 M of deionized urea, 1 M of NaCl, 20 mM of Tris-HCl pH 7.5 and 5 mM of BME, followed by a slow continuous-flow dialysis of the sample at 4°C against 1 M of NaCl, 20 mM of Tris-HCl pH 7.5 and 5 mM of BME, and finally, a stepwise dialysis to lower the salt concentration. Interestingly, I noted that no precipitation was visible at low salt concentration in the DNA-contaminated sample, while precipitation started to appear around 250 mM of NaCl in the non DNA-contaminated sample. These observation suggested a DNA-induced stabilisation of the protein, as one would suspect from its function.

Throughout the refolding process, DLS was used in order to monitor the behaviour of the protein. Unfortunately, only very large aggregates could be detected. This refolding procedure was optimised, using several refolding agents, such as L-arginine, oxidized and reduced glutathione, sarkozyl, etc. Refolding was also achieved in parallel at 4°C and room temperature. But after several weeks of attempts, no positive result were obtained. I decided thus to go back to the old native protocol and start optimisation again.

2.4 How Thermofluor® helped preventing aggregation

The protocol for MBD3 purification in native state was redesigned after unsuccessful attempts to refold the protein *in vitro*. Several optimisations were made at different levels of the protocol, starting with the culture. In order to slow down protein production and avoid excessive insolubility, I decided to cultivate the cells at 37°C until OD_{600nm} = 0.6, then lower the temperature to 15°C prior to induction with 0.4 mM final of IPTG for overnight cultivation. Considering the low amount of soluble material available, higher volumes of culture were required to achieve decent quantities of protein for further studies. I thus started to made use of our bioreactors (Techfors-S 20 L and Techfors 100 L, Infors-HT) for very large-scale production, and routinely treated 20 to 25 L of cells for each purification.

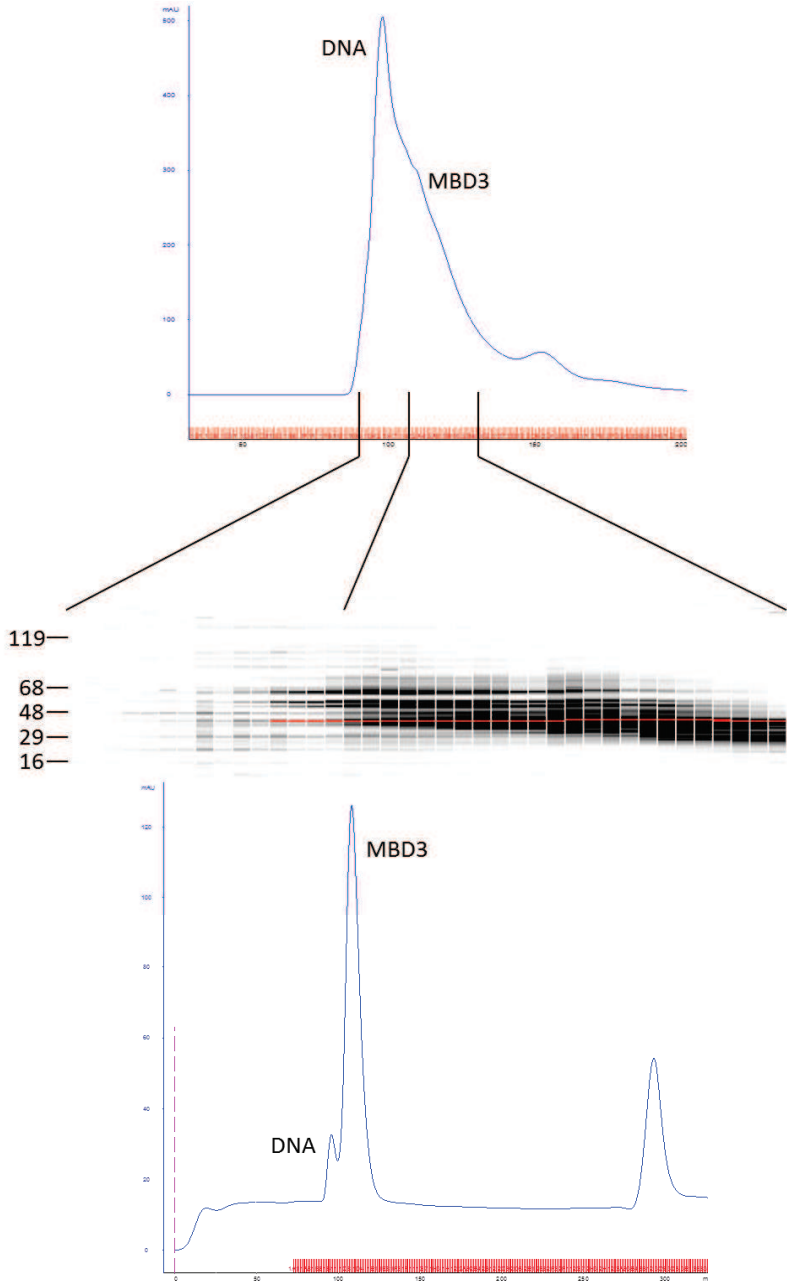


FIGURE 77

MBD3 N-His-3C: purification under denaturing conditions (2)

Top: gel-filtration profile (Sephacryl S200 26/60 HR)

Middle: Microfluidic capillary electrophoresis, LabChip GXII

Below: gel-filtration profile (Sephacryl S200 26/60 HR)

By adding DMSO and 1 M of NaCl to the dissolving buffer, DNA contamination could be removed as shown on the first GF profile above. After a second step of ion-exchange, a third gel-filtration confirmed the total absence of DNA in the protein sample.

The cell lysis technique was also readjusted, with chemical lysis being achieved with detergent and lysozyme, to avoid insofar as possible, excessive mechanical stress on the sample. After several comparative trials between NP-40, Triton-X, and CHAPS, the latter was preferred to achieve efficient cell lysis. A concentration of 2.5 mg of lysozyme per litre of culture were also added directly to the resuspended cell pellet, and incubated for 30 minutes at 4°C. High efficiency cell lysis however went together with high viscosity due to genomic DNA release. Sonication was thus still needed in order to disrupt DNA and no chemical-only lysis could be achieved.

Finally, the purification process itself could be continued after a first nickel-affinity step. Ion-exchange chromatography had been considered but dialysis against a low-salt buffer led to precipitation of the sample and significant losses. A second affinity step using a HiTrap Heparin 5 mL column was thus preferred. This chromatography technique combines the advantages of both affinity and ion-exchange chromatography. MBD3, as a DNA-binding protein, could interact with heparinized resin and be eluted at high salt concentration. A soft dialysis step to lower the salt concentration to around 250 mM prior to loading onto this column was enough to allow MBD3 to bind. The protein was eluted with a salt concentration gradient, with a peak being seen around 400 mM of NaCl at pH 7.5 and 300 mM of NaCl at pH 8.5. The quality of the sample was checked by SDS-PAGE and showed high purity. Thus, this step being successful, a final gel-filtration step was used to remove aggregates and recover native and functional protein. However, only one peak could be observed on the chromatogram, in the void volume, indicating full aggregation of MBD3.

After unsuccessful trials, leading consistently to fully aggregated proteins, I decided to carry out Thermofluor® experiments to optimise the buffer conditions in which MBD3 could be stabilised. Generally, this technique is only used on native and stable proteins. Binding of the fluorochrome Sypro Orange does not occur at low temperature as the protein is folded and not aggregated, and fluorescence appears only with thermally-induced denaturation. One could thus predict that working with an aggregated sample would lead to a high fluorescence signal even at low temperature unless the protein is stabilised by the buffer condition tested.

A 48-condition screen was designed to test four different buffers at different pH, together with increasing salt concentration. The composition of this screen is shown in the following table.

<table>
<tr><th></th><th>300 mM
Na/K</th><th>200 mM
Na/K</th><th>100 mM
Na/K</th><th>200 mM
$LiSO_4$</th><th>200 mM
Na/K</th><th>200 mM
Na/K</th></tr>
<tr><td>25 mM
MES pH 6.5</td><td rowspan="4"></td><td rowspan="4"></td><td rowspan="4"></td><td rowspan="4"></td><td rowspan="4"></td><td rowspan="4"></td></tr>
<tr><td>25 mM
HEPES pH 7.5</td></tr>
<tr><td>25 mM
Tris-HCl pH 8.0</td></tr>
<tr><td>25 mM
Tris-Maleate pH 7.4</td></tr>
<tr><td></td><td colspan="3">2 % glycerol
1 mM CHAPS
5 mM $MgSO_4$</td><td colspan="2">2 % glycerol
5 mM $MgSO_4$</td><td>1 mM CHAPS
5 mM $MgSO_4$</td></tr>
</table>

After mixing the aggregated protein sample together with the different conditions of the screen and addition of Sypro Orange, a temperature gradient from 20 to 95°C was carried out, and fluorescence was measured in the RT-PCR thermocycler. As explained earlier, working with aggregates causes a high fluorescence signal from the start of the experiment. Thus, the two parameters to be taken into account for proper interpretation of the results are the reduction of the fluorescence signal at room temperature and the shift of melting temperature towards medium-to-high temperatures. These two necessary requirements were nicely fulfilled in two conditions (*figure 78*):

- 25 mM of MES pH 6.5, 200 mM of NaCl/KCl, 2 % of glycerol, 1 mM of CHAPS and 5 mM of $MgSO_4$;
- 25 mM of Tris-Maleate pH 7.4, 100 mM of NaCl/KCl, 2 % of glycerol, 1 mM of CHAPS and 5 mM of $MgSO_4$;

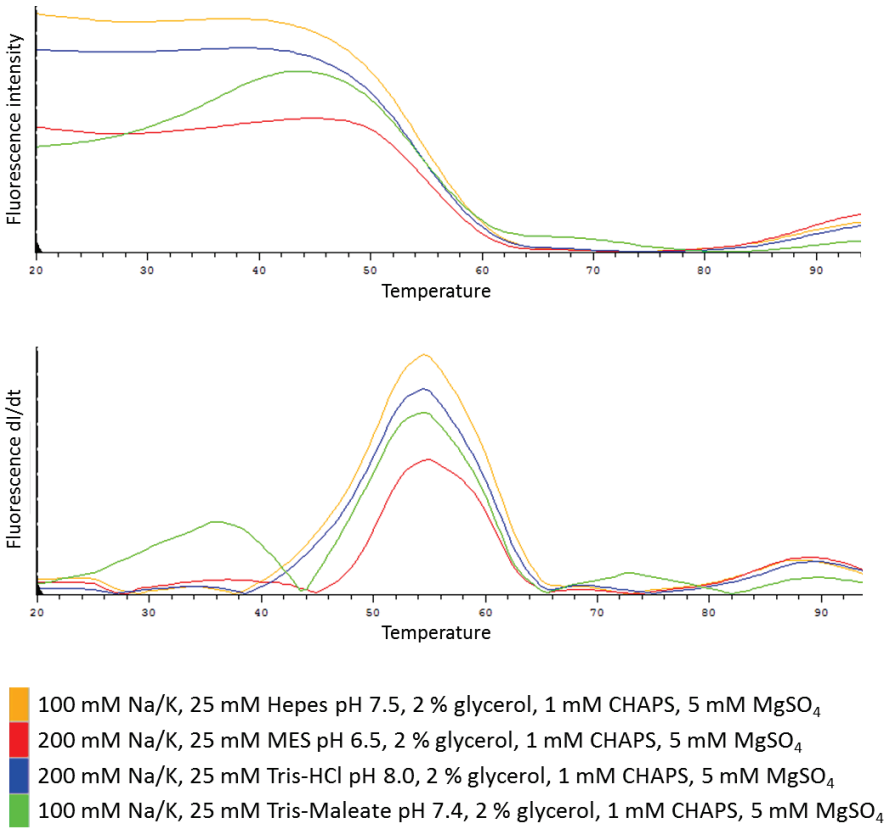


FIGURE 78

MBD3 N-His-3C: Thermofluor®

Four curves are shown on this Thermofluor® profile, each corresponding to another tested buffer. As expected when working with an aggregated protein, a high fluorescence signal could be observed even at low temperature. Two parameters are thus to be taken into account on this profile: first, the lowering of the fluorescence signal at low temperature; and second, the highest melting temperature, defined by dI/dt = 0. Two curves fulfilled these conditions: the green one (Tris-Maleate pH 7.4), with Tm = 42°C; and the red one (MES pH 6.5), with Tm = 43°C. The two others showed high fluorescence at low temperature and Tm = 39°C.

Following these results, I tried to purify MBD3 by changing the buffer to MES pH 6.5 instead of Tris-HCl. The rest of the protocol remained essentially the same, with a first step of affinity chromatography using Ni-NTA resin, a second step of affinity with a HiTrap Heparin column, and finally, a gel-filtration step. As expected from the previous observations, the lower the pH, the higher the salt concentration for MBD3 elution on heparinized resin. Indeed, in MES pH 6.5, MBD3 could be eluted between 500 and 600 mM NaCl, whereas 400 mM were enough to elute it in Tris-HCl pH 7.5 and 300 mM in Tris-HCl pH 8.5. More interestingly, the gel-filtration profile didn't show any aggregation of MBD3 but instead, two overlapping peaks (*figure 79*). As shown on the SDS gel, the first peak corresponded to the full-length MBD3 (elution volume ≈ 12.7 mL ≈ 70-80 kilo Daltons) whereas the second one contained smaller proteins (elution volume ≈ 16 mL ≈ 30 kilo Daltons), most certainly degradation products. Furthermore, DLS experiments were carried out on these two protein samples, and showed that the first peak corresponded to proteins between 3.6 and 3.9 nm in size for a mass of approximately 70 to 80 kilo Daltons, i.e., the size of an MBD3 dimer. As for the second peak, DLS measurements showed a small, 1.7 to 2.1 nm species, for a mass of 15-20 kilo Daltons, which corresponds to the SDS gel results (*figure 79*).

Together, these results led to several suppositions. First, MBD3 is more stable in low pH buffers like MES pH 6.5, and doesn't form any aggregates. Secondly, it is eluted on the gel-filtration as a dimer, which is consistent with the DLS measurements carried out, as well as the literature which mentions MBD3 dimerisation[16]. Thirdly, the second peak eluted on the gel-filtration column could correspond to a monomer of MBD3 but SDS gel and DLS measurements showed that it corresponds instead to smaller products. It could thus be that MBD3 is highly unstable as a monomer and degrades rapidly if not bound to DNA for example, whereas the MBD3 dimer could be a self-stabilised form, awaiting DNA to dissociate and bind. These new and encouraging results invited me to persevere in this direction and further optimise the purification process of MBD3.

2.5 Last optimisations and final purification protocol

Optimising the purification of a protein is very much a matter of trial and error. This was the case for MBD3. Several weeks were thus necessary to find the right conditions, using MES buffer pH 6.5, to achieve good quantities of pure protein.

The first optimisation that could be achieved was to get rid of the dialysis between the first two chromatography steps. Indeed, considering the instability of the protein over time, purifying it should require as little time as possible. One night of dialysis could thus be saved, made possible by the higher salt concentration needed to elute the protein on heparinized column. From 300 to 400 mM of NaCl required in Tris-HCl buffer, the changeover to MES pH 6.5 increased this concentration to 500 to 600 mM. More than the salt concentration itself, the important parameter in this heparin affinity step was the conductivity of the sample and buffer. I could observe over my different trials that MBD3 was always eluted around 39 mS/cm². The sample should then have a lower conductivity in order to bind to the resin. Knowing this, I could make use of a conductimeter to check for

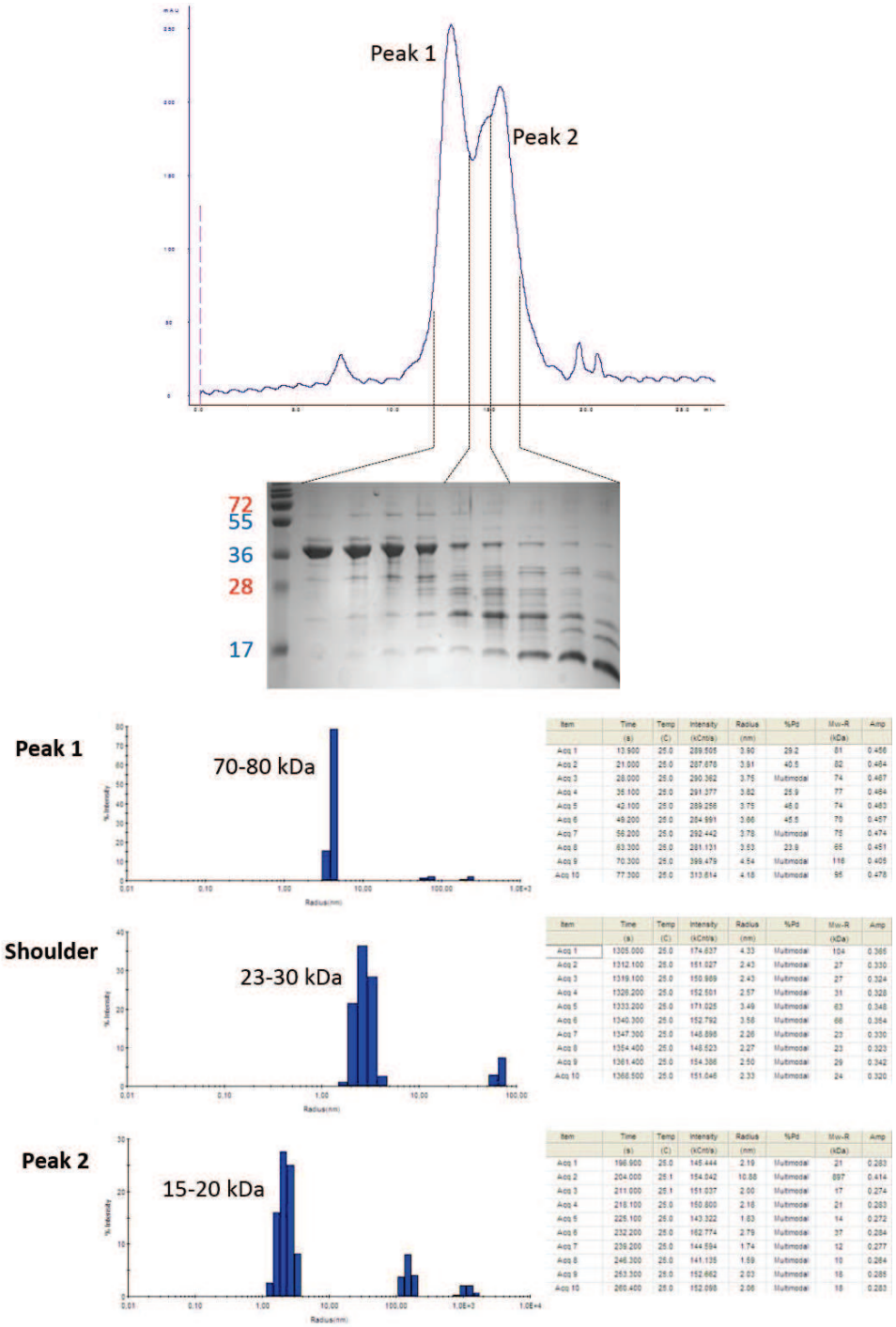


FIGURE 79

MBD3 N-His-3C: purification in MES pH 6.5

Top: gel-filtration profile (Superdex 200 10/300 GL)

Middle: SDS-PAGE, Coomassie staining

Below: DLS measurements

After affinity chromatography using heparinized resin, MBD3 could be purified by gel-filtration. The profile shows a first peak corresponding to MBD3 full-length (as seen on the gel), and a second peak with a shoulder, containing mainly degradation products.

DLS measurements were carried out at 20°C and showed, despite high polydispersity, a dimeric species in the first peak (70-80 kilo Daltons).

conductivity whenever I tested new buffer optimisations. Lysis was thus finally achieved in a medium concentration salt buffer (450 mM of NaCl), and the eluate recovered after nickel affinity could be directly loaded onto the HiTrap Heparin column, without prior removal of salt or imidazole. The conductivity of the sample was then 37 mS/cm², in the presence of 450 mM of NaCl and 300 mM of imidazole.

A second optimisation related to the use of magnesium in the buffer. When studying DNA-binding proteins, it is common to use magnesium as it is often involved in DNA stabilisation. It was therefore natural to add a small amount of $MgSO_4$ in the buffers, for future complex formation. However, and in spite of the use of protease inhibitor along the whole purification process, cleavage of MBD3 could be observed after a few days at 4°C. As shown on the gel, this cleavage appeared to be clean, leading to only two products (*figure 80*). This, together with the constant usage of protease inhibitor, suggests an autoproteolytic activity of MBD3 in the presence of magnesium. More interestingly, this cleavage was not observed when magnesium was replaced with EDTA, even after a couple of weeks at 4°C. MALDI-TOF analysis of these two products showed a peptide coverage from residues 2 to 102 (plus the 6xHis-tag and TEV cleavage site) for a 13.804 kilo Dalton fragment, and from 115 to 291 for a 19.662 kilo Dalton fragment.

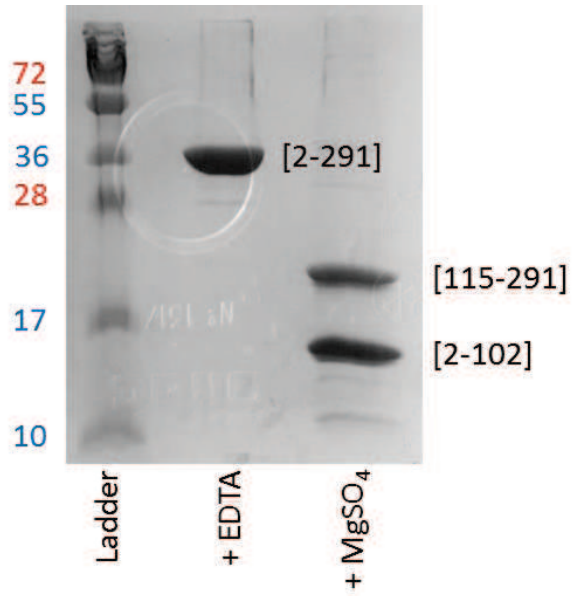


FIGURE 80

MBD3 N-His-3C: $MgSO_4$ vs. EDTA

SDS-PAGE, Coomassie staining

After a few days at 4°C, two samples of pure MBD3 containing either $MgSO_4$ or EDTA as an additive, were loaded on a SDS gel and further analysed by on-gel mass spectrometry.

The gel revealed a proteolysis effect in presence of magnesium, that could be avoided by addition of EDTA. Furthermore, mass spectrometry allowed to cover the cleaved domains.

Thirdly, resolution of the overlapping peaks separated on Superdex 200 10/300 GL size-exclusion column could be further improved by the use of two Superdex 200 10/300 GL columns connected in series. The final bed dimension was extended to 600 mm for the same diameter (10 mm), allowing a high-resolution separation. Three peaks could thus be nicely separated: Ve_1 ≈ 26.2 mL ≈ 80 kilo Daltons; Ve_2 ≈ 30.0 mL ≈ 28 kilo Daltons; and Ve_3 ≈ 32.1 mL ≈ 16 kilo Daltons. These results were consistent with the previously observed ones but allowed complete isolation of the MBD3 dimer.

Lastly, precipitation issues during purification could be avoided by constantly working at 4°C and below. The extreme sensitivity of MBD3 indeed caused major problems regarding precipitation, and effective ways to handle this protein had to be found. Unfortunately, most of the biophysical experimentations can only be carried out at room temperature, thereby severely limiting the possibilities of study. In particular, DLS and SEC-MALLS, which are two major techniques used to appreciate the size and homogeneity of a sample, could not be carried out on MBD3.

Summarizing the purification process of MBD3 full-length, 10 to 20 litres of cell culture were treated using a lysis buffer (450 mM of NaCl, 50 mM of MES pH 6.5, 3 mM of CHAPS, 2 mM of BME, PMSF and cOmplete EDTA-free). After complete resuspension, the sample was lysed using 2.5 mg of lysozyme/L of culture, and incubated for 30 minutes at 4°C, prior to brief sonication. The lysate was clarified by centrifugation at 50000 Xg for 1 hour, and the supernatant was allowed to bind to Ni-NTA resin. After one hour of incubation time, the resin was washed with a washing buffer (450 mM of NaCl, 50 mM of MES pH 6.5, 10 mM of imidazole, 3 mM of CHAPS and 2 mM of BME), then poured into a polypropylene gravity flow column. After further washing, the protein was eluted using an elution buffer (450 mM of NaCl, 50 mM of MES pH 6.5, 300 mM of imidazole, 2 mM of CHAPS and 2 mM of BME), and loaded onto a HiTrap Heparin 5 mL column. A stepwise-elution (450 mM, 550 mM and 610 mM of NaCl) was carried out using a low salt and high salt buffer (0 mM/1M of NaCl, 50 mM of MES pH 6.5, 5 mM of EDTA, 1 mM of CHAPS and 2 mM of BME). The recovered protein was finally concentrated using Vivaspin Turbo 15 (Sartorius) concentrators until 0.5 to 1 mL in volume, and injected in two pre-equilibrated Superdex 200 10/300 GL columns in series (GF buffer: 500 mM of NaCl, 50 mM of MES pH 6.5, 5 mM of EDTA, 0.5 mM of TCEP). The recovered pure protein could then be used for further biophysical and structural studies.

2.6 Binding studies on nucleosomes

Once MBD3 was purified, I could move forward and use it for binding assays together with reconstituted nucleosome particles. This work was achieved with the help of Dr. Kareem Mohideen, post-doctoral researcher in our team. All the following experiments were carried out at 4°C with pre-cooled material.

Our main dilemma in this work was to find the proper way for mixing both MBD3 and nucleosomes, since the first one was kept in 500 mM of NaCl, whereas the second was stored in a salt-free buffer (20 mM of Tris-HCl pH 7.5, 5 mM of DTT and 1 mM of EDTA). Mixing both as such

would result in MBD3 precipitation by a sudden decrease in the salt concentration, prior to binding and stabilisation on the nucleosome.

For the first trials, we decided to add salt to a final concentration of 400 mM in the nucleosome sample, then mix together nucleosomes and MBD3 in a 1:5 molar ratio. One µL of nucleosome at 9 mg/mL was used for this first experiment, supplemented with 0.8 µL of 5 M NaCl and 7.2 µL of Tris-HCl pH 7.5 (final volume = 9 µL). These 45 pmol of nucleosomes were then mixed with 225 pmol of MBD3 (i.e., 7.8 µg at 2.4 mg/mL = 3.2 µL), leading to a final volume of 12.2 µL and salt concentration of 426 mM. No precipitation was observed and salt could be lowered after a 30-minute incubation time, by addition of a low-salt buffer, to the desired final concentration. Three different salt concentration were thus tested: 400 mM, 350 mM and 300 mM of NaCl. After further incubation, a 5 % EMSA-gel was run in 0.25x TBE buffer, at low voltage. Finally, ethidium bromide staining allowed us to observe a clear shift on the gel between nucleosomes alone and nucleosomes in complex with MBD3, whatever the salt concentration (*figure 81*).

This successful result was reproduced using a new batch of MBD3, following the same protocol as previously described. Two different molar ratios were tested (1:5 and 1:10), as well as several final salt concentrations (400, 325, 250 and 175 mM of NaCl). As seen on the EMSA-gel, the result was not only reproducible, but also showed that the shifting pattern was the same for both ratios, suggesting the 1:5 complex was already saturated (*figure 81*). Plus, the lower the salt concentration, the higher the shift. This observation suggests an equilibrium between the bound and unbound form of the complex, shifted towards the bound form when the shift is more important. We could thus conclude that 175 mM of NaCl seemed to be an ideal salt concentration for the nucleosome-MBD3 complex.

2.7 Crystallisation and structural studies

Once we had defined the molar ratio and the lower salt concentration reachable for the MBD3-Nucleosome complex, we could carry out first crystallisation assays. However, the protocol for complex formation had to be further optimised. Indeed, biophysical studies like EMSA require only little amount of complex, at low concentration. That is not the case for crystallisation, which requires not only higher amounts of sample but also higher concentrations. Considering the instability of this complex over time, we decided to spare it by avoiding as much as possible concentration steps. Great efforts had thus to be made to concentrate both nucleosome and MBD3 separately. After several operating errors, leading to sample precipitation, I finally managed to get highly concentrated MBD3. Nucleosomal particles on the other hand also showed concentration issues, however different from that encountered with MBD3. Indeed, DNA has the natural tendency to stick to concentration membranes and significant amounts of nucleosomes would be lost. In both cases, several concentrors have been tested and helped improving the final yield. In particular, great differences have been observed between the commonly-used regenerated cellulose membranes and polyethersulfone membranes. The first ones allowed to reach around 3 to 5 mg/mL prior to MBD3

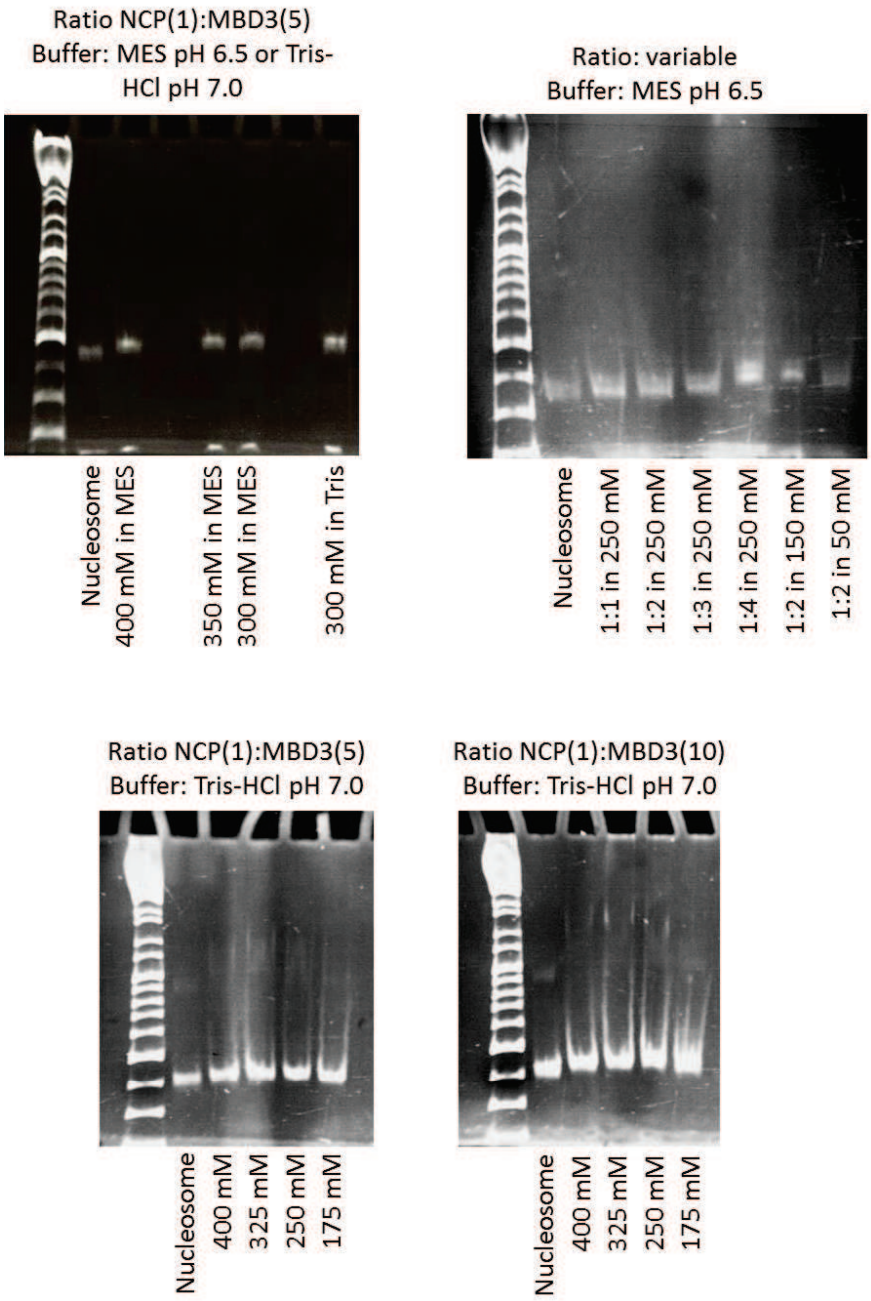


FIGURE 81

MBD3 N-His-3C: binding studies on nucleosome

EMSA gels, ethidium bromide staining

Top left: a first 1:5 complex was formed, and final salt concentration was lowered to 400, 350 and 300 mM. A shift is clearly visible, whatever the salt concentration, as compared to nucleosome alone.

Top right: different ratios were tested at final salt concentration 250 mM, and showed clear shifts by EMSA. The final salt concentration was also lowered to 150 and 50 mM and did not lead to a loss of shift, indicating a stable complex formation at low salt.

Below: a 1:5 and 1:10 ratio was tested, both in Tris-HCl buffer. Salt concentrations ranging from 400 to 175 mM were tested and showed shifts as compared to nucleosome alone. Both ratio showed the same migration pattern by EMSA, suggesting that the 1:5 complex is already saturated. Furthermore, the lower the salt concentration, the higher the shift, indicating a more stable complex at 175 mM of NaCl.

precipitation, while up to 20 mg/mL could be reached using the second ones, with high recovery. Similar concentrations could be reached for nucleosomes.

A second optimisation had to be made regarding the mixing of both nucleosomes and MBD3. Considering that a final concentration of 3 to 5 mg/mL of complex, with final salt concentration around 150-200 mM, was ideally desired, no large dilutions were tolerated to avoid concentrating steps of the complex. To this end, several trials were achieved, to find the best conditions in which the MBD3-nucleosome complex would be stabilised at high concentration. As a reminder, MBD3 was stored in 500 mM of NaCl while nucleosomes were stored in a no-salt buffer. First attempts were carried out by slowly adding MBD3 on top of nucleosomes to the desired ratio. However, this invariably led to precipitation, due to saline shock. Overnight incubation helped however dissolving the precipitate, but EMSA gels showed no shift. Further attempts were made by mixing nucleosomes on top of MBD3, little by little. Several minutes would be needed to gently mix both samples and avoid as much as possible destabilisation of MBD3. That way, precipitation could mainly be avoided. Molar excess of MBD3 compared to nucleosomes however led to a too-high salt sample, around 300 to 400 mM. Further dilutions were thus required using a no-salt buffer to lower the final salt concentration. Typically, an overnight incubation would then be carried out prior to crystallisation.

A first MBD3-Nucleosome complex was used for crystallisation trials. This sample contained 80 mM of NaCl final, and EMSA was carried out to check for complex formation. Crystallisation drops were set up, using the Nucleix commercial screen (Qiagen), in CrystalQuick 96 Greiner plates, with 50 µL reservoir and 200 + 200 nL drops. These plates were specially designed for in-plate X-rays diffraction in the Diamond Light Source (Oxfordshire, England). Drops were set up at room temperature and stored at 17°C for 5 days, then moved to 20°C. Quickly, within a few days, crystals grew in a dozen of different conditions. Most of them were needles, microcrystals or spherulites. Some common features could be extracted from these conditions:

- first, the pH value of the crystallisation conditions always ranged 6 to 6.5 in sodium cacodylate or MES;
- second, an alcohol was found in more than half the cases, either isopropanol or 2-methylpentane-2,4-diol (or MPD). In the case of isopropanol, its volatility makes it difficult to handle in small volumes for crystallisation and reproducibility might not be optimal. Plus, opening the plate to harvest crystals would undeniably led to damages due to isopropanol evaporation;
- finally, spermine and spermidine were also often present. These polyamine mimic DNA via their negative charges and could thus stabilise MBD3;

Among all these crystallisation conditions, some needles could grow up and became rods of 15-20 µm large and thick, and several hundreds of µm long (*figure 82*). This was in particular the case in two conditions:

- A7: 20 mM of magnesium chloride, 15 % of isopropanol and 50 mM of MES pH 6.0

- E4: 18 mM of magnesium chloride, 9 % of isopropanol, 2.25 mM of spermine and 50 mM of sodium cacodylate pH 6.5

Also, two 50x50 µm crystals appeared in the condition F5 (9 mM of magnesium chloride, 0.9 mM of spermidine, 2.25 mM of spermine, 1.8 mM of cobalt hexamine and 5 % of PEG 400), surrounded by microcrystals.

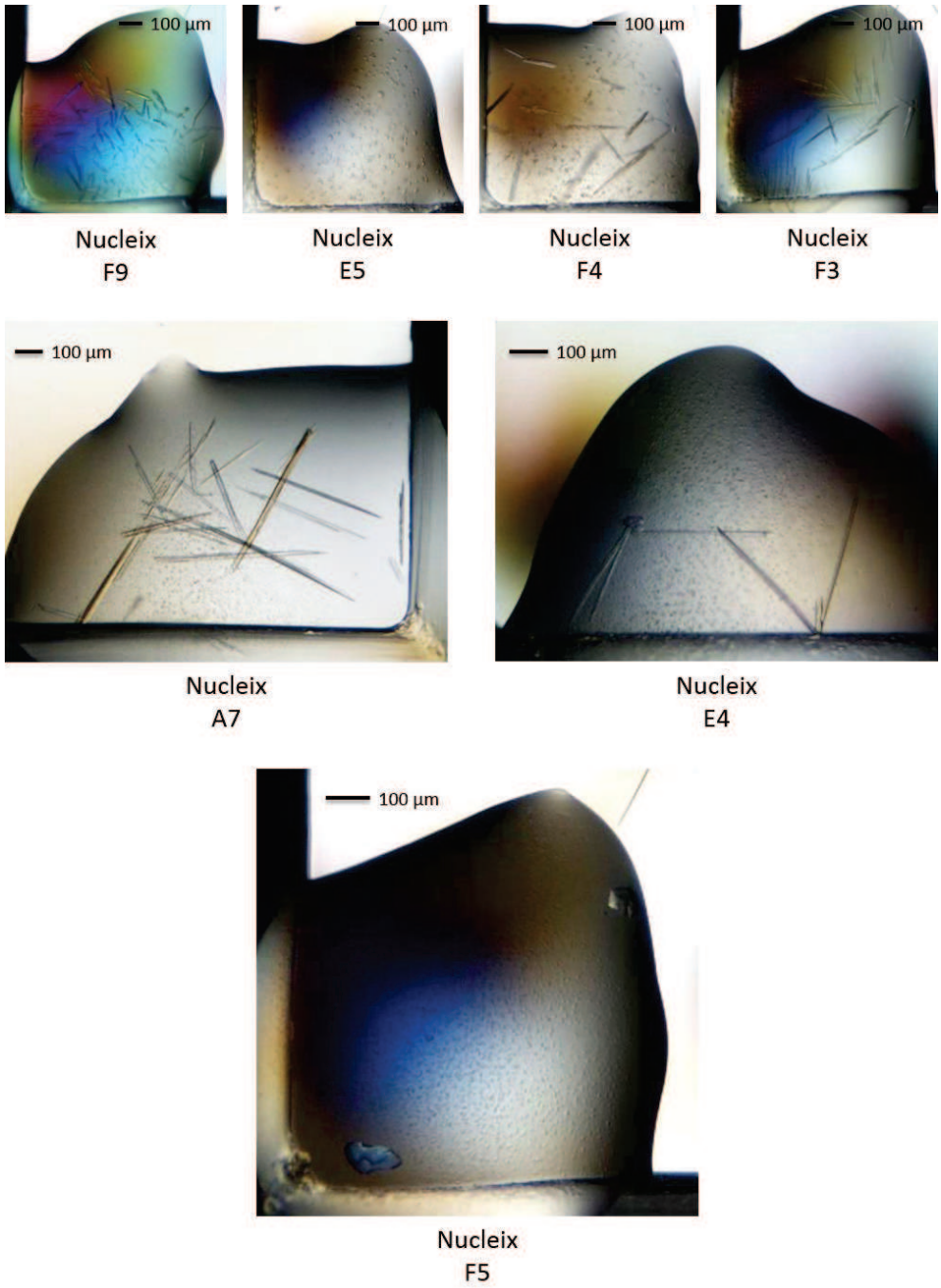


FIGURE 82

MBD3-Nucleosome complex: crystallization

Crystallization assays in CrystalQuick 96 Greiner plates, 200 + 200 nL drops, 50 µL reservoir, Nucleix screen. Complex buffer: 80 mM of NaCl, 50 mM of MES pH 6.5, 5 mM of EDTA, 5 mM of TCEP

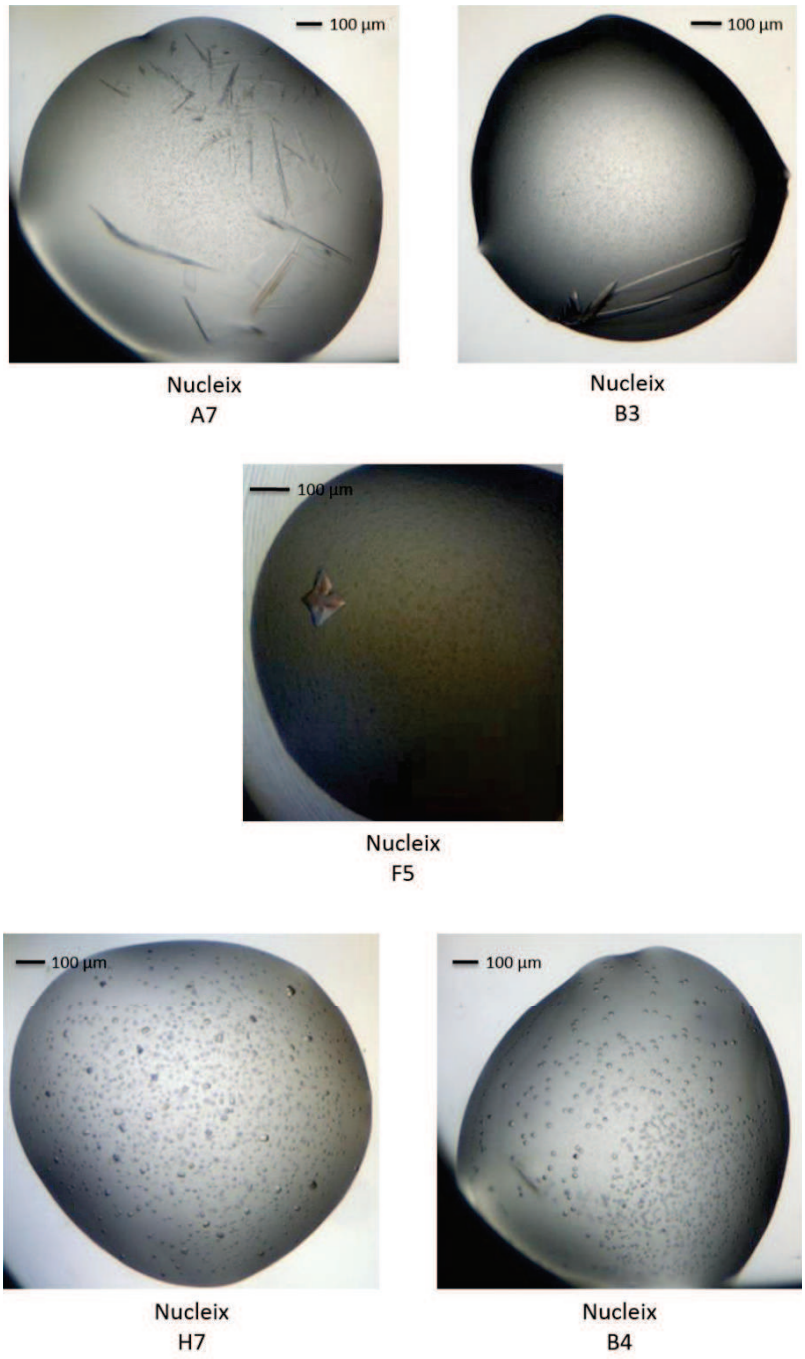


FIGURE 83

MBD3-Nucleosome complex: crystallization

Crystallization assays in MRC 2 plates, 200 + 200 nL drops, 50 µL reservoir, Nucleix screen.

Complex buffer: 180 mM of NaCl, 50 mM of MES pH 6.5, 5 mM of EDTA, 5 mM of TCEP, 1 mM of CHAPS

This first assay was followed by a second one, mainly similar with one difference in the final salt concentration: instead of 80 mM as previously, the final salt concentration was brought to 180 mM, which was more relevant with the prior binding assays that had been carried out. Again, an EMSA gel was run to check for proper complex formation. The same Nucleix screen was tested, although MRC 2 plates were used with 50 µL reservoir and 200 + 200 nL drops. Drops were set up at room temperature and stored at 17°C for 5 days, then moved to 20°C. The previous crystals could then be reproduced in the same conditions as well as new ones. In total, crystals grew in 16 different conditions. Again, most of them were needles and microcrystals (*figure 83*). However, rods and plates appeared in several conditions:

- A7: 20 mM of magnesium chloride, 15 % of isopropanol and 50 mM of MES pH 6.0
- B3: 40 mM of magnesium chloride, 5 % of MPD and 50 mM of sodium cacodylate pH 6.0

One 80x80 µm ninja star-shaped crystal could also be reproduced in the F5 condition, and microcrystals in two conditions grew up to around 10x10 µm:

- H7: 100 mM of sodium chloride, 0.5 mM of spermine, 25 % of MPD, 50 mM of sodium cacodylate pH 6.0 and 20 mM of magnesium acetate
- B4: 30 % of MPD, 50 mM of sodium cacodylate pH 6.0 and 40 mM of magnesium acetate

Finally, considering the issues that could be encountered with volatile alcohols as mentioned above, in-plate X-rays diffraction seemed the best way to handle this experiment. Crystals were thus reproduced in CrystalQuick 96 Greiner plates, with a 180 mM salt complex, and two different reservoir volumes: 50 µL and 70 µL. The latter would ensure a faster vapour diffusion and possibly quicker crystal growth. Drops were set up at room temperature and kept at 17°C for 5 days, then moved to 20°C. Again, crystals grew in a dozen of conditions, although interestingly, major differences could be observed between the two reservoir conditions. Thus, when using a 50 µL reservoir, rods up to 1 mm long x 50 µm large could be observed in three conditions (*figure 84*):

- A10: 5 % of PEG 4000, 50 mM of MES pH 6.0 and 5 mM of magnesium sulphate
- E3: 1.8 mM of cobalt chloride, 18 mM of magnesium chloride, 0.9 mM of spermine, 9 % of isopropanol and 50 mM of sodium cacodylate pH 6.5
- E6: 2mM of cobalt chloride, 10 mM of magnesium chloride, 10 % of isopropanol and 50 mM of di-sodium succinate pH 5.5

X-shaped crystals could also be observed in one condition:

- D11: 10 mM of magnesium chloride, 100 mM of potassium chloride, 30 % of PEG 400 and 50 mM of Tris-HCl pH 8.5

When using a 70 µL reservoir however, crystals grew in different conditions. Nice however small clover-shaped crystals were observed in four conditions (*figure 85*):

- B10: 15 % of PEG 400, 50 mM of sodium cacodylate pH 6.5 and 80 mM of magnesium acetate
- E4: 18 mM of magnesium chloride, 9 % of isopropanol, 2.25 mM of spermine and 50 mM of sodium cacodylate pH 6.5

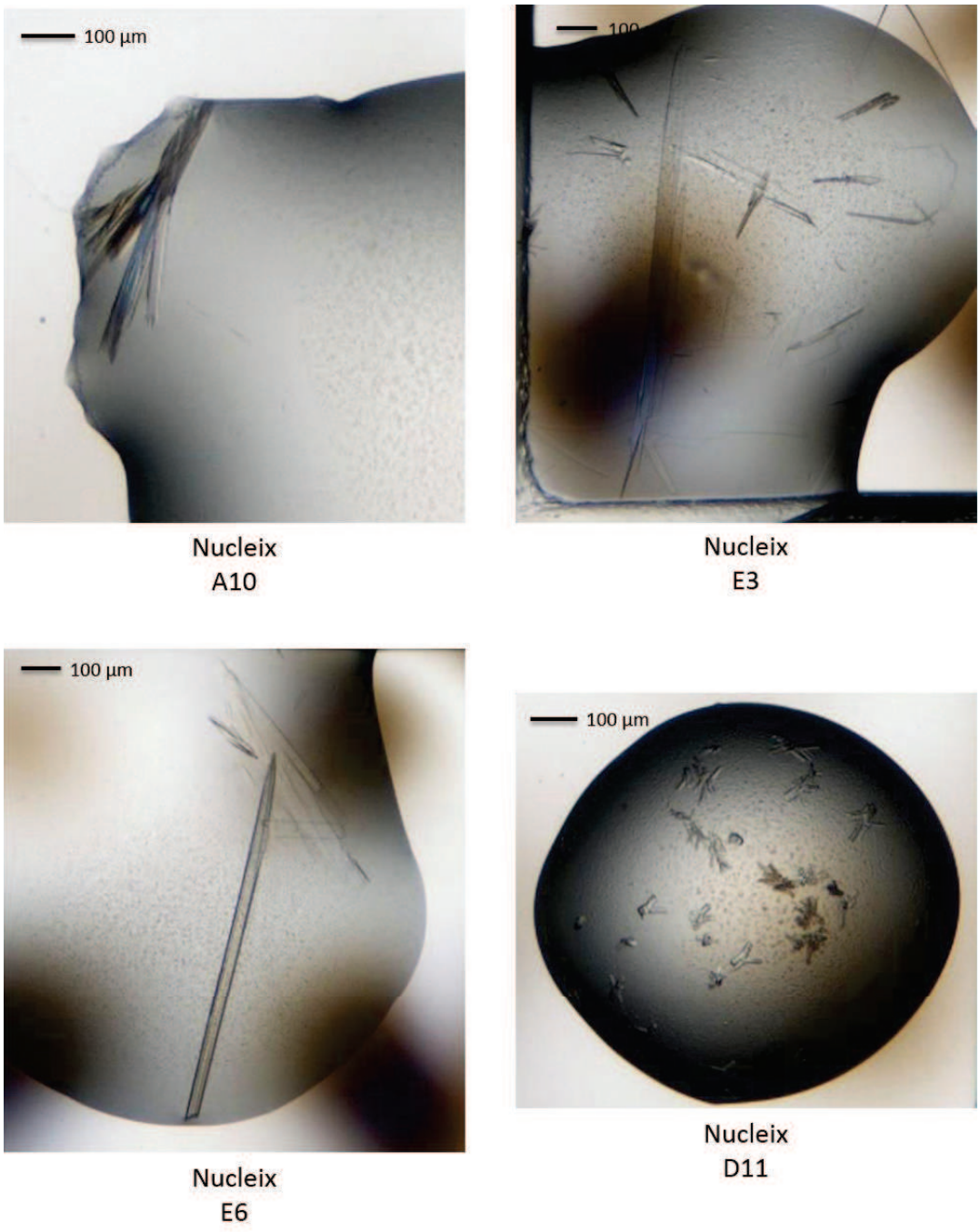


FIGURE 84

MBD3-Nucleosome complex: crystallization

Crystallization assays in CrystalQuick 96 Greiner plates, 200 + 200 nL drops, 50 µL reservoir, Nucleix screen. Complex buffer: 180 mM of NaCl, 50 mM of MES pH 6.5, 5 mM of EDTA, 5 mM of TCEP

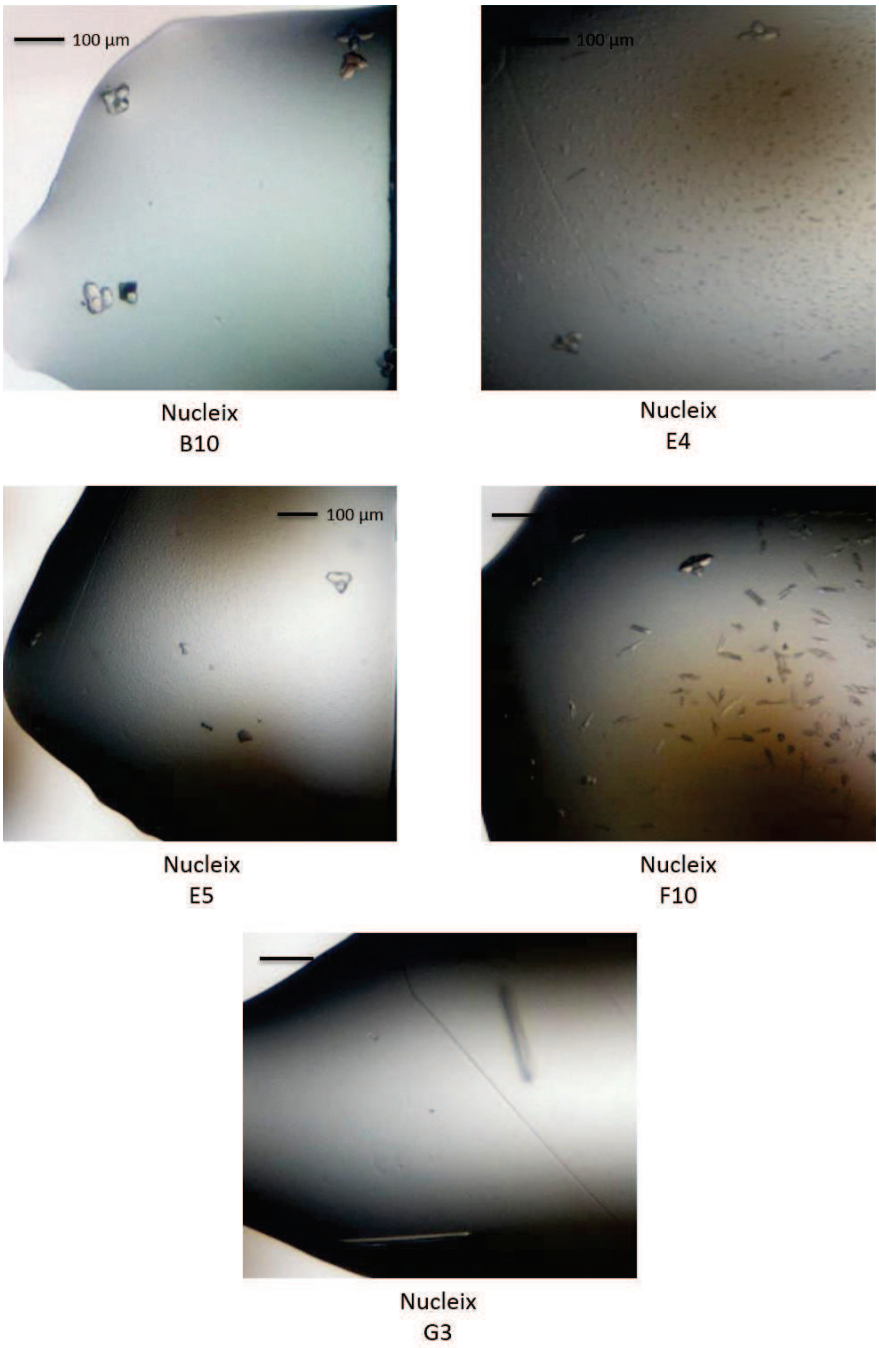


FIGURE 85

MBD3-Nucleosome complex: crystallization

Crystallization assays in CrystalQuick 96 Greiner plates, 200 + 200 nL drops, 70 µL reservoir, Nucleix screen.

Complex buffer: 80 mM of NaCl, 50 mM of MES pH 6.5, 5 mM of EDTA, 5 mM of TCEP

- E5: 0.9 mM of cobalt chloride, 18 mM of magnesium chloride, 2.25 mM of spermine, 4.5 % of MPD and 50 mM of sodium cacodylate pH 7.0
- F10: 18 mM of calcium chloride, 2.5 mM of spermine, 9% of isopropanol and 50 mM of sodium cacodylate pH 6.5

Also, two 250x20 µm rods grew in one condition:

- G3: 1 mM of cobalt chloride, 100 mM of magnesium chloride, 10 % of ethanol and 50 mM of sodium cacodylate pH 6.5

These crystals were all tested in-plate, at room temperature, in the Diamond Light Source on the 17th of December, 2013. Diffraction reached around 8-10 Å resolution for the best crystals, however, the rapid spread of free-radicals killed diffraction power within a couple of seconds. No dataset could thus be collected.

Based on crystallisation condition optimisation, new crystals could be obtained in 1 + 1 µL sitting drops. For this new experiment, drops were set up at 4°C, then immediately stored at 17°C. Within a couple of days, long 500x40 µm crystals could be observed. These crystals were treated with a cryoprotectant solution and mounted on loops. They were then tested on the PX II beamline in SLS on the 15th of February, 2014 (*figure 86*). Diffraction reached 8 Å and several incomplete datasets could be collected before diffraction decay. These allowed us to determine the space group and cell parameters of the crystal, being primitive monoclinic ($P2_1$) with a = 98.59 Å, b = 174.0 Å, c = 134.86 Å and β = 107.65°. Unfortunately, the low completeness of the dataset (52.5 %) did not allow us to further process this dataset, and molecular replacement could not be achieved.

	Overall	InnerShell	OuterShell
Low resolution limit	98.10	98.10	8.43
High resolution limit	8.00	25.30	8.00
Rmerge	0.072	0.033	0.589
Rmerge in top intensity bin	0.034	-	-
Rmeas (within I+/I-)	0.095	0.043	0.784
Rmeas (all I+ & I-)	0.095	0.043	0.784
Rpim (within I+/I-)	0.060	0.027	0.511
Rpim (all I+ & I-)	0.060	0.027	0.511
Fractional partial bias	-0.056	-0.036	-0.182
Total number of observations	6053	203	887
Total number unique	3506	107	518
Mean ((I)/sd(I))	4.5	12.3	1.2
Mn(I) half-set correlation CC(1/2)	0.996	0.993	0.195
Completeness	52.5	49.0	53.3
Multiplicity	1.7	1.9	1.7
Anomalous completeness	28.5	36.7	28.9
Anomalous multiplicity	1.2	1.1	1.1
DelAnom correlation between half-sets	-0.910	-0.944	-0.444
Mid-Slope of Anom Normal Probability	0.853	-	-

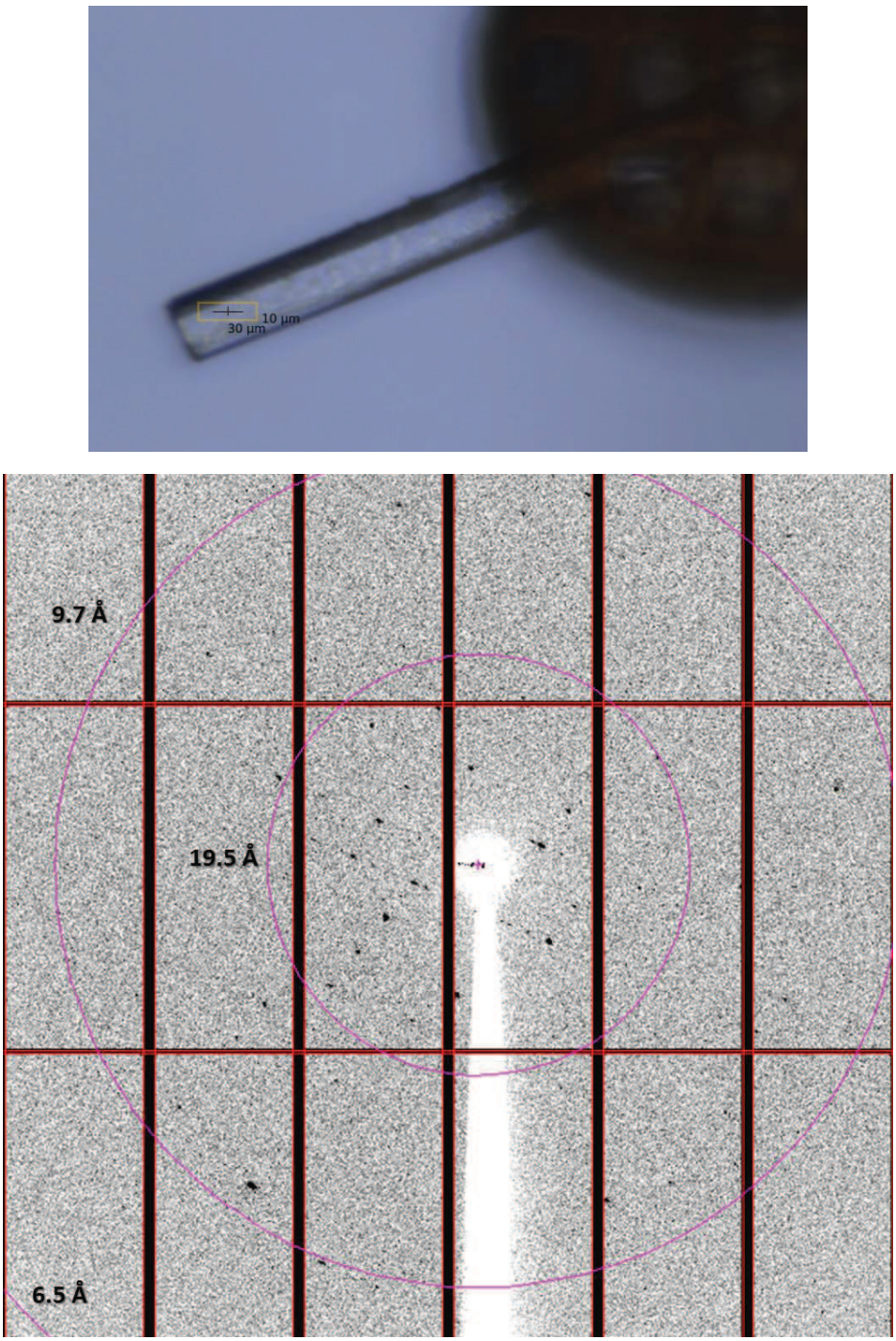


FIGURE 86

MBD3-Nucleosome complex: diffraction pattern

Top: MBD3-Nucleosome crystal mounted on a loop, in the X-ray beam. The yellow square represents the size of the beam (30x10 µm, PX II beamline, SLS, Villigen, Switzerland).

Below: diffraction pattern of the MBD3-Nucleosome crystal. Three resolution rings are shown in pink (19.5, 9.7 and 6.5 Å). Diffraction could reach up to 8 Å.

However, these preliminary results are encouraging as crystallisation conditions, crystal space group and cell parameters were different from the typical ones for core nucleosome crystals. Indeed, nucleosome core particles are routinely crystallised in 170 mM of $MnCl_2$, 120 mM of KCl and 40 mM of sodium-cacodylate pH 6.0, at room temperature, in a salting-in condition. They grow within 7 days to 3 weeks, and the crystals obtained are primitive orthorhombic ($P2_12_12_1$) with a = 106 Å, b = 182 Å and c = 110 Å. These major differences suggest thus that the crystals we obtained indeed contain an MBD3-nucleosome complex.

Finally, prior to pursue on crystallisation trials and optimisation, the stability of the MBD3-Nucleosome complex had to be improved. In spite of all the conditions tested, this complex would dissociate rapidly within a couple of days and lead to precipitation. One major question had thus to be addressed, that was whether MBD3 is monomeric or dimeric, and if indeed, it was correctly folded.

2.8 Is MBD3 a dimer or a monomer?

Stability of MBD3 was to some extent improved when in complex with nucleosomes. However, precipitation could still be observed after some days at 4°C, raising the issue of whether MBD3 was correctly folded or not. Plus, beside gel-filtration and one non-reproducible DLS measurement, no biophysical studies clearly stated whether MBD3 was a dimer or a monomer. Assumptions were only made based on a few experimental observations and from the literature. The latter however was also not based on biophysical studies.

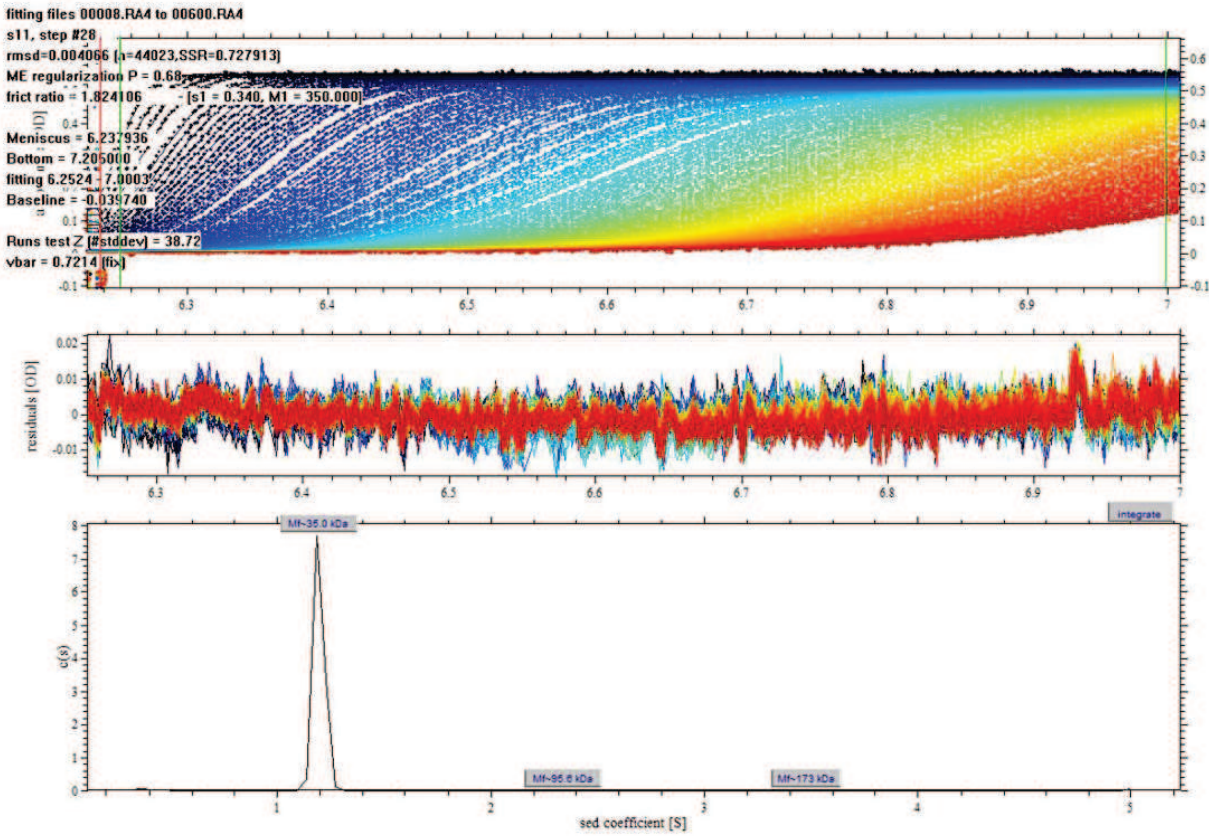
Thus, to address this question, I chose to carry out analytical ultracentrifugation on a freshly purified MBD3 sample. Sedimentation velocity analysis allowed the determination of several parameters of the sample (*figure 87*), such as the sedimentation constant s = 2.122 S. Then, based on $\overline{\nu}$ = 0.73 mL/g at 20°C and 0.7214 at 4°C, and considering the experimental friction ratio f/f0 = 1.824, the molecular weight of the particle could be approximated to 34.965 kilo Daltons. This result confirmed the absence of dimer in the sample, but was clearly in conflict with the gel-filtration profile.

Finally, using a theoretical friction coefficient corresponding to a compact sphere, the molecular weight of the particle could be approximated to 14.192 kilo Daltons. This clearly indicates a protein folding issue, all the more so considering the a/b ratio, that illustrates the length and width of the particle, was approximated to 12.98. Taken together, these results indicate that MBD3 is at least partially if not entirely unfolded.

This unfortunate result brought me to reconsider the whole strategy for structural study of this complex. Indeed, whether MBD3 was partially or fully unfolded when unbound to DNA, the MBD domain itself, at least, must fold during complex formation with the nucleosome, as clearly observed on EMSA-gels. However, this MBD domain of MBD3 only represents 25 % of the whole protein, and the 75 remaining percent must thus be responsible, at least in part, for the instability of

the protein in complex. We hypothesized that the unfolded part of MBD3, on its C-terminal end, could act as a rope puller, pulling the rest of the protein and unwinding it over time.

If this hypothesis proves to be true, then crystallography is not the best technique for structural studies. Indeed, crystal growth is a slow process that can take up from several days to weeks. It thus requires that the complex remains stable during this period, bringing us to the conclusion that unless very fast growth occurred, any crystals obtained would most probably contain only nucleosome.



sw = 1.195 S, sw(20,w) = 2.122 S
(c = 0.5241, 97.805 % of the total)
with best-fit friction ratio = 1.824: Mw = 34965 Da
(based on vbar = 0.721, vbar20 = 0.730, hydratation = 0.300, buffer density = 1.02615, relative viscosity = 1.7006)
minimum Mw (for compact sphere) = 14192 Da
Stokes Radius (20C) = 3.95 nm, a/b(oblate)=12.98, a/b(prolate)=11.47

FIGURE 87

MBD3 N-His-3C: analytical ultracentrifugation

Sedimentation velocity analysis was carried out at 4°C on a fresh MBD3 protein. An unexpectedly high f/f0 friction ratio of 1.824 was determine, suggesting a partial unfolding of the protein. This is confirmed by the a/b ratio determined, corresponding to the major over minor axis ratio if the protein was an ellipse.

From this point, I decided thus to follow three different routes: use in situ proteolysis to crystallise a folded domain of MBD3; study the MBD3-nucleosome complex by cryo-EM; and focus on the isolated MBD domain of MBD3. These three different parts of the project are further described in the upcoming chapters.

2.9 Mild-proteolysis assays

A proteolytic fragment or domain of a protein might crystallise more easily than a protein exhibiting unfolded structures. In the case of MBD3, this may be true since over 5000 different conditions were unsuccessfully tested to grow MBD3 crystals. To overcome this issue, I thus decided to make use of the Prote-Ace kit (Hampton) which provides a set of six different proteases, to generate small and hopefully folded fragments or domains of MBD3, for crystallisation.

This experiment was carried in two steps. First, a proteolytic screening was achieved on small aliquots of freshly purified MBD3. All six proteases (α-chymotrypsin, trypsin, elastase, papain, subtilisin and endoproteinase Glu-C) were tested, in a 1/1000-fold ratio, for overnight proteolysis at 4°C. Analysis of the products by SDS-PAGE showed a clean and complete cleavage of MBD3 with trypsin and subtilisin, leading to around 18 and 20 kilo Dalton products, respectively (*figure 88*).

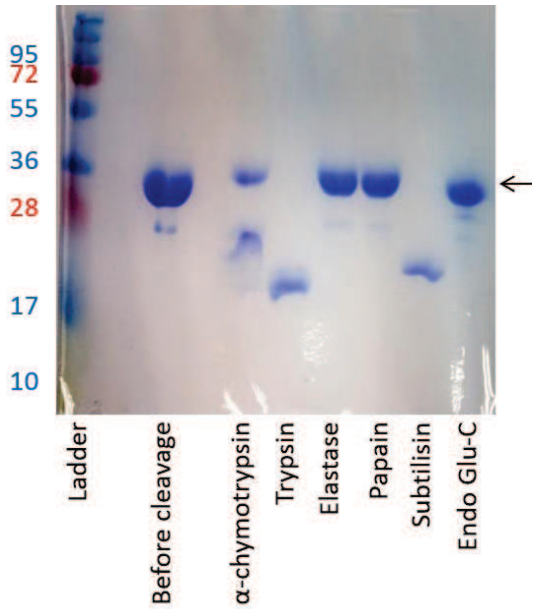


FIGURE 88

MBD3 N-His-3C: mild proteolysis

SDS-PAGE, Coomassie staining

Mild proteolysis was carried out on MBD3 full-length using the Proti-Ace kit (Hampton Research). Six different proteases were tested and two showed good result, with a clean cleavage leading to only one band: trypsin and subtilisin. This method is often used to cleave out unfolded parts of proteins and help crystallization of folded domains alone.

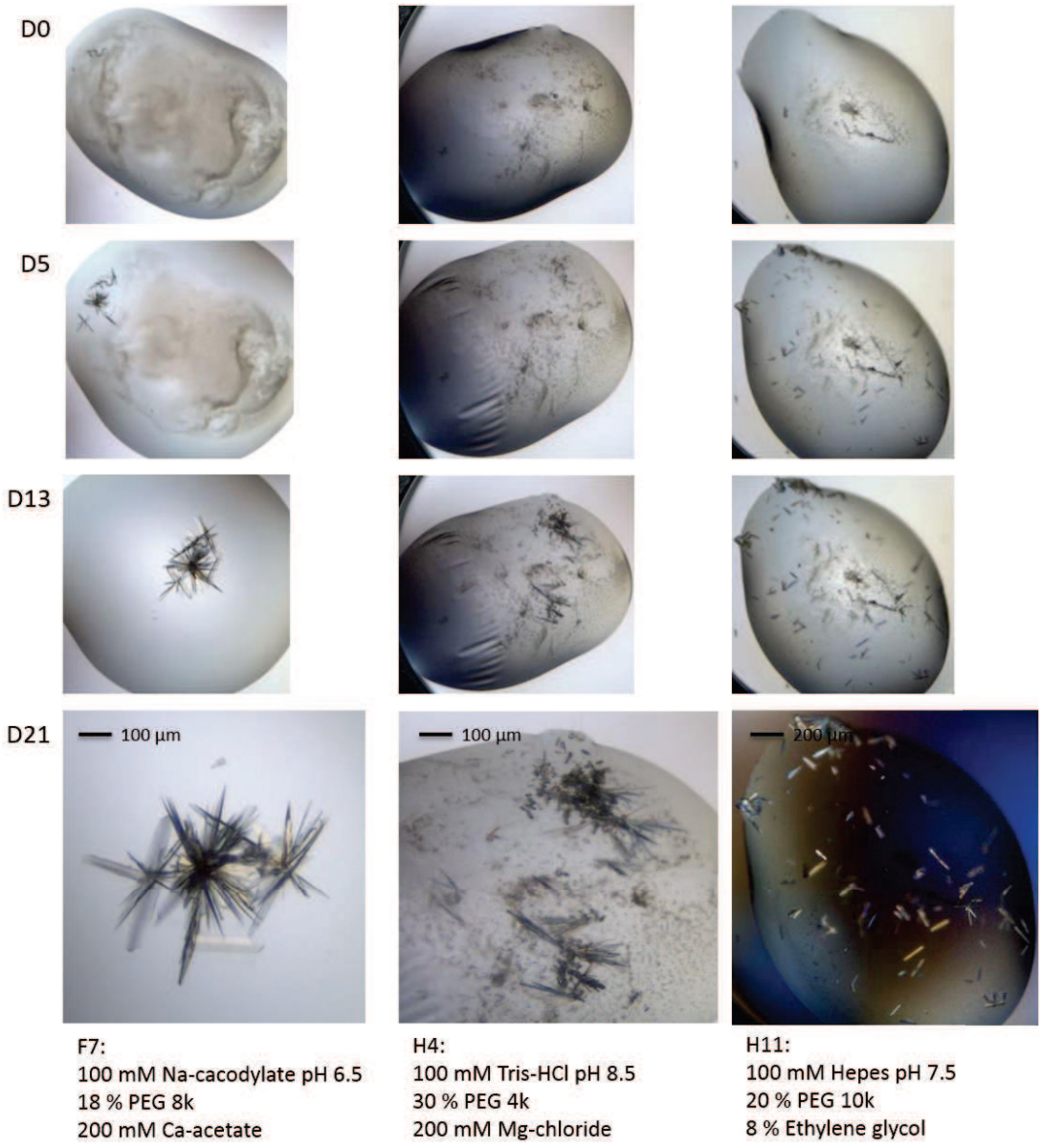


FIGURE 89

MBD3 N-His-3C: crystallization

Crystallization assays in MRC 2 plates, 200 + 200 nL drops, 50 µL reservoir, The Classics screen.
MBD3 buffer: 500 mM of NaCl, 50 mM of MES pH 6.5, 5 mM of EDTA, 0.5 mM of TCEP, 1 mM of CHAPS
Using *in-situ* mild proteolysis with subtilisin (1/1000), crystals could be obtained in three different conditions.

These two proteases were then further used to carry out *in situ* mild proteolysis of MBD3. Again, a 1/1000-fold dilution of each protease was used, and directly added in the crystallisation drops. Three different commercial screens were tested with a 6.0 mg/mL sample of MBD3: AmSO4 and The Classics (Qiagen); and Wizards I+II (Emerald BioSystems). Drops were sat up at 4°C using the Mosquito robot and stored at 4°C in the RockImager imaging system.

Three different conditions gave crystals in the Classics screen, in the presence of subtilisin (*figure 89*). They all started from a full precipitate which dissolved little by little within a few days, and allowed crystals to grow. These conditions contained:

- 100 mM of sodium cacodylate pH 6.5, 18 % of PEG 8000 and 200 mM of calcium acetate (condition F7);
- 100 mM of Hepes pH 7.5, 20 % of PEG 10000 and 8 % of ethylene glycol (condition H11);
- 100 mM of Tris-HCl pH 8.5, 30 % of PEG 4000 and 200 mM of $MgCl_2$ (condition H4);

Out of the three, the condition "F7" gave the nicest crystals. They were thus chosen for diffraction tests. During crystal fishing, I equilibrated the drop with a solution containing a 0.5X mix of both protein and reservoir concentrations (250 mM of NaCl, 25 mM of MES pH 6.5, 2.5 mM of EDTA, 50 mM of sodium cacodylate pH 6.5, 100 mM of calcium acetate and 18 % of PEG 8000). The cryoprotectant buffer was the same, supplemented with either 15 % of glycerol or 15 % of ethylene glycol. In both case, the crystals were plunged into a drop of cryoprotectant and behaved well during the treatment. They were finally flash-frozen in liquid nitrogen, and tested on the PX II beamline (SLS, Villigen, Switzerland) on the 29th of July, 2014. Unfortunately, diffraction patterns were characteristic of salt crystals, most probably acetate crystals.

The two other conditions were not tested since crystals were too small to be properly fished and mounted on loops. Plus, this experiment was replicated twice but led to a different proteolysis pattern, and crystals could not be reproduced in any of these three conditions.

2.10 MBD3-Nucleosome complex studies by cryo-EM

Cryo-EM gives the advantage of fixing the sample in a given hydrated state before studying it. In the case of MBD3-nucleosome, this technique turned out to be ideal, since the complex could be frozen immediately after formation, while still stable.

A first trial was carried out using a 6.4 mg/mL sample of 4:1 MBD3-nucleosome complex. This sample was prepared, incubated overnight, then diluted to the desired concentration and frozen. It contained 40 mM of MES pH 6.5, 200 mM of NaCl, 5 mM of EDTA and 500 µM of TCEP. Glow-discharged and no-discharged grids were used to freeze 2.5 µL of 0.3, 0.6 and 0.8 mg/mL sample, blotted for 0.5 second with filter paper to remove excess fluid, and then plunged in liquid ethane. These grids were finally checked using the Tecnai F30 Polara microscope. Among the different conditions tested, glow-discharged grids with 0.6 mg/mL of sample seemed to give the best result in terms of particle shape and distribution. Micrographs were thus recorded at 100 kV on the Falcon 1 camera, at 1.8 Å/pixel.

Out of this first data set, 12000 particles could be picked using EMAN2. These particles were split into two groups, treated separately, and nicely converged towards the same final 3-D reconstruction. A first 25 Å-map could thus be solved, showing a flat-shaped disc, and topped with an extra electron density (*figure 90*). Despite this low resolution, the disc showed a central groove as

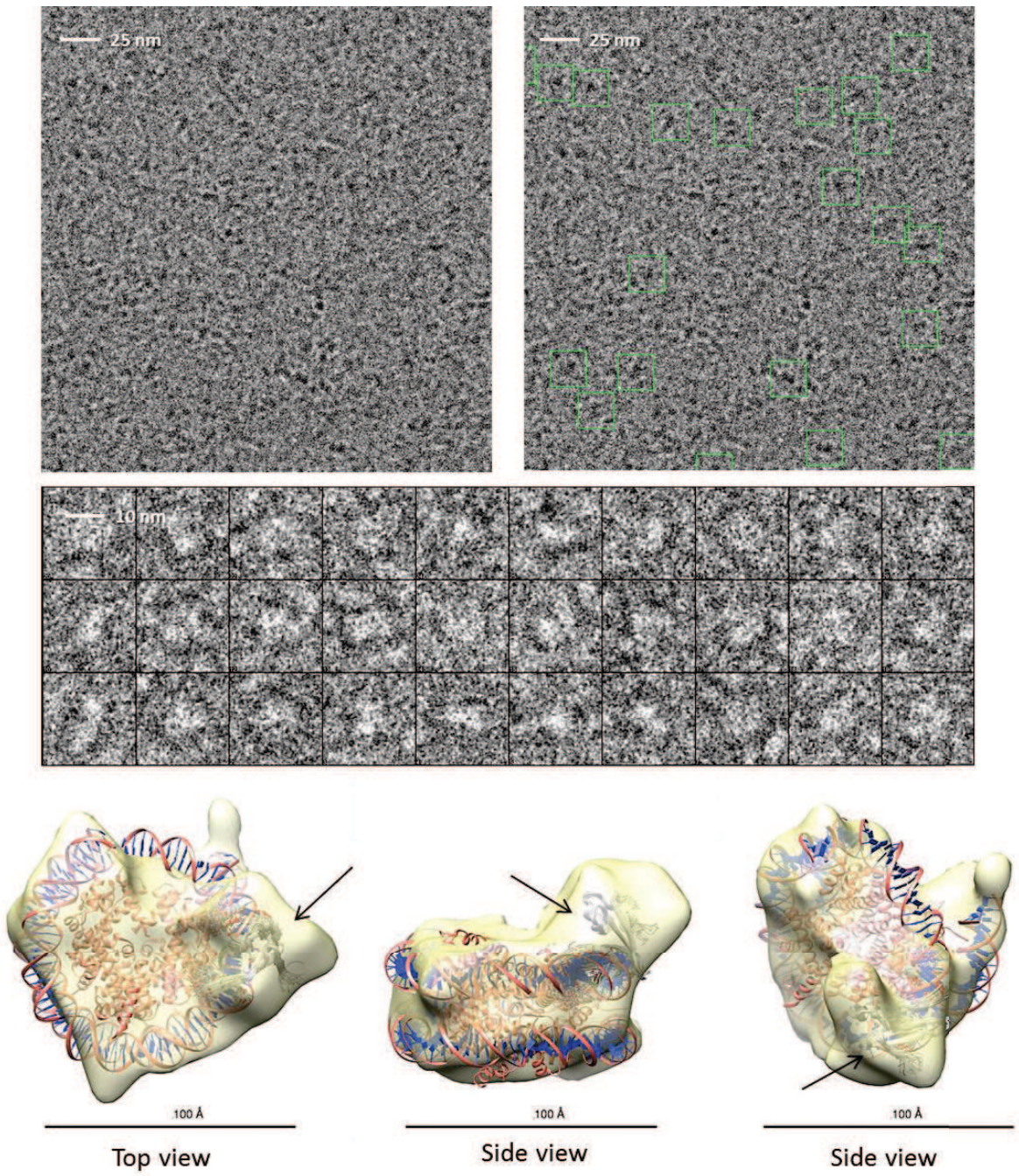


FIGURE 90

MBD3-Nucleosome: cryo-EM

MBD3-Nucleosome 4:1 complex, 0.6 mg/mL

Complex buffer: 180 mM of NaCl, 50 mM of MES pH 6.5, 5 mM of EDTA, 0.5 mM of TCEP

Data recorded on Tecnai F30 Polara, at 100 kV, Falcon 1 camera, 1.8 Å/pixel. A band-pass filtered electron micrograph is shown (top left) and the same one after particle picking (top right).

Although contrast is not optimal, 12000 particles could be picked and used to reconstruct a first density map at 25 Å resolution (EMAN2). Crystal structures fitted in the electron density: nucleosome (pdb: 3lz0) and $MBD2^{MBD}$ (pdb: 2ky8). The extra electron density highlighted by black arrows corresponds to the remaining part of MBD3 (C-terminal part after the MBD domain, residues 75 to 291).

well as DNA entry and exit points which allowed to nicely fit a nucleosome crystal structure (pdb: 3lz0) into this density. The extra-density on top of it corresponds to a 35-kDa molecule, which is consistent with MBD3, and the solution structure of $MBD2^{MBD}$ bound to DNA (pdb: 2ky8) could be placed into this density, showing a single molecule bound per nucleosome. According to the sequence of the Widom DNA, it fells nicely on a CpG island. Plus, the remaining extra electron density, corresponding to the rest of the protein that could not be fitted (the MBD domain represents only 25 % of the protein), seems to stretch out and cover to a certain extent the top of the nucleosome, interacting with histones H3 and/or H4. This suggestion makes sense since the C-terminal part of MBD3 contains a large amount of acidic residues as already mentioned. These residues could thus interact with basic histones and anchor the protein on nucleosomes rather than on free DNA. However, this is only a supposition, and whether this is a representation of reality or only an artefact still needs to be addressed.

Nevertheless, this first reconstruction had to be reproduced and optimised. Indeed, high amounts of MES buffer (40 mM), together with an excess of MBD3 (4:1), led to contrast issues that needed to be solved in order to get a better reconstruction. Further cryo-EM experiments were thus carried out to optimise the sample and collection conditions. A 2:1 MBD3-nucleosome complex at 0.6 mg/mL was prepared and frozen on grids. However, only long threads could be observed on the micrographs, characteristic of nucleosome unwinding. Another parameter that could be exploited to increase contrast was the buffer concentration. We thus decided to set a 6:1 MBD3-nucleosome complex, showing a nice shift on EMSA-gel, in only 20 mM of MES pH 6.5. This sample was frozen at 0.6 and 0.8 mg/mL on glow-discharged grids. In parallel, highly concentrated nucleosomes particles were diluted to 0.6 mg/mL in the same complex buffer (180 mM of NaCl, 20 mM of MES pH 6.5, 5 mM of EDTA and 500 µM of TCEP) and frozen on similar grids. These grids were checked using the new Titan Krios microscope, and interestingly, they revealed a complete destabilisation of nucleosome particles in MES buffer, which was not observed when in complex with MBD3. Plus, the contrast appeared to be better than previously, and a 2 TB data set could be collected using movie-mode data recording, on the Falcon 2 camera. This mode allows several full-resolution frames per second to be written at low electron dose, instead of one frame at higher electron dose, causing sample damage and information loss.

Another data set was collected using freshly reconstituted nucleosomes together with freshly purified MBD3, in similar conditions (glow-discharged grid, 6:1 ratio, same buffer composition, 0.3 mg/mL). Both of these datasets are currently being treated and will hopefully lead to a medium-to-high resolution structure of a full nucleosome-bound MBD3 protein.

3. Focusing on the MBD domain…

3.1 Choosing the right domain boundaries

Working with the isolated domain of MBD3 first required cloning of this domain into a pET28b expression vector with a 6xHis-tag and a 3C cleavage site in N-terminal. This work was performed by the molecular biology service of the IGBMC. In total, seven new vectors were designed, with three different domain boundaries:

- The first isolated domain corresponds to the MBD *stricto sensu*. This domain goes from residue 2 to 72.
- A second domain goes from residue 2 to 102. It corresponds to a fragment already observed in MALDI-TOF experiments and described earlier.
- The third domain goes from residue 2 to 133. It was defined by online prediction tools.

These three fragments were thus cloned as wild-types, but also including the F34Y mutation for each of them, as well as the R23M mutation for $MBD3_{2\text{-}72}$.

Expression tests were carried out for all seven for these vectors and all showed high levels of overexpression. However, due to time limitation, I started to work with the $MBD3_{2\text{-}72}$ construct only. The other constructs will be the subject of future work carried out in the lab.

3.2 Purification

Considering the already optimised protocol to purify MBD3 full-length, I decided to start from the same basis. However, I decided to reduce the salt concentration throughout the whole purification process, following the example of $MBD2^{MBD}$ purification that has recently been published[534]. The protocol was thus the following: 20 litres of cell culture were treated using a lysis buffer (300 mM of NaCl, 50 mM of MES pH 6.5, 4 mM of CHAPS, 2 mM of BME, PMSF and cOmplete EDTA-free). After complete resuspension, the sample was lysed using lysozyme and sonication. The lysate was clarified by centrifugation and the supernatant was incubated with Ni-NTA resin. After one hour, the resin was washed with a wash buffer (300 mM of NaCl, 50 mM of MES pH 6.5, 10 mM of imidazole, 4 mM of CHAPS and 2 mM of BME), then poured into a gravitational chromatography column. After further washing, the protein was eluted using an elution buffer (300 mM of NaCl, 50 mM of MES pH 6.5, 300 mM of imidazole, 2 mM of CHAPS and 2 mM of BME). The eluate was very cloudy and had to be centrifuged before loading onto a HiTrap Heparin 5 mL column. A stepwise-elution (350 mM, 400 mM, 450 mM, 500 mM, 550 mM, 600 mM, 650 mM and 700 mM of NaCl) was carried out using a low salt and high salt buffer (0 mM/1M of NaCl, 50 mM of MES pH 6.5, 5 mM of EDTA, 1 mM of CHAPS and 2 mM of BME). MBD3 could be recovered in several fraction going from 600 to 700 mM of NaCl. After concentration of the pooled fractions, the sample was injected in two pre-equilibrated Superdex 200 10/300 GL columns connected in series (GF buffer: 150 mM of NaCl, 50 mM of MES pH 6.5, 5 mM of EDTA, 1 mM of CHAPS and 500 µM of TCEP). The GF profile showed only one specie, eluted as an 8-11 kilo Dalton protein ($MBD3_{2\text{-}72}$ is 10.3 kilo Daltons including the

tags). Furthermore, quantification using Bradford reagent indicated over 35 mg of pure protein at the end of the purification process. This protein could thus be stored in small aliquots in 30 % of glycerol at -80°C.

It could then be used for further biophysical and structural studies, after a gel-filtration step to remove glycerol and possible aggregates due to freezing.

3.3 Biophysical studies

$MBD3_{2-72}$ was used to carry out binding assays on reconstituted nucleosomes. Unfortunately, after several trials, using different ratios of protein to nucleosome, and lowering the salt concentration, as tried previously for the MBD3 full-length study, no binding could be observed by EMSA. This first observation could corroborate the idea that MBD3 requires its C-ter domain to properly bind to nucleosomes.

To further delve into this idea, I decided then to carry out binding studies on small DNA oligos. To this end, I designed ten 11-bp oligos, based on the work of K. Günther, who identified MBD3-binding regions by ChIP-seq analysis, including several gene promoters such as ZSCAN22, MED23, AURKAIP1, RPL37 and CRABP1. I extracted CpG islands from these gene promoters and designed oligos. Different CpG-adjacent nucleotides were also selected to study their effect on binding efficiency. Scarsdale *et al.*[534] indeed observed a sequence preference for a guanine residue immediately following the CpG island in the case of MBD2, and showed that in several MBD2-binding promoters, the amount of CpGpG exceeded 60 %, while CpGpT represented only around 15 % of CpG sequences. The table below summarizes the ten DNA oligos designed.

Oligo name	Gene promoter	Sequence (5'-3')
ZSCAN22/1	*zscan22* (ZNF50, HKR2)	CAGGC**CG**GAGC
ZSCAN22/2		GAGGA**CG**AAGT
MED23/1	*med23* (CRSP3)	GACAC**CG**TCTC
MED23/2		GGTGT**CG**GCTG
AURKAIP1/1	*aurkaip1* (Aurora Kinase A-Interacting Protein)	GGTCT**CG**AACT
AURKAIP1/2		ACCTG**CG**GGTA
RPL37/1	*rpl37* (Ribosomal Protein L37)	ATCCT**CG**TCCT
RPL37/2		AGCCC**CG**CATT
CRABP1/1	*crabp1* (RBP5)	CCTCA**CG**CTTT
CRABP1/2		CCTTG**CG**AGCT

After annealing of these HPLC-purified oligos, these were used for preliminary ITC experiments. DNA was stored in an ITC buffer, also used for MBD3 gel-filtration (150 mM of NaCl, 50 mM of MES pH 6.5, 5 mM of EDTA and 500 µM of TCEP), to avoid buffering effects during ITC

injections. Five experiments were carried out. For each of them, MBD3 was placed in the sample cell while the DNA was gradually injected into the cell with the syringe. Measurements were carried out at 4°C, 10°C and 25°C, using 20 µM of protein and 200 µM of DNA (RPL37/1). However, apart from dilution effects, no titration could be observed, revealing no binding of MBD3 on DNA.

This result was confirmed by binding experiments carried out on the same protein sample by Christophe Papin (Ali Hamiche's lab, IGBMC), using radio-labelled DNA containing unmodified and modified CpGs (mCpG, hmCpG, fCpG and caCpG).

I carried out further optimisations to study $MBD3_{2-72}$ binding on DNA oligos. In particular, I took inspiration from the binding and structural studies of other MBD domains on DNA (including MBD2[534] and MBD4[533]). Similar to my experiments, these were carried at salt concentration ranging between 50 and 150 mM of NaCl. However, the major difference was the constant usage of Hepes-NaOH buffer at pH 7.4. Using MES pH 6.5, some electrostatic interactions could then be unstable and could result in low binding affinity.

Taking this into account, I decided to gel-filtre an aliquot of $MBD3_{2-72}$ using the two Superdex 200 10/300 GL columns connected in series, with the following GF buffer: 250 mM of NaCl, 20 mM of Hepes-NaOH pH 7.4 and 1 mM of $MgCl_2$. The gel-filtration chromatogram shows a typical profile, and this sample was used to carry out new binding assays using the MED23/2 DNA. A 1.5:1 ratio of protein to DNA was used, to allow saturation of the DNA. Thus, 1 µL of 1.92 mg/mL DNA (266 µM) was mixed with 6.88 µL of 0.6 mg/mL $MBD3_{2-72}$ (399 µM). The DNA oligo being stored in the previously described ITC buffer with 150 mM of NaCl, the final salt concentration of this mix was around 237 mM. With addition of a low-salt buffer, several salt concentrations could be tested, ranging from 237 mM to 50 mM. After a 2-hour incubation time at 4°C, a 20 % EMSA-gel was loaded and run for 5 hours at 2 W in 1x TBE buffer. Finally, ethidium bromide staining allowed the observation of clear shifts on the gel (*figure 91*). These shifts show several successive levels, according to the final salt concentration: thus, a first level is observed at 175 and 150 mM of NaCl, while a higher shift is seen at lower salt concentration. However, at 50 mM of NaCl, the shift heights starts to be lost and become similar to the 175/150 mM level. This successful result has led me to pursue efforts on these binding assays, and optimisation of the ratio and salt concentration, as well as new ITC studies prior to crystallisation tests, are now currently ongoing.

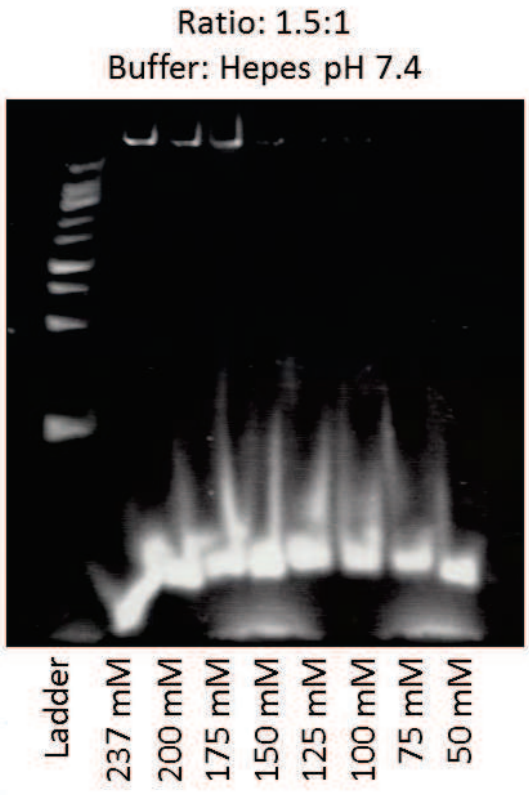


FIGURE 91

$MBD3_{2-72}$: binding studies on DNA oligos

EMSA gel, ethidium bromide staining

An 11-bp DNA oligo was used for binding studies with the MBD domain of MBD3. Both were mixed to a final salt concentration of 237 mM, then salt concentration was gradually lowered until 50 mM. The gel shows shifts in successive levels, along with the salt decrease. The higher shifts are observed at 125 and 100 mM of NaCl. Below, the complex seems to destabilize, especially at 50 mM of NaCl.

RESULTS – PART III

STUDY OF THE PROTEINS RBAP46/48

1. Context

The two proteins RbAp46 and RbAp48 share 90 % of sequence identity and both adopt the same tertiary structure. However, they have been shown to be implicated in different biochemical activities: RbAp46 associates with HAT1 to acetylate newly synthesised H4 histones, while RbAp48 is a subunit of CAF1, involved in loading histones onto newly replicated DNA. Both bind to helix H1 of histone H4, and the X-ray structure of RbAp46 in complex with a small peptide of histone H4[15] led the authors to conclude that this binding would destabilise the H3-H4 handshake motif. As a consequence, RbAp46 and RbAp48 would be able to bind to free histone H4 only.

As fully-fledged members of the NuRD complex, which role is to remodel chromatin through binding to nucleosomes, the above-mentioned roles of RbAp46 and RbAp48 appear to be inconsistent within the frame of NuRD. It is thus legitimate to question the conclusion reached by the authors. This is how I got interested in studying this interaction more thoroughly during my PhD. My goal was to define, using structural biology, whether these two histone chaperones were indeed unable to bind to nucleosomes or if the latter could exhibit some hitherto unknown structural plasticity.

Expression vectors were designed for baculovirus system expression and optimal cultivation conditions were determined as previously described. I could thus develop a purification protocol, to isolate RbAp46 and RbAp48 in a soluble and non-aggregated form. This work is detailed in the following chapter.

2. Purification

The following protocol has been optimised based on previous work done by Lejon *et al.*[412], and applies to both RbAp46 and RbAp48 (*figure 92*). Hi5 cells from 3 L of culture were harvested 48 hours post-infection by pelleting at 1200 Xg for 20 minutes at 4°C. These were then resuspended in a lysis buffer containing 1 M of NaCl, 20 mM of Tris-HCl pH 7.5, 5 % of glycerol, 4 mM of CHAPS, 10 mM of imidazole and 5 mM of BME, and supplemented with cOmplete EDTA-free tabs for protease inhibition. Lysis was performed by Dounce homogenization and sonication for 10 minutes at 60 % amplitude, alternating 1 second of sonication and 1 second of break. Cell debris were then pelleted at 100000 Xg for twice 45 minutes, at 4°C. The clarified lysate was transferred to Ni-NTA resin and rotated for 2 hours at 4°C to facilitate protein binding. The pelleted beads were then poured into a polypropylene gravity flow column and washed. The protein could finally be recovered using the same buffer containing 500 mM of imidazole. The sample was thus immediately buffer-exchanged by overnight continuous flow dialysis against dialysis buffer containing 50 mM of NaCl, 20 mM of Tris-HCl pH 7.5 and 5 mM of BME.

Following dialysis, the sample was applied onto a HiTrapQ 5 mL anion exchange column (GE Healthcare). A linear gradient was applied using a low salt and high salt buffer (0 mM/1M of NaCl, 20

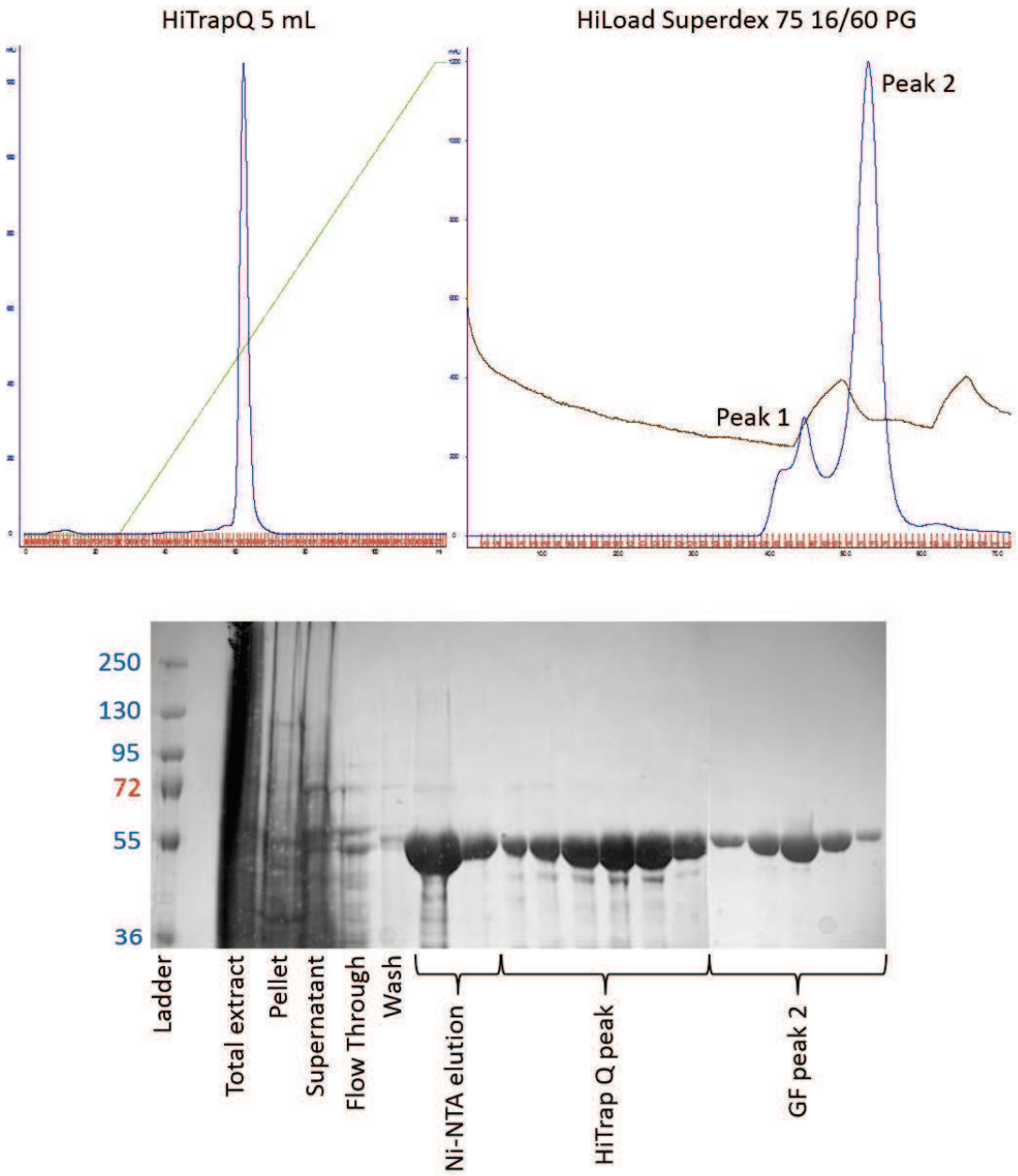


FIGURE 92

RbAp46/RbAp48: purification process

Top left: ion-exchange chromatography profile (HiTrapQ 5 mL)

Top right: gel-filtration profile (HiLoad superdex 75 16/60 PG)

Below: SDS-PAGE, Coomassie staining

Both RbAp46 and RbAp48 could be purified using the same protocol. After a nickel-batch affinity step, an ion exchange could be achieved, leading to pure protein. A final gel-filtration step allows to get rid of potential aggregates (peak 1) and recover pure and monomeric protein (peak 2).

mM of Tris-HCl pH 7.5 and 5 mM of BME) and both RbAp46 and RbAp48 were eluted at 270-300 mM of NaCl. Fractions containing the protein were pooled and dialyzed against 100 mM of NaCl, 20 mM of Tris-HCl pH 7.5 and 5 mM of BME before concentration in an Amicon Ultra 4 mL 30-kilo Daltons concentrator. The concentrated sample was finally injected onto a Superdex 75 16/60 PrepGrade column and eluted with 100 mM of NaCl, 20 mM of Tris-HCl pH 7.5 and 5 mM of BME, to remove potential aggregates. Further salt reduction has been tested and both RbAp46 and 48 showed to be stable in 20 mM of Tris-HCl pH 7.5 and 5 mM of BME.

On average, 0.5 mg/L of culture of pure and monomeric RbAp46 could be recovered, and 1.5 mg/L of culture for RbAp48. These highly pure proteins could be further characterised by biophysical methods as described hereafter.

3. Biophysical characterisation

Biophysical characterisation was intended to define the oligomeric state and polydispersity of RbAp46 and 48, as well as their molecular weight.

Gel-filtration offered a first estimation of the protein size: both were eluted at Ve ≈ 53 mL on a 120 mL column (*figure 92*), corresponding to a molecular weight of 60 kilo Daltons. This value is higher by about 18 % than the theoretical molecular weight of these two proteins, which remains acceptable.

MALDI-TOF analysis was also carried out and confirmed the presence of RbAp46 and RbAp48 (*figure 93*). However, the C-terminal part of each protein could not be covered, leading to uncertainty regarding the integrity of the protein. Finally, two minor contaminants in the RbAp48 sample could be identified as endogenous proteins from *Autographa californica nuclear polyhedrosis virus*.

Finally, further characterisation was carried out using DLS. RbAp46 was dialysed against 20 mM of Tris-HCl pH 7.5 and 5 mM of BME after gel filtration, and appeared to be monomeric, with an approximated molecular weight between 55 and 60 kilo Daltons (*figure 94*). Plus, it showed an excellent polydispersity index between 11.5 and 12 %, making it thus ideal for structural studies.

4. Binding assays and structural studies

Once these proteins were purified and biophysically characterised, I could move forward and use them for binding assays together with reconstituted nucleosome particles. This work was achieved with the help of Dr. Kareem Mohideen.

Binding studies were carried at room temperature since both RbAp proteins and nucleosome appeared to be stable at 20°C and analysed by EMSA. A 2:1 and 4:1 ratio of RbAp over nucleosomes was used, and both were mixed in a no-salt condition (20 mM of Tris-HCl pH 7.5 and 5 mM of BME). For this first study, Widom DNA-nucleosomes were used, but unfortunately, EMSA gel

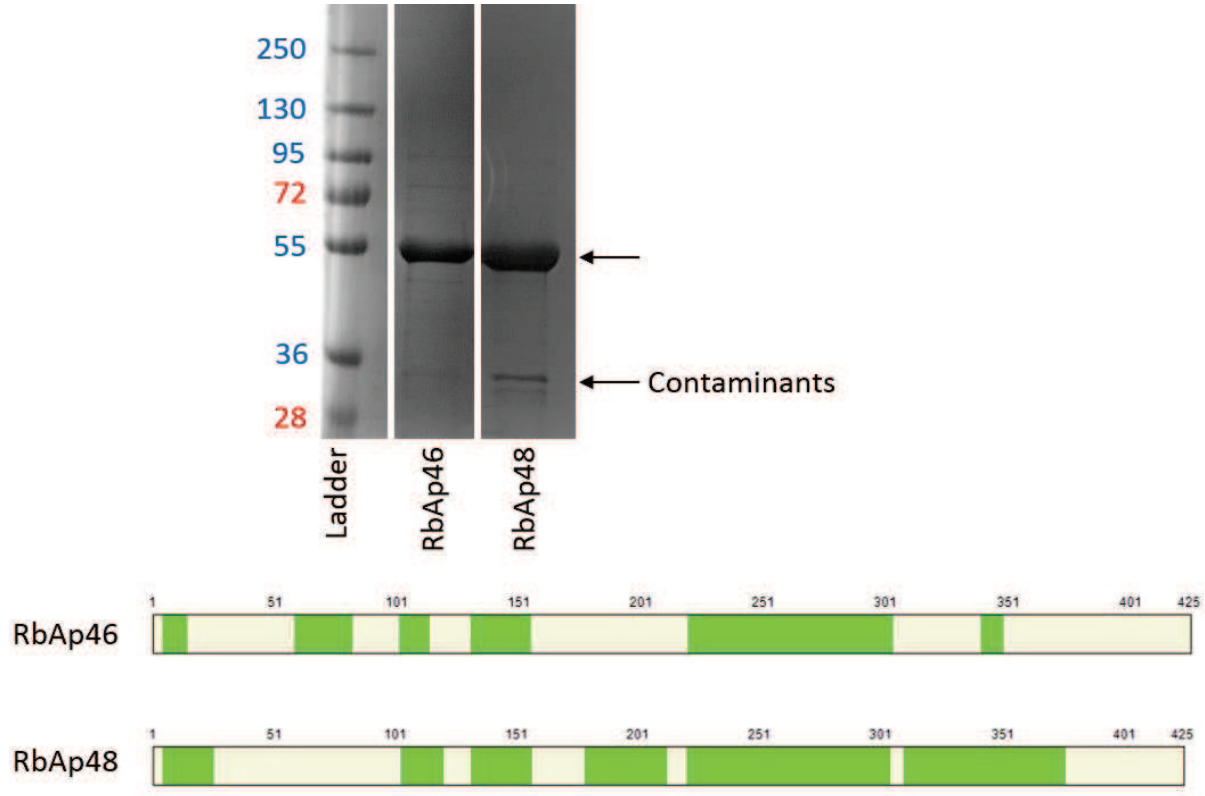


FIGURE 93

RbAp46/RbAp48: MALDI-TOF

Top: SDS-PAGE, Coomassie staining

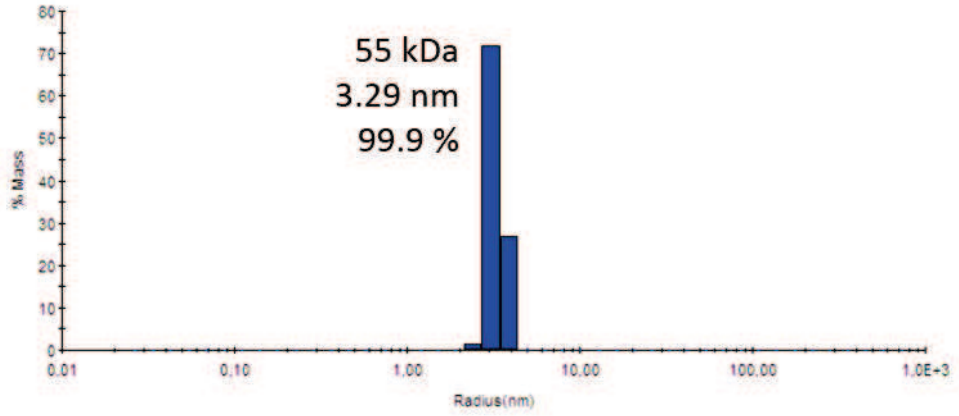
Below: peptide coverage map after MALDI-TOF analysis

The tryptic peptide analysis of both RbAp46 and RbAp48 shows integrity of the N-terminal end but does not cover the C-terminal end. Furthermore, two contaminants co-purified with RbAp48 appeared to be endogenous proteins from the baculovirus.

FIGURE 94

RbAp46: dynamic light scattering

DLS was carried out at room temperature on a fresh RbAp46 sample, and showed excellent polydispersity (around 11.5-12 %). The molecular weight of RbAp46 could be estimated to 55-60 kilo Daltons, with a minor part of aggregates (0.1% in mass, 15.7 % in intensity).



didn't show any binding. For the record, this DNA is an artificial sequence optimised to tightly bind to histone octamer.

Regarding the previous statements made by Murzina *et al*. on the binding of RbAp proteins to histone H4, we thus decided to test another DNA sequence with lower affinity towards the histone octamer, to allow a certain measure of freedom and rearrangement upon binding. The natural α-satellite "α32" DNA was thus used to reconstitute new nucleosomal particles and carry out further binding studies. Using the same protocol for complex formation, the EMSA gel showed, surprisingly, binding of RbAp46 and RbAp48 to nucleosomes (*figure 95*). Plus, the 4:1 complex showed a higher shift on the gel, in particular in the case of RbAp46. This exciting result tends to support and corroborate our first speculation on the role of the RbAp proteins within the NuRD complex.

To confirm this idea, crystallisation assays have been carried out, in order to get a structural evidence of the RbAp-Nucleosome complex. A 3:1 ratio was used to form new RbAp46-Nucleosome complexes around 3 to 4 mg/mL, and, after a 30-minute incubation, 100 + 100 nL crystallisation drops were set up. Over 1500 different salting-out conditions were tested, using commercial screens, in MRC 2 plates. In parallel, well-known salting-in nucleosome crystallisation conditions were also tested on this sample, in 1 + 1 µL hanging drops.

After a few days, crystals appeared in the latter conditions, among which a hexagonal bipyramid and twisted rods (*figure 95*). In both cases, these shapes have never been observed in nucleosome crystals. Their diffraction was tested in the SLS, on the PX III beamline, and showed medium resolution spots, to 4.8 Å. Data collection was thus achieved and processed by molecular replacement. Unfortunately, no structure could be solved due to high mosaicity, bad electron map and missing features.

Others crystals were also obtained in a dozen of various commercial screen conditions, although no common features could be extracted from these conditions (*figure 95*). One condition gave particularly good looking crystals, in the Nucleix screen:

- A3: 20 % of MPD, 100 mM of magnesium acetate and 50 mM of MES pH 5.6

These crystals were tested in-plate on the PX III beamline in the SLS, and showed a diffraction pattern reaching 4 Å resolution. However, only a couple of images could be collected prior to crystal damages and loss of diffraction power, and no complete dataset could be collected.

Further experiments are thus required in order to optimise crystallisation of this complex. In particular, the very promising crystals above-mentioned will have to be reproduced and cryoprotected prior to X-ray diffraction, with the hope of collecting a full dataset.

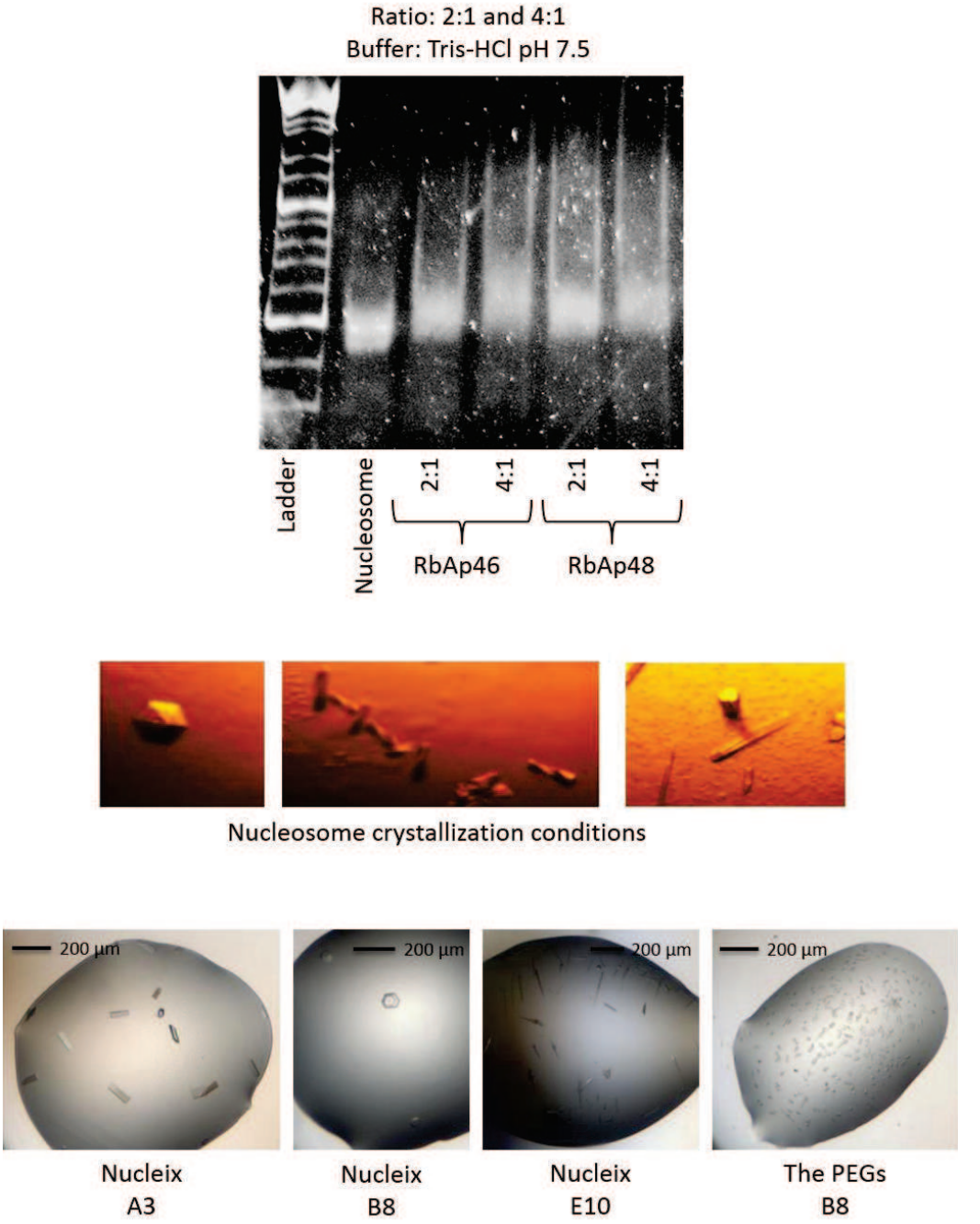


FIGURE 95

RbAp46: binding to nucleosomes and crystallization

Top: EMSA gel, ethidium bromide staining

Binding assays were carried out with 2:1 and 4:1 ratios of protein over nucleosome. Both RbAp46 and RbAp48 were tested in a no-salt buffer, and showed clear binding.

Crystallization assays were carried out on the RbAp46-Nucleosome complex and crystals grew in several conditions.

DISCUSSION AND OUTLOOK

During these 4 years of PhD, I've had the great opportunity to set up a new project focused on chromatin regulation, which appeared to be original and innovative in a team which mainly worked on nuclear receptors and translational processes.

Chromatin remodelling is a vast and still poorly understood field, yet an important, if not major, process for cellular function. In our body, every cell possesses the same genetic material, and yet, there is not much in common between a skin cell, a neuron, a myocyte or a lymphocyte. Each of these acquired very different functions, however from the same initial instruction manual, by a strict time-space selection of genes to be expressed or repressed. Chromatin remodelling is involved in this evolutionary progress, which allowed single-celled organisms to evolve towards multicellular organisms, exhibiting increased complexity due to cell compartmentalisation and differentiation.

This remodelling process can be carried out in distinct ways: for example, in an ATP-dependent way, using the energy released by ATP hydrolysis to destabilise the structure of chromatin; or by covalent modifications of the chromatin components (histones and/or DNA), which can cause a direct change in the chromatin packaging (in the case of acetylation/deacetylation for example), or through the recruitment of enzymes, transcription factors and other partners to further regulate transcription.

The NuRD complex is one out of many complexes involved in this remodelling process. It has been discovered in 1998 and has, since then, proven to be a very unique complex, carrying two *a priori* opposite remodelling activities: an ATP-dependent chromatin remodelling activity and a deacetylase activity. Furthermore, it has been shown to be the main form of histone deacetylase in the cell, as well as a very broad and general transcriptional repressor; however recent studies seem to highlight a much more complex and multifaceted role. Whatever it be, this complex hasn't lift the veil to date on all its specificities, and tremendous work still needs to be achieved, especially from a structural point of view.

It was within this contest that, in collaboration with the group of Ali Hamiche, we decided in 2010 to start this challenging project. I joined the team of Bruno Klaholz at that time as a PhD student to implement this project with two main ideas in mind. First, the study of each isolated subunits of the complex, as well as stable subcomplexes within NuRD or with chromatin components like nucleosomes. Indeed, only few structural data were available at that time and most of the NuRD subunits had not been studied. A great work was, and still is, to be done regarding this issue. Second, the study of the whole endogenous complex purified from human cells. Here again, despite the 16 years that have passed since the discovery of the NuRD complex, fundamental issues have still not been solved. In particular, the question of the stoichiometry of each subunit within the NuRD complex is still open, not to mention its overall structure which is completely unknown.

* * * *

My first work was thus to clone each NuRD subunit gene into expression vectors in order to produce recombinant proteins. With the aim in mind to express several genes to produce stable

subcomplexes, we made our choice for the baculovirus expression system. Regarding the good overexpression of MBD3, RbAp46 and RbAp48, I selected these three proteins as main topic of study. Of particular interest, these three subunits have been shown to directly interact with chromatin components, either DNA, histones or full nucleosomes which are reconstituted routinely in our laboratory.

The case of MBD3 is especially noteworthy, in the sense that this protein is a perfect example of recent evolution. It belongs to the MBD family, binding methylated CpG islands, and shares 77 % of sequence identity with its paralogue MBD2. However, in mammals, appearance of a point mutation in MBD3 (Y34F) has led to a loss of specificity towards methylated DNA. This mutation is not observed in lower vertebrates, suggesting a redundant role of MBD2 and MBD3 in these organisms; and in invertebrates, only one protein named MBD2/3 exists. From a pathological point of view, another point mutation (R23M) has been observed in two patients, displaying late and unfunctional language acquisition. I thus started to study the human MBD3 protein with the aim of understanding the molecular mechanism of MBD3 binding to DNA. Apart from the wild-type human MBD3, the study of the two previously described point mutations within the MBD domain (the counter-evolutive F34Y and the pathological R23M) was designed to help understanding the binding specificity of this protein towards unmodified DNA and its implication in neurological diseases.

The absence of biochemical data on MBD3 made its purification a tedious work. Although overexpression was nicely achieved in insect cells, limited solubility as well as precipitation and aggregation once purified were common issues. Several months of trials were thus necessary to obtain sufficient amounts of pure and soluble protein to carry out structural studies. But in spite of the 3500 crystallisation conditions tested, none gave protein crystals. Finally, a mass spectrometry analysis of the sample showed that the protein produced was a shortened isoform of MBD3, called MBD3Δ, which lacks the MBD domain. I therefore decided to undertake new clonings of the main isoform of MBD3.

* * * *

MBD3 cDNA being unavailable from libraries, a synthetic gene was designed and subcloned into expression vectors. Again, a baculovirus vector was built in view of co-infections to produce stable subcomplexes. In parallel, bacterial vectors were designed to express rapidly and at lower cost huge amounts of protein for isolated studies. Again, insolubility and aggregation caused delays in structural studies implementation. However, Thermofluor® contributed significantly to solve aggregation issues and what could be characterised as an MBD3 dimer could be purified and isolated.

Attempts to characterise this dimer gave ambiguous and uncertain results. Though literature mentions dimerisation of MBD3, no reliable data were published. Therefore, I made use of several biophysical methods to solve this issue. First, gel-filtration profiles of MBD3 constantly showed elution of the protein in a volume corresponding to a 70-80 kilo Daltons species, i.e., a dimer of MBD3. DLS experiments, in turn, showed a monodisperse species with an estimated molecular

weight between 70 and 80 kilo Daltons. Finally, analytical ultracentrifugation tore these previous results to shred by showing a single monomeric and partially unfolded species. We therefore concluded that MBD3 probably requires partners in order to get stabilised and to fold properly, explaining in particular the very high instability of this protein. Working with isolated and unstabilised protein required indeed organisation of strict working conditions and installation of a working bench at 0-2°C.

In the meantime however, binding and structural studies have been carried out using nucleosomes. In spite of MBD3 instability on its own, binding conditions could be defined and revealed by EMSA. A refolding procedure had to be settled down to allow MBD3 to interact with nucleosomes and to properly fold. This could be achieved by slow lowering of the salt concentration by addition of a low-salt buffer. Although this worked nicely for biophysical studies like EMSA, which require only low concentrations of this complex, this process had to be further optimised to achieve crystallography studies. By juggling with highly concentrated samples, slow addition of low-salt buffer and incubation times, we finally managed to obtain MBD3-Nucleosome complexes at suitable concentrations for crystallography. Over a dozen of conditions gave crystals, mostly needles or plates, but also some 3-D crystals, morphologically suitable for X-rays diffraction tests. In plate-tests at room temperature were carried out at the Diamond Light Source (Oxfordshire, England), and showed diffraction for some of these crystals, however weak and rapidly lost upon radiation. Whether these crystals contain the MBD3-Nucleosome complex or only nucleosome is currently unknown. Indeed, this complex has shown to be unstable over some days, which is incompatible with proper crystal growth. However, crystals could be obtained in one optimised condition, and after cryoprotection treatment, data collection at the SLS showed diffraction up to 7 Å. A first dataset could be collected, and allowed to solve the space group and cell parameters, which appeared to be different from that of the core nucleosome. But the completeness level of the dataset was too low (52.5 %) to achieve molecular replacement and solve a 3-D structure. Further experiments and optimisations will thus be needed.

To overstep the reasons set out earlier, regarding the instability of the MBD3-Nucleosome complex over time, we decided to make use of our expertise in single particle cryo-EM. The MBD3-Nucleosome complex could be frozen directly after complex formation, making sure to preserve the interactions monitored by EMSA. A first data collection led to a preliminary low-resolution 3-D reconstruction, around 25 Å, nicely showing a circular and flat shape corresponding to the nucleosome, and topped with an extra-density in which the NMR solution structure of the MBD domain of the MBD2 could be fitted. This extra-density shows a clear interaction on the side of the nucleosome, expectedly with DNA, but more surprisingly, seems also to spread on the face of the nucleosome to reach the H3-H4 dimer. Plus, it is interesting to mention that the Widom DNA used on these nucleosomes contains 13 CpG, yet only one extra-density is found on the nucleosome. It appears that this extra-density is located on a CpG, raising thus the question as to why would MBD3 have a specific binding site on the nucleosome. One assumption could be that MBD3 not only interacts with the DNA but also with histones as suggested earlier. However, details of this

interaction need to be confirmed. To this end, new data were collected on our newly installed Titan Krios microscope, after optimisation of the sample concentration as well as buffer condition. Strikingly, micrographs revealed a stabilisation of nucleosomes in presence of MBD3, while in absence, nucleosomes tended to fall apart in similar buffer conditions. Again, a medium-to-high resolution structure of this complex will enable us to address this question and understand the mode of interaction between MBD3 and the nucleosome. New data collections are currently being processed to this end and should help describing the MBD3 binding site on the nucleosome.

* * * *

Considering the unfolded state of MBD3, I decided in parallel to work on the isolated MBD domain of MBD3. After cloning and expression, I could optimise a purification protocol, leading to very high yields of pure protein, however still unfolded in absence of DNA. Based on the same implementation carried out for MBD3 full-length, I could study binding properties of this MBD domain towards nucleosome and DNA oligos. Interestingly, the isolated MBD domain did not show any affinity for nucleosomes as compared to the full-length protein. On the contrary, binding assays on small DNA oligos turned out to be positive. Salt concentration, buffer type and pH need to be optimised to characterise MBD3 in interaction with unmodified and modified DNA oligos (methylated, hydroxymethylated, formylated and carboxylated).

* * * *

To sum up my studies on MBD3, I should stress the very challenging work it has been. Yet, hypothesis and suggestions have been made, based on observations and a fine understanding of the behaviour of this protein. With the first exciting results coming up now, these questions will finally be addressed. In particular, we observed that the full-length MBD3 binds to nucleosomes while surprisingly, the MBD domain didn't show any affinity for nucleosomes. Considering that the MBD domain of MBD3 represents only 25 % of the whole protein and the function of the remaining 75 % remains unknown. First cryo-EM reconstruction suggests an interaction between the C-terminal end of MBD3 and histones H3 and/or H4. This could explain the lack of affinity of the MBD domain alone towards nucleosomes. As part of the chromatin remodeller NuRD, MBD3 has to be recruited either directly or indirectly to chromatin components, and in particular to the nucleosome. The role of MBD3 in nucleosome recognition has thus to be seriously considered, as a mechanisms to distinguish between CpG islands whether they are found in nucleosomal DNA or not. Currently processed data are expected to tell us more about this interaction interface.

Besides, the promising study of the MBD domain of MBD3 in complex with DNA oligos designed from gene promoter sequences will hopefully lead to high resolution structure of this complex and lift the veil on the molecular specificities ruling unmodified DNA recognition. Furthermore, the future study of the two MBD3 mutants F34Y and R23M should give detailed insights into the specificity of the interactions. In particular, the R23 residue has not been shown to be a crucial residue for CpG island recognition but still is located close by. The hypothesis of an

abolished affinity towards CpG islands is however very unlikely, regarding the embryonic lethality observed in knockout mouse. It is thus possible that this R23M mutation imparts new functional properties to MBD3 without affecting its primary role.

* * * *

My work has also been focused on RbAp46 and RbAp48. After expression vector design, I have been able to produce and purify these two histone chaperones from insect cells. An easy-to-implement and common purification protocol has been set up and first binding assays with nucleosomes have been carried out. The great stability of both RbAp46 and RbAp48 allowed to completely remove salt and EMSA gels showed binding of both proteins on reconstituted nucleosomes with the natural α-satellite DNA. This unexpected result was inconsistent with previously published data which suggested, although without experimental evidences, that these chaperones could only bind H3-H4 dimers at best, but were unable to bind to tetramers, octamers or full nucleosomes. As we could show interactions with the nucleosome, crystallisation trials were carried out and a dozen of conditions gave crystals among which some were suitable for X-rays diffraction. However, only medium resolution data could be obtained, which did not allow 3-D structure determination. These results are however very promising, first of all because obtaining crystals of factor-bound nucleosomes is very difficult, and because these crystals have been obtained in previously unseen conditions for the crystallisation of the core nucleosome. Further binding studies using biophysical tools and new crystallisation trials are currently ongoing, to try to address experimentally the binding of these chaperones to full nucleosomes.

* * * *

To conclude, this ambitious project consisted in studying the NuRD complex from a structural point of view, in order to gain insights into its functions. My work has helped implementing this long-term project, which is now on track, with respect to MBD3, RbAp46 and RbAp48 which should reveal new insights to help drawing a scheme of NuRD functions. To go a step further, stabilisation of the full-length MBD3 will require to characterise interaction partners that are likely to bind to the protein and induce its proper folding. However, crystallisation of MBD3-Nucleosome and RbAp46-Nucleosome complexes shows the feasibility of structural analysis of nucleosome with NuRD subunits.

In parallel, new studies should be carried out with other subunits of NuRD, in particular CHD4, the core ATPase if the NuRD complex, which remains a very challenging protein. This 218-kilo Dalton protein has been well studied from a functional point of view, but the structural aspects have been overlooked. Thus, to date, only NMR structures of the two PHD domains and the chromodomain of this protein have been published.

The MTA2 protein is also a key protein, especially considering its involvement in cancer. It has indeed be shown to make breast cancer cells insensitive to oestrogens and tamoxifen. Recently, the publication of two crystal structures of MTA1 has brought MTA2 back to centre stage. These

structures correspond to the ELM2 and SANT domains of MTA1 (at the N-terminal end) bound to HDAC1; and a small peptide of the C-terminal end of MTA1 bound to RbAp48. Taken together, these isolated structures suggest the existence of a stable subcomplex of NuRD, including HDAC1, RbAp48 and MTA2. Furthermore, the BAH domain of MTA2 has been shown to interact with histone H3 and could thus be complexed onto nucleosomes. All these subunits are now available for baculovirus expression, and coinfections could lead to *in vivo* subcomplex reconstitution.

Finally, the most exciting part of this project lies in the study of the whole NuRD complex. To this end, the team of Ali Hamiche has designed a HeLa cell line, stably expressing a tagged MTA2 subunit. By TAP-tag purification, the endogenous NuRD complex can thus be purified and further studied. In particular, crosslinking protein interaction analysis by mass spectrometry could reveal the still poorly known interaction network between the different NuRD subunits; and cryo-EM can be used to study the structure of this 1-mega Dalton complex. In this respect, first trials have already been carried out, and the whole NuRD complex could be purified, at the choice, either with or without nucleosomes, as revealed by SDS-PAGE. Moreover, recent studies indicate that the NuRD complex is highly stable and requires above 1 M of salt to dissociate. But despite that, first observations in cryo-EM showed essentially dissociated complex subunits. Sample preparation conditions will thus have to be optimised in the near future to carry on this promising study.

PUBLICATIONS AND ORAL COMMUNICATIONS

Publications

J-F. Ménétret, H. Khatter, A. Simonetti, I. Orlov, A. G. Myasnikov, Vidhya KV, S. Manicka, **M. Torchy**, K. Mohideen, A-S. Humm, I. Hazemann, A. Urzhumtsev, B. P. Klaholz. Integrative structure-function analysis of large nucleoprotein complexes. RNA structure and folding, de Gruyter, **2013**, D. Klostermeier & C. Hammann (Eds.); invited review/book chapter.

Morgan P. Torchy, Ali Hamiche, Bruno P. Klaholz. Structure and function insights into the chromatin-remodelling complex NuRD. *Submitted*.

Conference contributions

Morgan P. Torchy, Kareem Mohideen, Sinthuja Peiris, Isabelle Hazemann, Arnaud Depaux, Ali Hamiche, Bruno P. Klaholz. Structure-function analysis of the chromatin remodelling complex NuRD. Poster. 11th EMBL Conference: Transcription and Chromatin. Heidelberg, Germany. August 2014.

11th EMBL Conference: Transcription and Chromatin. Heidelberg, Germany. August 2014.

Structure-function analysis of the chromatin remodeling complex NuRD

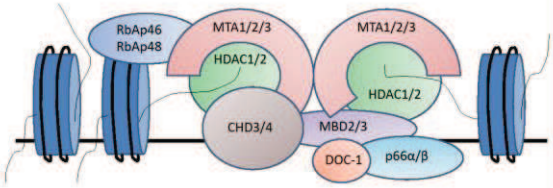
Morgan P. Torchy, Kareem Mohideen, Sinthuja Peiris, Isabelle Hazemann, Arnaud Depaux, Ali Hamiche, Bruno P. Klaholz

IGBMC (Institute of Genetics and Molecular and Cellular Biology), Department of Integrated Structural Biology, Illkirch, 67404, France

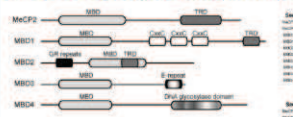
NuRD: Nucleosome Remodelling and histone Deacetylation

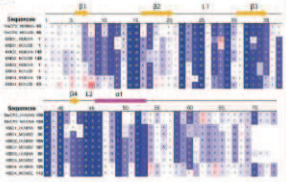
Discovered in 1998, it is to date the only known complex coupling two independent chromatin-remodelling activities: an ATP-dependent remodelling activity and a histone deacetylation function.
NuRD is highly conserved among superior eukaryotes, and is expressed in a large variety of tissues. It forms a large macromolecular assembly that consists of different proteic subunits, each exhibiting several homologs and isoforms. This leads to a horde of coexisting NuRD complexes, depending on the cellular, tissue, physiological or pathological context.

Composition : CHD3 or CHD4 (Chromodomain, Helicase, DNA binding domain)
HDAC1 or HDAC2 (Histone Deacetylases)
MTA1, MTA2 or MTA3 (Metastasis Associated)
MBD2 or MBD3 (Methylated CpG-Binding Domain)
RbAp46 or RbAp48 (Retinoblastoma Associated protein)
p66α or p66β (GATA Zinc Finger Domain protein 2)
DOC-1 (Deleted in Oral Cancer)

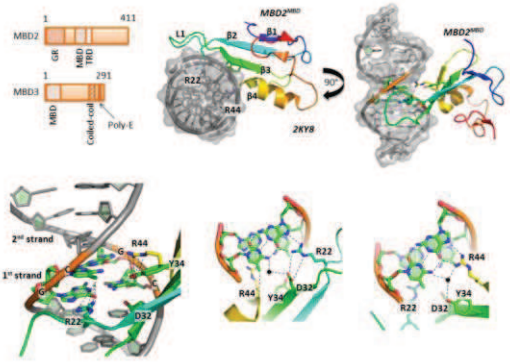


MBD: Methyl-CpG binding domain

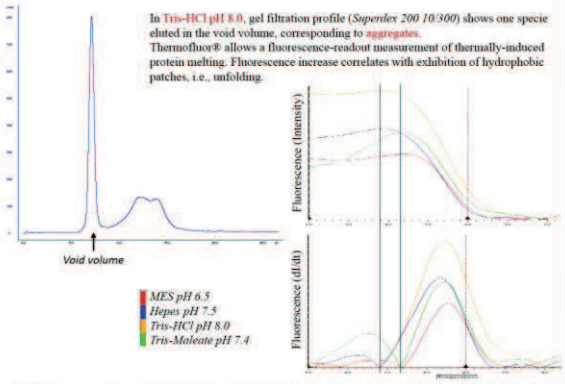




MBD proteins are able to specifically recognize and bind symmetrically methylated CpG islands.
However, due to an evolutionary mutation (Y34F), MBD3 in mammals have lost this ability and is only able to bind to non-modified CpG islands.



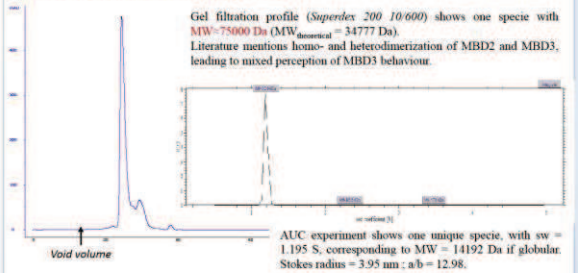
MBD3: buffer optimization using Thermofluor®



In Tris-HCl pH 8.0, gel filtration profile (*Superdex 200 10/300*) shows one specie eluted in the void volume, corresponding to aggregates.
Thermofluor® allows a fluorescence-readout measurement of thermally-induced protein melting. Fluorescence increase correlates with exhibition of hydrophobic patches, i.e., unfolding.

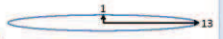
→ **Melting temperature is shifted towards slightly higher temperatures in MES pH 6.5 and Tris-Maleate pH 7.4. These buffers appear thus to be a better choice for MBD3 stabilization.**

MBD3: a partially unfolded protein



Gel filtration profile (*Superdex 200 10/600*) shows one specie with MW=75000 Da ($MW_{theoretical}$ = 34777 Da).
Literature mentions homo- and heterodimerization of MBD2 and MBD3, leading to mixed perception of MBD3 behaviour.

AUC experiment shows one unique specie, with sw = 1.195 S, corresponding to MW = 14192 Da if globular.
Stokes radius = 3.95 nm ; a/b = 12.98.



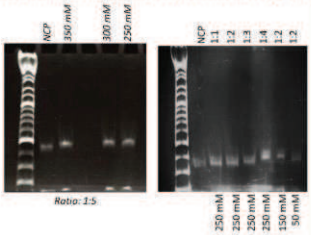
→ **In solution, MBD3 does not dimerize, and is partially unfolded.**

MBD3: complexes on nucleosomes

MBD3 was complexed with in-house reconstituted nucleosomes, and final salt concentration was optimized using EMSA gels.

The low molecular weight of MBD3 (34 kDa) makes it difficult to properly observe shifts on gel. Other technics such as MALLS, gel filtration or AUC are not resolutive enough to differentiate between a 200 and 230 kDa specie.

→ **Nevertheless, shifts are observed for MBD3 on nucleosomes, in a salt concentration range from 100 to 350 mM.**

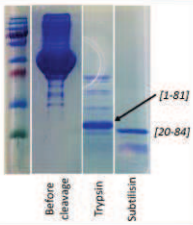


MBD3: mild proteolysis

MBD3 was incubated with different proteases to check for mild proteolysis.
Out of six different proteases (Chymotrypsin, Trypsin, Elastase, Papain, Subtilisin and Endo Glu-C), only two showed proteolysis: trypsin and subtilisin.

Mass spectrometry showed that the cleaved parts of MBD3 correspond to the N-terminal end, i.e., the MBD domain [1-72].

→ **These results suggest that secondary structures exist in the N-ter domain of MBD3, in absence of DNA or other partner.**

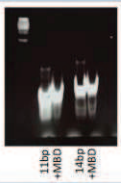


MBD3: MBD domain on DNA

The MBD domain of MBD3 [2-72] was complexed with DNA oligos (11 and 14bp). Despite the 1:1 ratio, shifts can be observed.

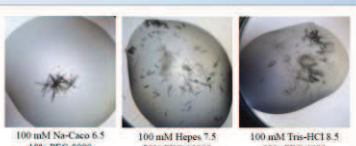
The same was done on nucleosome, with different ratios of $MBD3^{MBD}$ in excess : no shift could be observed.

→ This suggests that the C-ter part of MBD contains motifs to stabilize the protein on nucleosomes.



MBD3: crystallization trials

Crystallization of MBD3 alone was, for now, unsuccessful.
During *in situ* mild proteolysis with subtilisin, crystals were obtained, but further optimization is needed.



100 mM Na-Caco 6.5
18% PEG 8000
200 mM Ca^{2+}-Acetate

100 mM Hepes 7.5
20% PEG 10000
8% Ethylen Glycol

100 mM Tris-HCl 8.5
30% PEG 4000
200 mM $MgCl_2$

Structure and function insights into the chromatin-remodelling complex NuRD

Morgan P. Torchy, Ali Hamiche, Bruno P. Klaholz

Submitted to *Cell Mol Life Sci*. In review. 2014.

In an organism, every cell contains the same genetic material. Nevertheless, evolution has made possible the selective expression of some genes, and the repression of some others, and thus allowed cell specialization. With the establishment of differential expression patterns, cells can differentiate and organisms can develop. For a given cell, various normal and pathological processes can occur, such as reactions to stress stimulation (nutrient deficiency, hypoxia, lack of growth factors, etc.), pathologies related to deregulations of gene expression (cancers, etc.) or simply the progress of the cell cycle. This modulated gene expression is made possible by chromatin remodelling, a process that is thought to be related with the accessibility of the DNA of target genes to transcription factors or RNA polymerase in particular.

In 1942, Conrad Waddington coined the term "epigenetic", the branch of biology which studies "the causal interactions between genes and their products, which brings the phenotype into being". Indeed, genes and more generally, chromatin, are targeted by covalent modifications, which can be recognized by proteic effectors, allowing the recruitment of enzymes and other partners involved in chromatin remodelling. In 1998, several groups described a complex exhibiting an ATP-dependent remodelling activity, similar to that of ySWI/SNF from *Saccharomyces cerevisiae*, and coupled to a histone deacetylation function. This complex, called NURD, NRD, Mi-2 complex, and finally, NuRD, standing for "Nucleosome Remodelling and histone Deacetylation", is, to date, the only known complex coupling two independent chromatin-remodelling activities[1-4]. One possible reason for that could be that the ATP-remodelling activity is necessary for the Histone Deacetylase (HDAC) subunits to access their target[5]. This idea is supported by the observation that in absence of ATP, deacetylation is only possible on histone octamers, and not on nucleosomes. The binding site of HDACs could be somehow protected by the DNA, and thus inaccessible. Experiments carried out to determine whether ATP could stimulate deacetylase activity did not show any significant effect on free histone octamers. By contrast, when nucleosomes were tested, ATP was shown to stimulate deacetylase activity by two-fold: without ATP, 30-35% of acetylated H4 histones were deacetylated, while in the presence of ATP, 60-70% were[2].

The NuRD complex is highly conserved among superior eukaryotes, and is expressed in a large variety of tissues. It forms a large macromolecular assembly that consists of different proteic subunits; however, different homologs and isoforms have been described for each of those subunits, leading to a horde of coexisting NuRD complexes, depending on the cellular, tissue, physiological or pathological context. Moreover, the stoichiometry of the different subunits remains an open question. Recently, the development of a new label-free quantitative mass spectrometry method, applied to the analysis of NuRD, suggested that it is composed of one CHD3 or CHD4 protein (Chromodomain, Helicase, DNA binding domain), one HDAC1 or HDAC2, three MTA1/2/3 (Metastasis Associated), one MBD3 (Methylated CpG-Binding), six RbAp46/48 (Retinoblastoma Associated protein), two p66α or p66β and two DOC-1 (Deleted in Oral Cancer)[6]. Those data are nevertheless in contradiction with the structural analysis of the HDAC1/MTA1 complex showing a dimerization of MTA1, suggesting the presence of two MTA1/2/3 and two HDAC1 or HDAC2 in NuRD[7]. The specificities of each isoform, together with the sharing of competences such as the opposite activities of deacetylation and remodelling, ensure that NuRD is a major actor in various biological processes, like embryonic development, cellular differentiation, haemato- and lymphopoeisis, tumour growth inhibition, or the general repression of transcription. Furthermore, it directly interacts with various partners, like the lysine specific demethylase 1 (LSD1/KDM1A), Ikaros, Aiolos, Helios, B-cell lymphoma 6 (BCL6), the oestrogen receptor α (ERα/NR3A1) or Oct4/Sox2/Klf4/c-Myc (OSKM). This highlights the very broad and general role of NuRD, especially given that it is the most abundant form of deacetylase in mammals.

The aim of this review is to give an up-to-date and comprehensive overview of the NuRD complex and the structure-function relationships of its different subunits. Great efforts have been made these past few years to lift the veil regarding biochemical, genetic and structural data to fully understand the precise action of a given NuRD complex *in vitro* but also in its environment, as justified by the quasi-ubiquitous role that it plays. In this regard, numerous studies focus on isolated subunits, and the results obtained are extrapolated to the whole complex, leading to a multitude of scopes of activities, which

need to be placed back into the context of the entire complex.

CHD3/4: the ATP-dependent chromatin-remodelling

ATP-dependent chromatin-remodelling enzymes are helicases which utilize the energy brought by the hydrolysis of ATP to destabilize interactions between DNA and histone proteins that constitute the core of the nucleosome. The chromatin structure is thus altered, by displacement of nucleosomes along the DNA, assumingly to make specific sequences available, or by eviction or replacement of histones. These enzymes are part of the SF2 superfamily and Snf2 family[8]. In this group, the CHD subfamily is composed of a characteristic pattern, with two tandem chromodomains at the N-terminal part, in addition to the ATPase domain. In yeast, only one CHD protein has been identified, yCHD1, while four in *Drosophila melanogaster* (dCHD1-4) and nine in mammals (hCHD1-9) exist. yCHD1 is closely related to d/hCHD1 and d/hCHD2 with approximately 35% of overall sequence identity. The other CHDs on the contrary share only sequence similarity with yCHD1 within their defined domains, their extremities being highly variable.

In the context of NuRD, CHD3 and CHD4, also called Mi-2α and Mi-2β, are the two homologs found to ensure ATP-dependent chromatin remodelling. The latter is the most abundant in the NuRD complex, although it seems that both proteins can coexist within the same complex. At least three molecular species can thus be found: Mi-2α/NuRD, Mi-2α/Mi-2β/NuRD and Mi-2β/NuRD. This raises the question whether this protein is present in the NuRD complex in two or else copies, which is still unclear. The Mi-2 protein was initially identified as an autoantigen in patients affected with dermatopolymyositis[9,10]. About one quarter of these patients are positives to anti-Mi-2 antibodies. While the correlation between those tumour developments and the presence of anti-Mi-2 antibodies hasn't been proven formally, in 20 to 25% of the cases, the patients develop an ovarian, colorectal, lung, pancreatic, stomachic or lymphatic cancer[8,11].

CHD3 and CHD4 are large ATPases, with a molecular mass of about 220 kilo Daltons. Their domain organization (fig. 1a) comprises two conserved plant homeodomains (PHD) fingers, two tandem chromodomains (Chromatin Organization Modifier), and a

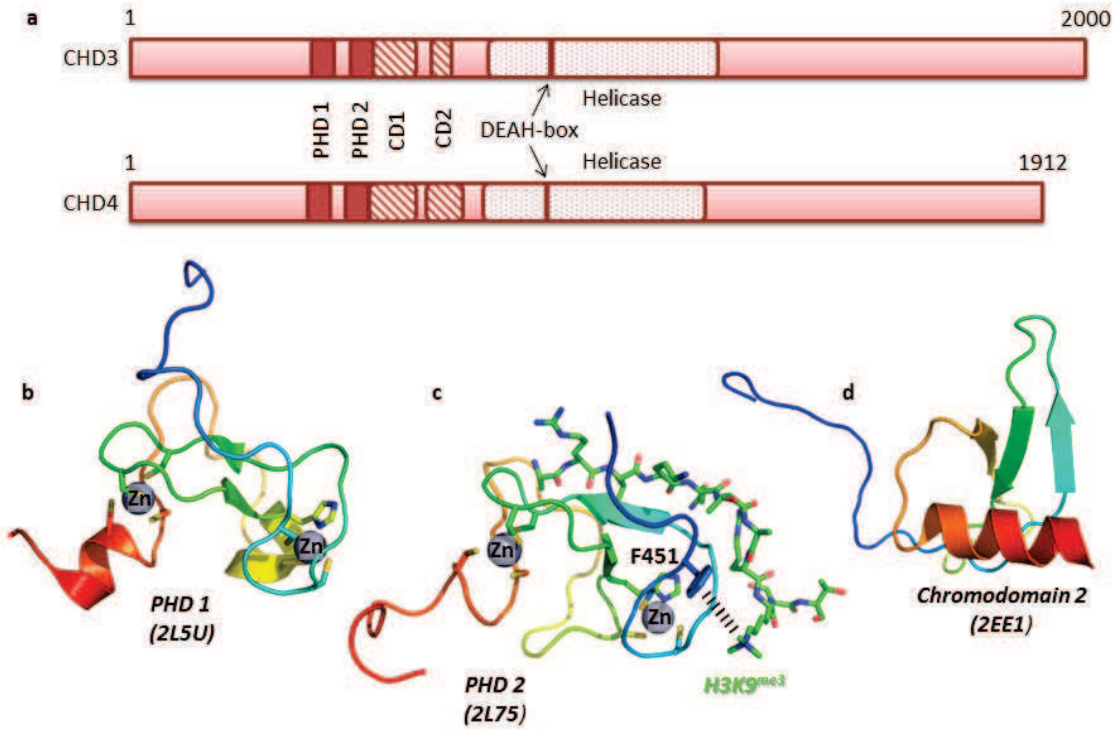


Figure 1, a: schematic description of the CHD3 and CHD4 domains. CD: ChromoDomain; DEAH-box: Asp-Glu-Ala-His-box. b,c,d: the published NMR structures of the two PHD domains and the second chromodomain of CHD4 are represented with their pdb accession number. Zinc ions are represented by grey spheres. In particular, one can notice the π-cation stacking interaction allowing discrimination between the methylated and non-methylated state of H3K9 by residue F451 of the second PHD domain of CHD4.

SWI2/SNF-like helicase domain[12]. They are highly conserved among the vegetable and animal kingdoms, although absent in yeast. The activity of Mi-2 proteins from three different species (*Drosophila melanogaster; Xenopus laevis; Homo sapiens*) were shown to be stimulated by chromatin but not by free DNA or histones[1,13,14]. This implies that these enzymes are implicated in the recognition of the nucleosome rather than of its individual components. NMR solution structures of individual chromodomain (fig. 1d) and the two PHD domains have been determined (fig. 1b,c), revealing a bivalent mode of binding to histone H3 tail[15,16]. Indeed, the two PHD domains of CHD4 are able to bind two distinct H3 tails, within a single nucleosome or on adjacent nucleosomes[17]. The post-translational modifications of those tails govern the binding affinity of CHD4: H3K9 trimethylation promotes the binding of the enzyme (figure 1), while H3K4 methylation abolishes it[18].

Additionally, two isoforms of CHD3, CHD3.1 and CHD3.3, exhibit a C-terminal SUMO-interaction motif (SIM) allowing them to interact with the sumoylated form of the KRAB-associated protein-1 (KAP-1), a major component of heterochromatin. KAP-1 phosphorylation by the ataxia telangiectasia mutated protein (ATM), as observed in the case of DNA double strand breaks, inhibits this interaction with CHD3 and leads to chromatin decompaction[19-21].

The Mi-2 proteins have also shown their crucial role in the development of some model organisms. In *Caenorhabditis elegans*, both CHD3 and CHD4 are implicated in the Ras signalling pathway, regulating cell fate in the hermaphrodite vulva and male tail[22]. In *Arabidopsis thaliana*, the CHD3 homolog PICKLE is implicated in the auxin signalling pathway, required for lateral root initiation and development[23]. In human, both CHD3 and CHD4 interact with transcription factors Ikaros, Aïolos and Helios, and target NuRD to specific promoters involved in lymphocytic development and proliferation[24-26]. Among those genes, one could mention CD179b, for progenitor B cells to precursor B cells differentiation; dntt, required for the V-DJ recombination; or CD4 and CD8a, for thymocytes maturation. These data suggest that Mi-2 could also have an important role in mammal development, but the lack of genetic models remains today a crucial bottleneck to further study these enzymes.

HDAC1/2: deacetylating histone lysines

During a ligand screen aiming at blocking the tumorigenic effect of the v-sis gene on 3T3 fibroblasts, a new molecule, trapoxin, was discovered[27]. Trapoxin-treated cells were shown to be hyperacetylated and their deacetylation function was inhibited, but the binding target remained unknown[28]. Finally, in the mid-1990's, this question was solved using trapoxin as a bait to purify its target by affinity chromatography. Mass spectrometry studies revealed that it was a homolog of the yeast deacetylase Rpd3p[29]. Since then, eighteen histone deacetylases have been identified and divided into three classes, HDACs I, II and III. The first one comprises the nuclear HDACs 1-3 and HDAC 8, based on a strong homology with yRpd3p. The second class gathers the cytosolic and nuclear HDACs 4-7 and HDACs 9-11. Finally the third class, called sirtuins, comprises SIRT 1 to 7, homologs of the ySir2 histone deacetylase, and can be either cytosolic, nuclear, nucleolar or mitochondrial.

In the NuRD complex, the subunits ensuring histone deacetylation are HDAC1 and HDAC2 (fig. 2a). These 55 kDa proteins are highly conserved and ubiquitous in all eukaryotes. They share 83% of sequence identity, and their double knock-out in T-cells or embryonic stem cells leads to a decrease by half of the total deacetylase activity of these cells[30]. They are thus the two predominant enzymes in terms of histone deacetylation activity in mammalian cells. Though HDAC1 and HDAC2 don't exhibit any DNA-sequence specificity, it's been suggested that they could interact with coactivators and corepressors to target DNA in a more specific manner[31].

Sequence alignments of class I HDACs showed major differences in the C-terminal domain, which is entirely missing in HDAC8. This domain is required in HDAC1 and 2 to bind to partners in the context of proteic complexes, and is furthermore post-translationally modified to regulate their catalytic activity (reviewed in Segré & Chiocca, 2011). Nevertheless, the first crystal structure of a HDAC, that of HDAC8 in complex with different inhibitors, paved the way for structural understanding of the class I HDACs[32,33]. These proteins are composed of a single α/β domain (fig. 2b), consisting of an eight-stranded-parallel-β-sheet at the centre of thirteen α-helices. These secondary structures are connected through long loops, thus creating the catalytic core domain of these enzymes. The active site consists of a long tunnel with a minimum depth of 8 Å, also referred to as lipophilic tube leading to the catalytic machinery. This tunnel is occupied by the four carbons of the side chain of the acetylated lysine, stabilized by hydrophobic contacts with residues G151, F152, H180, F208, M274 and F306 (HDAC8 numbering). All these residues are conserved among the class I HDACs, with the exception of M274

being a leucine in all other class I HDACs. Finally, the end of the tunnel accommodates a zinc ion, chelated by five coordination bonds in a trigonal bipyramidal fashion, and stabilized by the carboxylic oxygen of residues D178 and D267, and by the Nδ1 atom of the H180 side-chain. The carbonyl oxygen of the acetyl moiety carried by the acetylated lysine, as well as a water molecule, occupy the two other coordination sites. More recently, the structures of HDAC2 in complex with inhibitors[34,35], and the one of HDAC1 in complex with the ELM (Egl-27 and MTA1 homology) and SANT (Switching-defective protein 3, Adaptor 2, Nuclear receptor co-repressor, Transcription factor IIIB) domains of MTA1[7] (described later in this paper) shows the same global structure of the core HDAC protein.

It has been observed that inhibitors of the hydroxymate class, in the manner of SAHA (Suberoylanilide hydroxamic acid) or trichostatin A, bind to the catalytic site in roughly the same way as acetylated lysines, with fast binding kinetics and nanomolar K_d range over a large majority of class I and II HDACs. This is explained by the direct access of the ligand through the lipophilic tube, chelating the zinc ion with its hydroxamic group (fig. 2c). In contrast, inhibitors of the benzamide class, like entinostat and mocetinostat, are also located in the lipophilic tube, but their thiophene group is accommodated in a deeper pocket, named "foot pocket" (fig. 2d). This pocket is formed by flipping and shifting of the two M31 and L140 residues (HDAC2 numbering). Those two residues are conserved among HDAC1-HDAC3 but not HDAC8 and class II HDACs, giving rise to a higher specificity of this class of inhibitors. Finally, the central secondary amide moiety of these inhibitors chelates the zinc ion, locking the molecule in place. This explains the slower kinetics of benzamides, compared to hydroxymates, together with the higher specificity for class I HDACs, and in particular, HDAC1 and HDAC2.

The biochemical and genetic properties of HDAC1 and 2 make it difficult to understand their specific functions within the NuRD complex. Indeed, although deacetylation is largely associated with gene repression, knock-out experiments showed that several genes become repressed in the absence of HDAC 1 or HDAC2[36-39]. This suggests that these

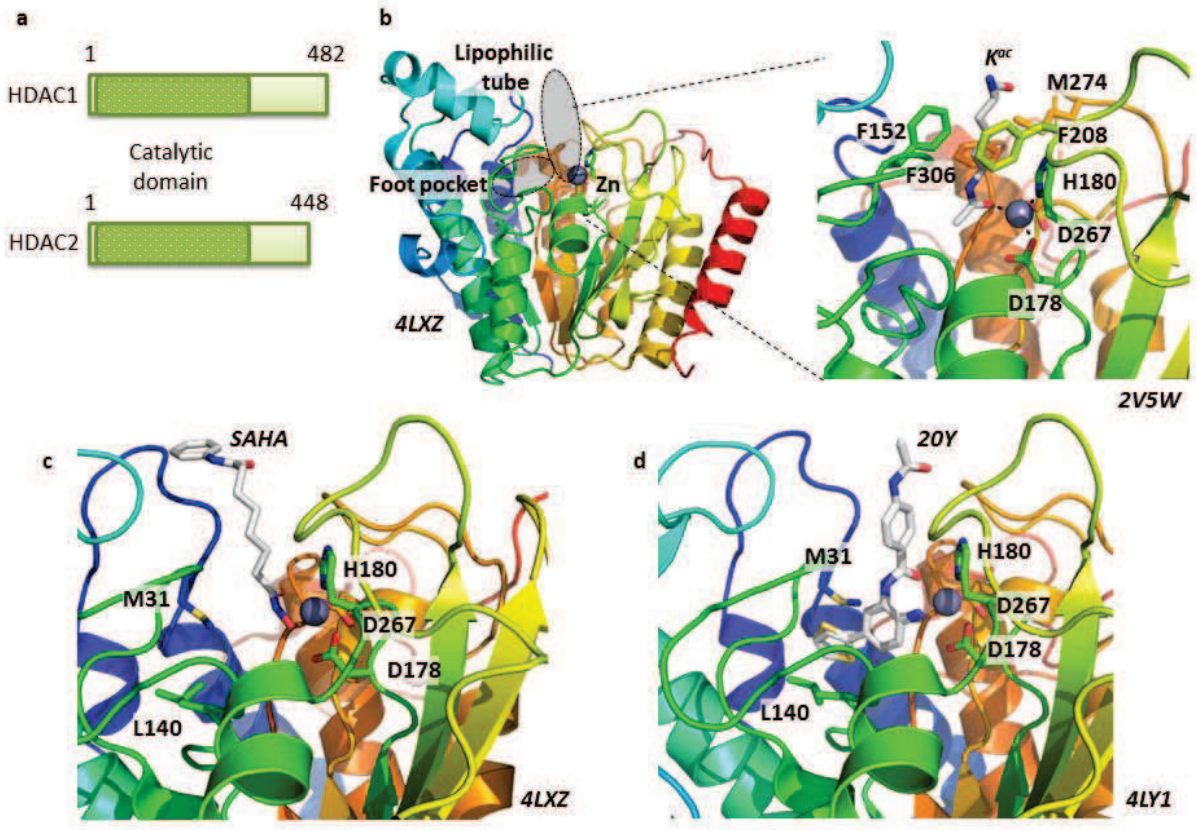


Figure 2, a: schematic description of the HDAC1 and HDAC2 domains. b: the global X-ray structure highlights the lipophilic tube as well as the foot pocket, and shows crucial residues for the zinc ion coordination and for substrate interaction (K^{ac}: acetylated lysine). Zinc ions are represented by grey spheres. c,d: the two structures of HDAC2 in complex with a hydroxymate (SAHA) and a benzamide (20Y: 4-acetylamino-N-2-amino-5-thiophen-2-ylphenylbenzamide) show the M31 and L140 residues, forming the gate of the foot pocket.

two enzymes could also have a role in gene activation. Further studies carried out by treating embryonic stem (ES) cells with trichostatin A showed both a decreased expression of pluripotency-related genes and an increase of lineage-specific genes, indicating a negative as well as positive regulation activity. By chromatin immunoprecipitation (ChIP), it has been shown that these enzymes can localize at some transcriptionally active loci in human[40], mouse[41] and yeast[42], corresponding to DNase I hypersensitive sites. In particular, HDAC1 has been detected in promoter regions, on pluripotency genes in ES cells (like fgf4, mbd3, nanog, oct4, sox2, tbx3 or zfp42) and trophoblast-lineage genes in trophoblast stem cells (like bmpr1a, cdkn1c, cdx2, elf5, hand1, msx2 or tcfap2c)[41], while HDAC2 is present in both promoters and gene bodies.

A commonly observed phenomenon when knocking-out HDAC1 and HDAC2 is the decrease of cell proliferation[38,43-45]. The loss of these enzymes induces an overexpression of the kinases p21/WAF1/CIP1[43,46] and p57/Kip2[38] inhibitors, preventing G1/S phase transition. HDACs inhibitors have been tested in numerous cases of cancers, with the aim of limiting tumour growth[47], but most of these inhibitors, in the manner of SAHA (approved and commercialized under Vorinostat or Zolinza) are large-spectrum inhibitors of class I and II HDACs and therefore lead to significant side-effects. Studies carried out on mice showed that the use of specific HDAC1 and/or HDAC2 inhibitors, like benzamides described above are equally efficient with respect to antiproliferative effects, but with potentially reduced side-effects[44,48]. Given their biochemical and genetic identity, it is not surprising that HDAC1 and HDAC2 are redundant enzymes: knock-outs of these showed no deleterious phenotype, the remaining enzyme complementing the missing one[30,37-39,43,49-52]. Why this redundancy exists remains to be addressed.

MTA1/2/3: reading histone tails and promoters

MTA proteins were the last ones to be characterized within the NuRD complex. The first representative in this family, temporarily called p70, then MTA1, was isolated after the observation of its differential expression pattern observed by cDNA library screening using the 13762NF rat mammary adenocarcinoma metastatic system[53]. But despite the overexpression of this protein, one had to wait for the discovery of NuRD and the presence of MTA proteins in this complex to start understanding the role of this family[2,4].

Phylogenetic studies suggested that the mta gene underwent duplications to lead to the three loci found in vertebrates (mta1 on chromosome 14q, mta2 on chromosome 11q and mta3 on chromosome 2q), mta2 being the nearest relative to the ancestral non-vertebrate gene[54]. Those three genes encode the three proteins MTA1, MTA2 and MTA3, and also three alternative-splicing products: MTA1S, MTA1-ZG29p and MTA3L[55]. The three canonical MTA proteins have a molecular weight of 80, 70 and 65 kDa, respectively, and share 68% of sequence homology between MTA1 and MTA2 and 73% between MTA1 and MTA3. This strong homology is especially due to the N-terminal domains, the C-terminal parts being more variable. With the exception of MTA1-ZG29p, all the MTA proteins possess various highly conserved domains (fig. 3a): a bromo adjacent homology domain (BAH; 70% of identity between $MTA1^{BAH}$ and $MTA2^{BAH}$ and 76% of identity between $MTA1^{BAH}$ and $MTA3^{BAH}$), an Egl-27 and MTA1 homology domain (ELM; 76% of identity between $MTA1^{ELM}$ and $MTA2^{ELM}$ and 78% of identity between $MTA1^{ELM}$ and $MTA3^{ELM}$) and a SANT domain (87% of identity between $MTA1^{SANT}$ and $MTA2^{SANT}$ and 94% of identity between $MTA1^{SANT}$ and $MTA3^{SANT}$). The role of these domains has not been fully studied yet in the context of MTA proteins. Nevertheless, some functional insights come from related proteins. For example, the SANT domains in Ada2 and SMRT seem to interact primarily with unmodified histone tails[56,57] and BAH of Rsc2 is implicated in histone H3 binding[58], while $ORC1^{BAH}$ recognizes H4K20me2[59].

Expression regulation for the mta genes is still little-known to date, however, preliminary results are available. For example, heregulin, a growth factor witch binds to the human epidermal growth factor receptors 3 and 4 (HER3 and HER4) transmembrane receptors, is able to induce MTA1 expression in breast cancer cells[60]. It has also been shown that the c-Myc proto oncogene could bind directly to the mta1 gene to activate its expression[61]. Moreover, MTA1 is overexpressed in hypoxia, and is responsible for hypoxia inducible factor 1 (HIF-1) stabilization by deacetylation, becoming then resistant to degradation by the 26S proteasome[62]. Additionally, MTA proteins are intimately linked to the oestrogen receptor ER[63,64], in breast cancer and mammary gland development[65]. The short MTA1S isoform, which is produced by alternative splicing inside a cryptic site of exon 14[66], directly interacts with ER and is responsible for its sequestration in the cytoplasm[66]. MTA1 also blocks ER-driven gene activation, by antagonizing the effect of oestradiol[60], while MTA2 can make breast cancer cells insensitive to oestrogens and tamoxifen, by deacetylation of ER itself[63]. Finally, the promoter of mta3 is directly

activated by ER-α, thanks to the presence of a half response-element ERE, and MTA3 seems to be involved in repression of some genes involved in invasive growth, like Snail[67] or Wnt4[68]. Consequently, MTA1 and MTA3 seem to have an opposite role. Expression patterns of those two proteins support this idea: MTA3 is largely expressed in healthy epithelial cells, and its expression decreases along with tumour growth, until complete shutdown at the carcinoma stage; on the contrary, MTA1 is gradually expressed, concomitantly with tumorigenesis. Finally, isoform MTA1-ZG29p is a product of the mta1 gene, including only the seven last exons. For this reason, it doesn't exhibit the three domains described previously, and its location seems to be restricted to zymogenic granules in the pancreas[69].

Recently, a first 3-Å-resolution crystal structure of HDAC1 in complex with MTA1 has been published[7] (fig. 3b,c). It shows the ELM and SANT domains of MTA1 (residues 162-335, i.e., one quarter of the protein), wrapping around HDAC1, with an interaction interface of 5185 Å² surface area. Three regions can be distinguished: the first one corresponds to the SANT domain of MTA1, composed of three α-helices (H1 to H3, fig. 3c). The interface with HDAC1 forms a positively charged pocket, which can accommodate an inositol tetraphosphate molecule (Ins[1,4,5,6]P_4) to stabilize this highly basic interaction, through residues K31, R270 and R306, among others (fig. 3b). This observation had previously been made on a HDAC3-SMRT^{SANT} complex, copurified from mammalian cells[70]. Further studies showed that mutations of the MTA1^{SANT} residues involved in coordination of Ins[1,4,5,6]P_4 lead to a reduced interaction between the SANT domain and HDAC1. However, MTA1 can still be tethered to HDAC1 in absence of Ins[1,4,5,6]P_4, through interaction of the ELM domain as described later. Studies on the HDAC3-SMRT showed a link between ageing of the complex, loss of Ins[1,4,5,6]P_4 moiety and decreased HDAC activity. However, addition of exogenous Ins[1,4,5,6]P_4 recovered the HDAC activ-

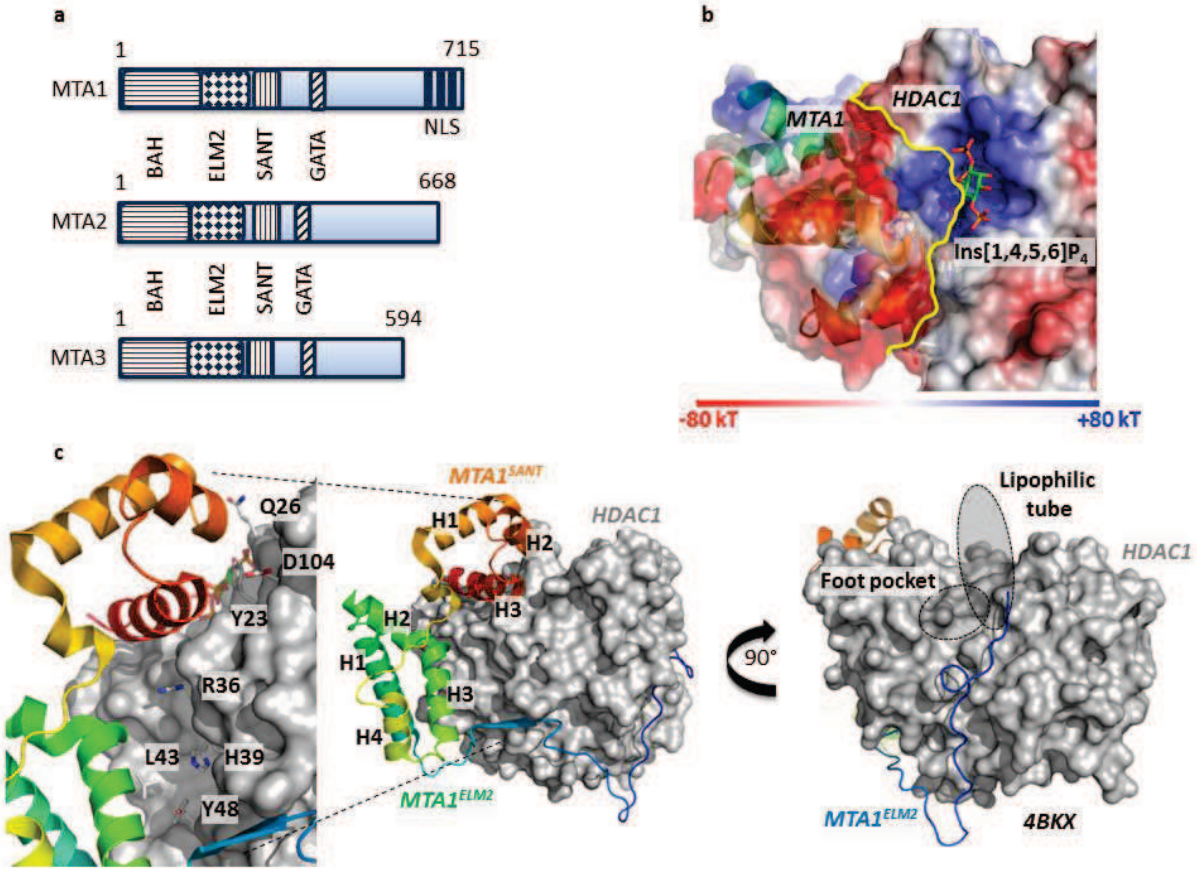


Figure 3, a: schematic description of the MTA1, MTA2 and MTA3 domains. NLS: Nuclear localization sequence. b: the structure shows how an inositol phosphate molecule can accommodate in the basic pocket (in blue) formed at the interface between HDAC1 and MTA1. In yellow, the limitation of the HDAC1-MTA1 interface. Structure superimposed on 4A69 (HDAC3-SMRT). Negative, neutral and positive surface electrostatic potentials are displayed in red, white, and blue, respectively. c: the global X-ray structure highlights the lipophilic tube as well as the foot pocket of HDAC1 (represented in grey; see also figure 2), and shows how MTA1 peptide (represented in colour) is wrapped around the deacetylase (in orange, the SANT domain; in green/blue, the ELM2 domain).Crucial residues involved in the HDAC1-MTA1 interaction are annotated.

ity with level higher than endogenous complexes. Similarly, the same observation has been made on the HDAC1-MTA1 complex, with an activation K_d around 5 µM. These elements tend to confirm Ins[1,4,5,6]P_4 as having a regulatory role of class I HDACs *in vivo*.

The second region correspond to three-quarters of the C-terminal region of the ELM domain, folded with four α helices (H1 to H4). The isolated ELM domain shows no folded secondary structure in circular dichroism, implying a radical structural reorganization upon binding to HDAC1[7]. Helices H1 and H3 mediate the interaction interface with HDAC1 (1278 Å²). Simultaneously, this region is responsible for dimerization of two MTA1 proteins, mediated by interactions between helices H1 and H4, and to a lesser extent, H2, of the two MTA1 molecules. Up to twenty-eight apolar residues (fourteen for each monomer) are involved in this dimerization, with an important interaction interface of 2332 Å². This is a rather clear confirmation that this dimerization interface is physiologically relevant, and that in terms of stoichiometry, the NuRD complex probably contains two MTA proteins, as well as two HDAC proteins. Finally, a third region corresponds to the N-terminal part of the ELM domain. It comprises a specific and conserved motif (EIRVGxxYQAxI), and forms an extended loop conformation. This long thirty-amino-acid-chain runs on the surface of HDAC1, inside a long apolar groove.

MBD2/3: DNA-binding and the connexion to methylation

The study of Methylated CpG-Binding domain Proteins (MBPs) started in 1989, after the fortuitous discovery of two proteins binding to methylated DNA. At this time, Bird and collaborators were seeking proteins able to bind to non-methylated DNA and likely to protect CpG islands from methyltransferases. Electromobility shift assays from mouse liver nuclear extracts showed the presence of two proteins, called MeCP1 and MeCP2 (Methylated CpG-binding Protein 1 and 2)[71,72]. MeCP2 was the first to be purified from mouse brain extracts. This 53 kDa protein exhibits what has then been described as a 90 residues N-terminal MBD domain, as well as a C-terminal transcription repression domain (TRD)[73,74] (fig. 4a). Sequence similarity searches in databases identified four other proteins, MBD1, MBD2, MBD3 and MBD4, all very conserved in vertebrates[75]. Among those MBPs, MBD2 and MBD3 share the highest sequence identity (77%). Furthermore, a single homolog of these two proteins, MBD2/3, is found in invertebrates. It is encoded by a single gene, in contrast to vertebrates where this gene probably underwent a duplication event. Indeed, mbd2 and mbd3 genes have a very similar genomic structure, varying only by the size of their introns. This supports the idea that MBD2 and MBD3 are probably the ancestral representatives of this family[76,77]. Later, the protein MeCP1 initially discovered along with MeCP2 turned out to be a MBD2/HDAC1 complex[78].

With a mass of approximately 43 and 33 kDa respectively, MBD2 and MBD3 are the smallest subunits of the complex, which are exclusive yet interchangeable within NuRD[79]. While MBD2 binds to methylated DNA[75], MBD3 has lost this ability in mammals. Indeed, the appearance of this class was accompanied by two point mutations in the mbd3 gene, leading to the incorporation of two new amino acids in positions 30 and 34 (a histidine and a phenylalanine, instead of a lysine and a tyrosine, respectively). This abolishes the selectivity of this protein for methylated DNA[80-82]. While the very first studies fifteen years ago credited MBD2 with only a transient role in the complex, being in particular a NuRD recruiter to methylated DNA before its eviction and replacement by MBD3[31,83], other studies since have shed light on a MBD2/NuRD complex, biochemically and functionally distinct from the MBD3/NuRD complex[79,84]. In that sense, MBD2 knock-out experiments showed only little effects at the phenotype level, whereas MBD3 knock-out leads to embryonic lethality[85].

Recently, it has been proposed that MBD3 and, to a lesser extent, MBD2, were able to specifically bind to hydroxymethylated CpG islands. Notably, MBD3 seems to colocalize with TET1 (ten-eleven translocation methylcytosine dioxygenase 1), the protein responsible for hydroxylation of methylcytosines[86]. Additional experiments however failed to show an interaction between MBD3 and hydroxymethylated DNA[87]. Instead, MBD2 and MBD3 appear to be preferentially localized at CpG-rich transcription start sites (TSS). At TSS's, MBD2 predominantly binds methylated CpG islands, leading to a repression of gene expression; whereas MBD3 binds to non-methylated DNA, and is associated with active transcription[88,89]. Recently, NMR spectroscopic dynamic analysis suggested that MBD3 could have a counterbalancing role, binding in a competitive manner to non-methylated CpG islands, avoiding thus an abusive repression of those active genes by MBD2[90].

Several X-ray and NMR structures of MBDs in complex with DNA have been solved, revealing a common interaction pattern for all the MBPs[90-96]. In particular, two solution structures of MBD2 and one

solution structure of MBD3 have been solved, highlighting a quasi-structural identity between the two[90,95,97]. The MBD is characterized by an α/β sandwich, composed of an N-terminal four-stranded antiparallel β-sheet (β1: residues 6-8 in MBD3; β2: residues 15-20; β3: residues 32-37; β4: residues 41-43), and a C-terminal α-helix (residues 47-53). This α-helix is kept antiparallel against the β4 strand by hydrophobic contacts. Furthermore, the MBD exhibits three loops L1, L2 and C-terminal hairpin. L2 connects the α-helix and the C-terminal hairpin and is well defined in solution. In contrast, the long L1 loop between β2 and β3, composed of a dozen of residues, is more flexible. This appears to be a necessary prerequisite for binding to DNA (fig. 4b). Indeed, seven residues of this loop make contacts with one of the DNA strand, at the level of the major groove. The other DNA strand interacts mainly with residues in the α-helix and L2-loop.

Recognition of a CpG island is independent for each methylcytosine, as guessed by the absence of symmetry in the interaction domain. Arginines 22 and 44, which are conserved among all MBPs, interact with symmetrically arranged guanines inside a CpG island. The guanidinium group of R22 makes an hydrogen bond with the O6 and N7 atoms of the guanine base of the first DNA strand (fig. 4d); while the guanidinium group of R44 exhibits the same interaction pattern with the guanine base of the second DNA strand (fig. 4e). Both arginines lie in a plane with their interacting guanines, stabilized and locked by direct hydrogen bonding of residue D32 and water-mediated hydrogen bonding of residue Y34. This flat orientation allows the two arginine residues to pack against the methylated cytosine bases neighbouring their interacting guanines, and permitting weak van der Waals interactions (fig. 4c). Finally, the carbonyl group of R44 forms a weak CO-HC hydrogen bond with the methyl group of the cytosine base on the second DNA strand; while Y34 forms a water-mediated hydrogen bond to recognize the methylated cytosine of the first DNA strand. The integrity of those residues is crucial to ensure the

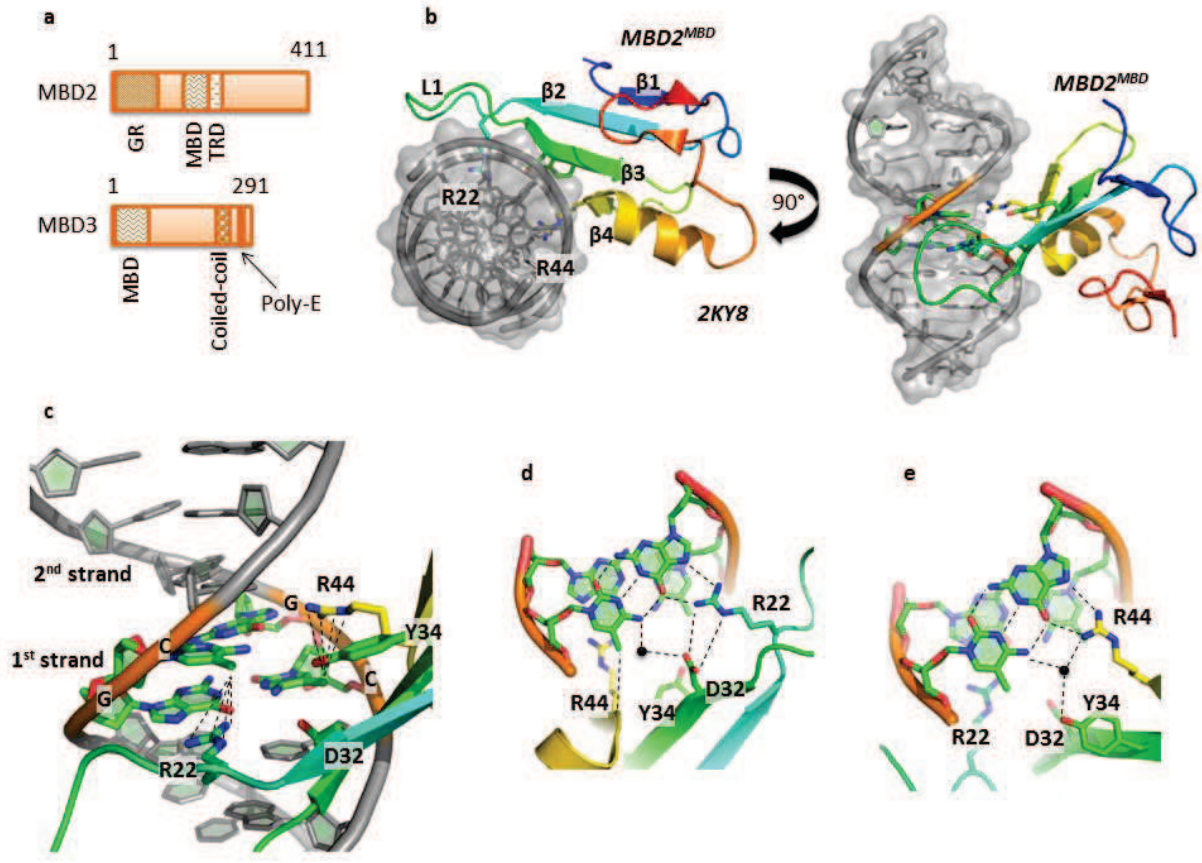


Figure 4, a: schematic description of the MBD2 and MBD3 domains. GR: Glycine-Arginine-rich region; MBD: Methyl-CpG Binding Domain; TRD: Transcription Repression Domain; Poly-E: Poly-glutamate. b: the NMR structure of MBD2^{MBD} shows the MBD domain of MBD2 interacting with a symmetrically methylated CpG island within an 11-bp DNA. c: a close-up of the interaction interface highlights the crucial residues, and shows Van der Waals forces between the methylated cytosines and the arginines. d,e: complementary CG-base pairs and their specific hydrogen bonds with MBD2 are shown. Water molecules engaged in water-mediated hydrogen bonds are represented by black dots.

binding to methylated DNA, as proven by mutagenesis experiments. In particular, Y34 turned out to be a key-residue in the recognition of the methylation state. Its mutation into a phenylalanine, as found in mammals, leads to a loss of affinity of methylated CpG islands. On the contrary, *Xenopus laevis* MBD3 doesn't exhibit this evolutionary mutation, and is thus still able to bind to methylated DNA. Also, the crystal structures of $MBD4^{MBD}$ in complex with different modified DNA show that Y96 (Y34 in xMBD3, T34 in m/hMBD3) is flipped out of the DNA interface, and is only making water-mediated hydrogen bonds with the phosphate backbone of the first DNA strand. This leads to a loss of specificity of MBD4 towards methylated DNA, at the cost of an increased binding of 5mCG/TG and 5mCG/hmCG islands[94].

MBPs are ubiquitous proteins, nevertheless exhibiting strong disparities depending on the cellular type and development stage. For example, in embryonic stem cells, MBD3 is the only predominantly expressed MBP. At the blastula stage of organismal development, MBD2 and MBD4 become detectable, and finally MeCP2 after the blastocyst implantation[98,99]. In adults, expressions patterns depend on the cellular type: MBD3 (along with MeCP2 and MBD1) is highly expressed in the brain, notably in the olfactory bulb, cerebellum, hippocampus and prefrontal cortex[100,101], while MBD2 has an almost opposite expression pattern, with mRNA quantities up to twenty times higher in some tissues, such as breast cells or cultured HeLa cells[102].

Recently, the central role of MBD3 in somatic cell reprogramming and cellular differentiation was suggested, interacting in particular with OSKM proteins, transcriptions factors responsible for maintaining totipotent state until blastocyst stage.[103-106]. However, opposite data obtained out of two different reprogramming systems suggest a context-dependent role of MBD3 in reprogramming, albeit further studies will be needed to confirm this theory. Another functional aspect, although in a completely different context, is the role of MeCP2 in the Rett syndrome. Seeing as mutations in the mecp2 gene are responsible for this neurodevelopmental disorder, lethal in men and causing neurological and psychiatric conditions in women, it has been thus suggested that mutations in other MBPs could also be linked to neurologic disorders. In this respect, the DNA of 226 Caucasian and Afro-American autistic patients and their relatives was thus analysed and alterations were found in mbd1-4 genes in 198 of them[107]. Interestingly, one of those alterations was found in exon 1 of the mbd3 gene. It corresponds to a point mutation (G>T at 1,543,563 in locus 19p13.3), leading to the incorporation of a new amino acid inside the MBD domain (R23M). This mutation, inducing the loss of a positive charge, has been observed in two Afro-American half-brothers, displaying late and unfunctional language acquisition. This mutation seems to be inherited from their disease carrier maternal grandmother, suggesting a sex-related effect. This residue is semi-conserved in MBD2 were it correspond to K167. Though published structures haven't shown any relevant role of this arginine in DNA binding, it is located right after the crucial R22 residue binding the CpG island (Fig. 4D). A new structure of the mutated gene will thus be needed to answer the question raised by the phenotype observed in R23M patients.

Finally, a two-hybrid screening on MBD2 highlighted in 2002 the interaction of two proteins, baptized p66α and p66β, and later, GATAD2A and GATAD2B, respectively (GATA Zinc Finger Domain Containing 2A/B)[108]. These proteins have shown to interact and colocalize with MBD2 and MBD3[108]. Moreover, the overexpression of both p66 proteins induces an increase of repressive action by MBD2; whereas p66 knock-outs allow a partial recovery of MBD2-repressed genes[109]. However, it is still not clear today whether these proteins are bona fide subunits of the NuRD complex, or occasionally just interacting partners.

RbAp46/48: ensuring a stable platform and binding histones

RbAp46 and RbAp48 (also called Rbbp7 and Rbbp4, respectively), were first identified because of their interaction with the tumor suppressor factor retinoblastoma (Rb)[110-112]. Later, studies showed their affinity for histones, and their presence in various deacetylation and remodeling complexes[54,113-115]. Although those two proteins share 90% of sequence identity[111], they exhibit different biochemical activities. Thus, RbAp46 associates with other proteins, notably histone acetyltransferase 1 (HAT1), involved in *de novo* histone H4 acetylation, on its lysine 5 and 12 residues[5,116]. This acetylation pattern is conserved among all eukaryotes, from yeast to human[117]; whereas RbAp48 is an essential chaperone for histone H3-H4 tetramer deposition on newly replicated DNA[118], and is especially found in the assembly complex CAF-1 (Chromatin assembly factor 1), with p150/CHAF1A and p60/CHAF1B. Nevertheless, RbAp46 and RbAp48 can be jointly found inside complexes, for example in association with HDAC1 and/or HDAC2, within the Sin3A or NuRD complexes, where they promote gene repression, including the one regulated by Rb[31,110,119]; they are also found within the polycomb repressive complex

(PRC2 and PRC3), with the histone-lysine N-methyltransferase EZH2, to methylate H3K27 or H1K26[120]; or in the nucleosome remodeling factor (NURF) complex, along with ISWI (SNF2L in human), where RbAp proteins are called NURF55[121].

RbAp46/48 are 48 kDa proteins that share a WD40 repeat sequence (fig. 5a). The great stability of those proteins allowed to date to solve seven crystal structures: two RbAp46 structures in complex with a histone H4 peptide, at 2.4 and 2.6 Å resolution[122] (fig. 5b); a structure of RbAp48 alone, at 2.3 Å resolution[123]; a structure of RbAp48 in complex with a FOG-1 (Friend of GATA) peptide at 1.9 Å resolution[124] (fig. 5d); and three structures of RbAp48 in complex with an MTA1 peptide at 2.5 and 2.15 Å resolution, respectively[125] (fig. 5c). Predictably, the RbAp proteins showed a structure similar to other WD40 proteins: a donut-shaped seven-bladed β-propeller, with a long N-terminal α helix (residues 9 to 28), lying on the seventh blade of the barrel, and a short C-terminal α-helix (residues 405 to 409), which is placed above and seems to extend the N-terminal helix. Finally, one particularity of these WD40 proteins is the presence of a seventeen residues loop, negatively charged, inside the sixth blade of the barrel, called PP loop (because of two successive prolines P362 and P363) (fig. 5b).

The crystal structure of RbAp46 in complex with a small histone H4 peptide shows an interaction interface of approximately 700 Å². This H4 peptide corresponds to residues 25 to 42 of the human isoform, i.e., the first α-helix of the histone fold and a part of the N-terminal tail. Though the structures previously described in other WD40 proteins highlightened an interaction interface on the front of the barrel, or even sometimes, inside it, histone H4 preferentially binds in a unique pocket located on the side of the barrel, and formed by the PP loop and the long N-terminal helix. Thus, hydrophobic residues I34, L37 and A38 in helix α1 of histone H4 interact with a hydrophobic patch composed of residues F29, L30, F367, I368 and I407 of RbAp46 (fig. 5b). A complex network of salt bridges and hydrogen bonds is also described between Q27, K31, R35, R36, R39 and R40 of histone H4; and E356, D357, D360, G361, P362, L365, N406, I407 and D410 of

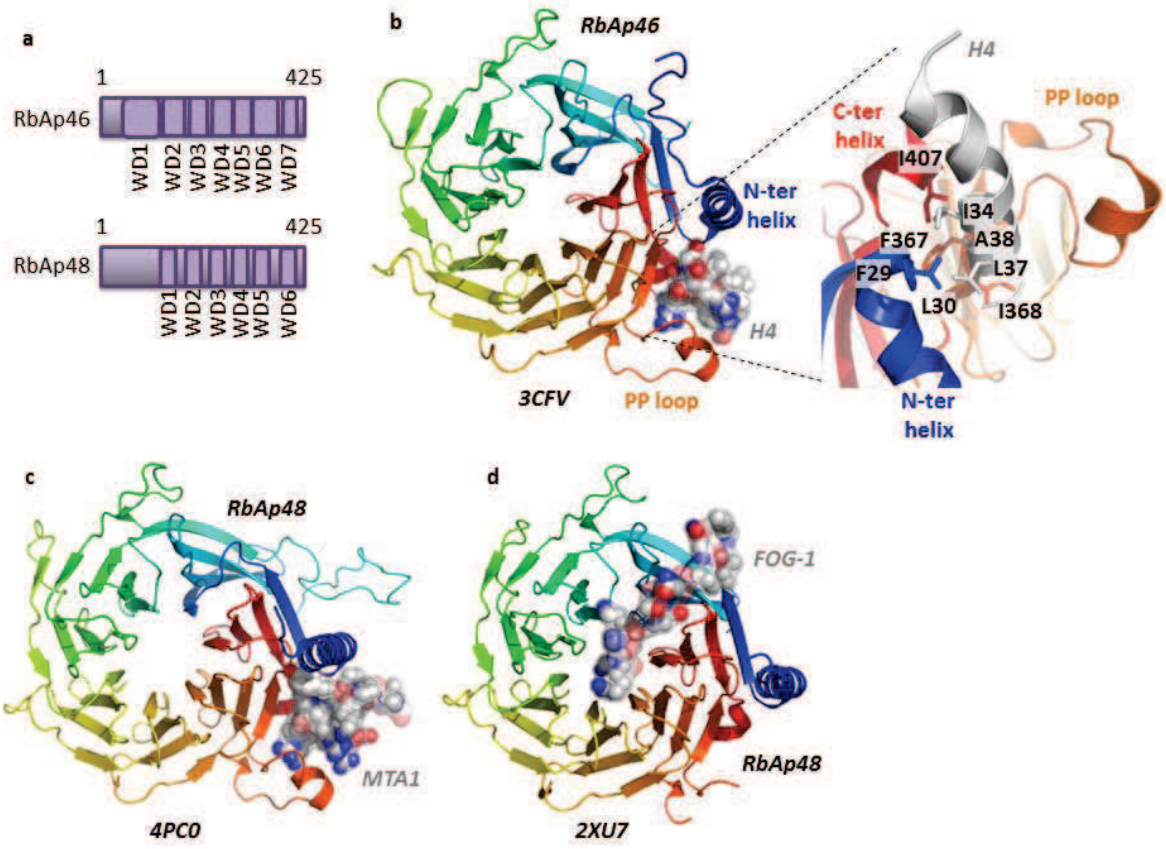


Figure 5, a: schematic description of the RbAp46 and RbAp48 domains. WD: Tryptophan-Aspartate domain. b: RbAp46/H4 complex shows a binding interface located on the side of the barrel, in a pocket formed by the PP loop and the long N-terminal helix. Crucial hydrophobic residues involved in the RbAp46-H4 interaction are annotated. c: the structure of the RbAp48/MTA1 complex shows a noticeably similar interaction interface, on the side of the barrel. d: the RbAp48/FOG1 complex shows a binding interface on the top of the barrel, extending towards the central channel.

RbAp46[122]. All these residues are conserved in RbAp48 and the yeast homolog p55, suggesting that the binding mechanism of these three proteins with histone H4 is similar. Finally, it has been shown that, in order to promote a proper interaction with RbAp46, the α1 helix of histone H4 must partially open, abolishing interactions with the α2 helix as well as those with histone H3, in particular through residues I34, L37 and A38. This observation raises thus the question of whether RbAp proteins interact with the nucleosome inside proteic complexes, or they suggest an accrued flexibility of the nucleosome[122]. Recently, pulsed electron-electron double-resonance (PELDOR) experiments have shown that RbAp48 can interact with a H3-H4 dimer, but not with a $(H3\text{-}H4)_2$ tetramer[126].

The structure of RbAp48 in complex with the fifteen N-terminal amino acids of the GATA-1 cofactor FOG-1, involved in erythroid and megakaryocytic cell differentiation, shows a binding interface located on the face of the barrel, which extends into the central channel[124] (fig. 5d). This interaction is different from that observed in the RbAp46/H4 complex, and is highly specific, because eight out of the thirteen residues in FOG-1 are involved in hydrogen or ionic bonds with RbAp48. In particular, this interface is composed of numerous acidic residues of RbAp48 (E231, E319, E179, E126, E395, E41), allowing the stabilization of a basic triade of FOG-1 (R3, R4 and K5). This interaction pattern can be extrapolated to RbAp46, as it shares those same conserved residues.

Finally, the RbAp48 structure, in complex with a short peptide of the C-terminal end of MTA1, shows a very similar binding site to the one observed in the complex with H4[125] (fig. 5c). This suggests that RbAp46/48 cannot simultaneously interact with MTA1 and histones.

Misregulations of RbAp46 and RbAp48 seem to be linked to tumorigenesis in several localizations, among which mammary and cervical tissues[127-129]. They indeed were shown to directly interact with the nuclear receptor ERα, and to affect ERα-regulated-gene expression[130]. For example, siRNA silencing experiments in MCF-7 cells were carried out, and Sox9 transcription factor and G2-cyclin gene activity were recorded. These genes are normally repressed by ERα in presence of estradiol, but it was shown that RbAp46 leads to their activation in presence of estradiol; in contrast, RbAp48 appears to maintain their repression in the absence of a ligand. Furthermore, a prolonged estradiol exposure of those cancer cells leads to a two to three-fold increase of RbAp46 levels. Together, these data suggest that RbAp46 could be a mediator favoring a continue ERα activity, while RbAp48 could ensure the basal repression of these genes in the absence of a ligand. Previous studies corroborate this idea, showing that repression of RbAp48 is involved in cervical cancer formation[131], and that an increase of RbAp46 levels prevents breast cancer development[127,129,132]. RbAp48 therefore appears to be a key therapeutic target for cervical cancer treatment[133]. Indeed, it has been shown that RbAp48 expression is favored by radiotherapy irradiations, and that SiHa, HeLa and Caski cells were radiosensitive, the more the level of RbAp48 is high.

In another context, a recent study carried out on eight human beings, aged 33 to 88, showed a differential expression pattern of RbAp48 in their brain, reduced along with the age, specifically in the dentate gyrus, a subregion of the hippocampus known for its life-long neurogenesis, and foreseen to be the seat for episodic memory[134]. Additional studies carried out on mice confirmed the role of RbAp48 in the memorization process. A young knock-out mouse has indeed less potential in memorizing new objects and environments; on the contrary, lentivirus-induced re-expression of RbAp48 in old mice helped increase their cognitive capacities. Those phenomena seem to be closely related to the activity of the RbAp48-binding partner CREB-binding protein(CBP)/p300, nuclear receptor-bound transcription coactivators increasing gene expression through their intrinsic histone H4 and H2B acetyltransferase activities.

DOC-1: the overlooked tumor-suppressor

Recently, copurification experiments carried out on recombinant MBD2 and MBD3-expressing stable cell lines revealed the presence of a new 12 kDa subunit called CDK2AP1 (Cdk2-associated protein 1) or DOC-1 (Deleted in oral cancer-1) inside both NuRD/MBD2 and NuRD/MBD3 complexes[79]. As its name suggests, this protein, a putative tumour suppressor interacting with CDK2, is inhibited in oral and colorectal cancers[135,136]. Later, mass spectrometry experiments confirmed the presence of this protein in NuRD[6,137].

The role of DOC-1 is still unclear; nevertheless, it was shown that overexpression in 293T cells would lead to a partial arrest of the cell cycle phase G1/S, together with a significant growth retardation[138]. This can be offset against the consequences of MBD2 overexpression promoting cell proliferation. This suggests thus that an opposite role for those two proteins exists inside the NuRD complex.

Structure-function relationship within NuRD and future prospects

From a functional point of view, it remains unclear today why evolution chose to assign two enzymatic activities within a single complex. Indeed, even though HDACs have proved their capacities to activate a subset of genes, these subunits still persist in being considered as general repressors, which raises the question of the apparent contradiction with the ATP-dependent remodeling activity of CHD3 and CHD4, known to allow breathing of the chromatin and thus, potentially activate gene expression. A long date proposal suggests that ATP-dependent remodeling of the chromatin is a prerequisite to allow other subunits of the NuRD complex, in particular HDACs, to access their substrate. However, this has never been clearly proven, and further experiments will be needed to confirm the mechanism underlying the function of NuRD.

Finally, from a structural point of view, it is intriguing how so many different proteins can interact with a complex. Structural studies of the whole NuRD complex will be needed to address the accessibility of factors to this macromolecular complex, and determine the molecular basis of inter-proteic interactions, such as with factors involved in cancer progression. Furthermore, relatively little is known about the intramolecular interactions within the entire NuRD complex, as illustrated by the remaining open question of the stoichiometry. Some works nevertheless constitute the blueprint for a better comprehension of the NuRD architecture, in the manner of the HDAC1/MTA1 complex or RbAp46/H4 complex structures. This indeed suggests the presence of two MTA and two HDACs subunits within the complex, as well as potentially two RbAp46/48 per nucleosome. Whether the latter work in synergy with CHD3/4 to destabilize histone octamer, as suggested by the binding of RbAp46 to H3-H4 dimer only, is also a remaining question to be answered.

1 Wade, P. A., Jones, P. L., Vermaak, D. & Wolffe, A. P. A multiple subunit Mi-2 histone deacetylase from Xenopus laevis cofractionates with an associated Snf2 superfamily ATPase. *Curr Biol* **8**, 843-846 (1998).

2 Xue, Y. *et al.* NURD, a novel complex with both ATP-dependent chromatin-remodeling and histone deacetylase activities. *Mol Cell* **2**, 851-861 (1998).

3 Tong, J. K., Hassig, C. A., Schnitzler, G. R., Kingston, R. E. & Schreiber, S. L. Chromatin deacetylation by an ATP-dependent nucleosome remodelling complex. *Nature* **395**, 917-921 (1998).

4 Zhang, Y., LeRoy, G., Seelig, H. P., Lane, W. S. & Reinberg, D. The dermatomyositis-specific autoantigen Mi2 is a component of a complex containing histone deacetylase and nucleosome remodeling activities. *Cell* **95**, 279-289 (1998).

5 Verreault, A., Kaufman, P. D., Kobayashi, R. & Stillman, B. Nucleosomal DNA regulates the core-histone-binding subunit of the human Hat1 acetyltransferase. *Curr Biol* **8**, 96-108 (1998).

6 Smits, A. H., Jansen, P. W., Poser, I., Hyman, A. A. & Vermeulen, M. Stoichiometry of chromatin-associated protein complexes revealed by label-free quantitative mass spectrometry-based proteomics. *Nucleic Acids Res* **41**, e28 (2013).

7 Millard, C. J. *et al.* Class I HDACs share a common mechanism of regulation by inositol phosphates. *Mol Cell* **51**, 57-67 (2013).

8 Hill, C. L. *et al.* Frequency of specific cancer types in dermatomyositis and polymyositis: a population-based study. *Lancet* **357**, 96-100 (2001).

9 Seelig, H. P. *et al.* The major dermatomyositis-specific Mi-2 autoantigen is a presumed helicase involved in transcriptional activation. *Arthritis Rheum* **38**, 1389-1399 (1995).

10 Ge, Q., Nilasena, D. S., O'Brien, C. A., Frank, M. B. & Targoff, I. N. Molecular analysis of a major antigenic region of the 240-kD protein of Mi-2 autoantigen. *J Clin Invest* **96**, 1730-1737 (1995).

11 Callen, J. P. & Wortmann, R. L. Dermatomyositis. *Clin Dermatol* **24**, 363-373 (2006).

12 Woodage, T., Basrai, M. A., Baxevanis, A. D., Hieter, P. & Collins, F. S. Characterization of the CHD family of proteins. *Proc Natl Acad Sci U S A* **94**, 11472-11477 (1997).

13 Brehm, A. *et al.* dMi-2 and ISWI chromatin remodelling factors have distinct nucleosome binding and mobilization properties. *EMBO J* **19**, 4332-4341 (2000).

14 Wang, H. B. & Zhang, Y. Mi2, an auto-antigen for dermatomyositis, is an ATP-dependent nucleosome remodeling factor. *Nucleic Acids Res* **29**, 2517-2521 (2001).

15 Kwan, A. H. *et al.* Engineering a protein scaffold from a PHD finger. *Structure* **11**, 803-813 (2003).

16 Mansfield, R. E. *et al.* Plant homeodomain (PHD) fingers of CHD4 are histone H3-binding modules with preference for unmodified H3K4 and methylated H3K9. *J Biol Chem* **286**, 11779-11791 (2011).

17 Musselman, C. A. *et al.* Bivalent recognition of nucleosomes by the tandem PHD fingers of the CHD4 ATPase is required for CHD4-mediated repression. *Proc Natl Acad Sci U S A* **109**, 787-792 (2012).

18 Musselman, C. A. *et al.* Binding of the CHD4 PHD2 finger to histone H3 is modulated by covalent modifications. *Biochem J* **423**, 179-187 (2009).

19 Goodarzi, A. A., Kurka, T. & Jeggo, P. A. KAP-1 phosphorylation regulates CHD3 nucleosome remodeling during the DNA double-strand break response. *Nat Struct Mol Biol* **18**, 831-839 (2011).

20 Ivanov, A. V. *et al.* PHD domain-mediated E3 ligase activity directs intramolecular sumoylation of an adjacent bromodomain required for gene silencing. *Mol Cell* **28**, 823-837 (2007).
21 Lee, D. H. *et al.* Phosphoproteomic analysis reveals that PP4 dephosphorylates KAP-1 impacting the DNA damage response. *EMBO J* **31**, 2403-2415 (2012).
22 von Zelewsky, T. *et al.* The C. elegans Mi-2 chromatin-remodelling proteins function in vulval cell fate determination. *Development* **127**, 5277-5284 (2000).
23 Fukaki, H., Taniguchi, N. & Tasaka, M. PICKLE is required for SOLITARY-ROOT/IAA14-mediated repression of ARF7 and ARF19 activity during Arabidopsis lateral root initiation. *Plant J* **48**, 380-389 (2006).
24 Georgopoulos, K., Winandy, S. & Avitahl, N. The role of the Ikaros gene in lymphocyte development and homeostasis. *Annu Rev Immunol* **15**, 155-176 (1997).
25 Kim, J. *et al.* Ikaros DNA-binding proteins direct formation of chromatin remodeling complexes in lymphocytes. *Immunity* **10**, 345-355 (1999).
26 Sridharan, R. & Smale, S. T. Predominant interaction of both Ikaros and Helios with the NuRD complex in immature thymocytes. *J Biol Chem* **282**, 30227-30238 (2007).
27 Itazaki, H. *et al.* Isolation and Structural Elucidation of New Cyclotetrapeptides, Trapoxin-a and Trapoxin-B, Having Detransformation Activities as Antitumor Agents. *J Antibiot* **43**, 1524-1532 (1990).
28 Kijima, M., Yoshida, M., Sugita, K., Horinouchi, S. & Beppu, T. Trapoxin, an Antitumor Cyclic Tetrapeptide, Is an Irreversible Inhibitor of Mammalian Histone Deacetylase. *Journal of Biological Chemistry* **268**, 22429-22435 (1993).
29 Taunton, J., Hassig, C. A. & Schreiber, S. L. A mammalian histone deacetylase related to the yeast transcriptional regulator Rpd3p. *Science* **272**, 408-411 (1996).
30 Dovey, O. M. *et al.* Histone deacetylase 1 and 2 are essential for normal T-cell development and genomic stability in mice. *Blood* **121**, 1335-1344 (2013).
31 Zhang, Y. *et al.* Analysis of the NuRD subunits reveals a histone deacetylase core complex and a connection with DNA methylation. *Genes Dev* **13**, 1924-1935 (1999).
32 Somoza, J. R. *et al.* Structural snapshots of human HDAC8 provide insights into the class I histone deacetylases. *Structure* **12**, 1325-1334 (2004).
33 Vannini, A. *et al.* Crystal structure of a eukaryotic zinc-dependent histone deacetylase, human HDAC8, complexed with a hydroxamic acid inhibitor. *Proc Natl Acad Sci U S A* **101**, 15064-15069 (2004).
34 Bressi, J. C. *et al.* Exploration of the HDAC2 foot pocket: Synthesis and SAR of substituted N-(2-aminophenyl)benzamides. *Bioorg Med Chem Lett* **20**, 3142-3145 (2010).
35 Lauffer, B. E. *et al.* Histone deacetylase (HDAC) inhibitor kinetic rate constants correlate with cellular histone acetylation but not transcription and cell viability. *J Biol Chem* **288**, 26926-26943 (2013).
36 Aguilera, C. *et al.* c-Jun N-terminal phosphorylation antagonises recruitment of the Mbd3/NuRD repressor complex. *Nature* **469**, 231-235 (2011).
37 Montgomery, R. L. *et al.* Histone deacetylases 1 and 2 redundantly regulate cardiac morphogenesis, growth, and contractility. *Genes Dev* **21**, 1790-1802 (2007).
38 Yamaguchi, J. *et al.* Histone deacetylase inhibitor (SAHA) and repression of EZH2 synergistically inhibit proliferation of gallbladder carcinoma. *Cancer Sci* **101**, 355-362 (2010).
39 Zupkovitz, G. *et al.* Negative and positive regulation of gene expression by mouse histone deacetylase 1. *Mol Cell Biol* **26**, 7913-7928 (2006).
40 Wang, Z. *et al.* Genome-wide mapping of HATs and HDACs reveals distinct functions in active and inactive genes. *Cell* **138**, 1019-1031 (2009).
41 Kidder, B. L. & Palmer, S. HDAC1 regulates pluripotency and lineage specific transcriptional networks in embryonic and trophoblast stem cells. *Nucleic Acids Res* **40**, 2925-2939 (2012).
42 Kurdistani, S. K., Robyr, D., Tavazoie, S. & Grunstein, M. Genome-wide binding map of the histone deacetylase Rpd3 in yeast. *Nat Genet* **31**, 248-254 (2002).
43 Lagger, G. *et al.* Essential function of histone deacetylase 1 in proliferation control and CDK inhibitor repression. *EMBO J* **21**, 2672-2681 (2002).
44 Senese, S. *et al.* Role for histone deacetylase 1 in human tumor cell proliferation. *Mol Cell Biol* **27**, 4784-4795 (2007).
45 Wilting, R. H. *et al.* Overlapping functions of Hdac1 and Hdac2 in cell cycle regulation and haematopoiesis. *EMBO J* **29**, 2586-2597 (2010).
46 Zupkovitz, G. *et al.* The cyclin-dependent kinase inhibitor p21 is a crucial target for histone deacetylase 1 as a regulator of cellular proliferation. *Mol Cell Biol* **30**, 1171-1181 (2010).
47 Marks, P. A. & Xu, W. S. Histone deacetylase inhibitors: Potential in cancer therapy. *J Cell Biochem* **107**, 600-608 (2009).
48 Rosato, R. R., Almenara, J. A. & Grant, S. The histone deacetylase inhibitor MS-275 promotes differentiation or apoptosis in human leukemia cells through a process regulated by generation of reactive oxygen species and induction of p21CIP1/WAF1 1. *Cancer Res* **63**, 3637-3645 (2003).
49 Dovey, O. M., Foster, C. T. & Cowley, S. M. Histone deacetylase 1 (HDAC1), but not HDAC2, controls embryonic stem cell differentiation. *Proc Natl Acad Sci U S A* **107**, 8242-8247 (2010).
50 Dovey, O. M., Foster, C. T. & Cowley, S. M. Emphasizing the positive: A role for histone deacetylases in transcriptional activation. *Cell Cycle* **9**, 2700-2701 (2010).
51 LeBoeuf, M. *et al.* Hdac1 and Hdac2 act redundantly to control p63 and p53 functions in epidermal progenitor cells. *Dev Cell* **19**, 807-818 (2010).

52 Ma, P., Pan, H., Montgomery, R. L., Olson, E. N. & Schultz, R. M. Compensatory functions of histone deacetylase 1 (HDAC1) and HDAC2 regulate transcription and apoptosis during mouse oocyte development. *Proc Natl Acad Sci U S A* **109**, E481-489 (2012).
53 Pencil, S. D., Toh, Y. & Nicolson, G. L. Candidate metastasis-associated genes of the rat 13762NF mammary adenocarcinoma. *Breast Cancer Res Treat* **25**, 165-174 (1993).
54 Bowen, N. J., Fujita, N., Kajita, M. & Wade, P. A. Mi-2/NuRD: multiple complexes for many purposes. *Biochim Biophys Acta* **1677**, 52-57 (2004).
55 Yaguchi, M. *et al.* Identification and characterization of the variants of metastasis-associated protein 1 generated following alternative splicing. *Biochim Biophys Acta* **1732**, 8-14 (2005).
56 Boyer, L. A. *et al.* Essential role for the SANT domain in the functioning of multiple chromatin remodeling enzymes. *Mol Cell* **10**, 935-942 (2002).
57 Yu, J., Li, Y., Ishizuka, T., Guenther, M. G. & Lazar, M. A. A SANT motif in the SMRT corepressor interprets the histone code and promotes histone deacetylation. *EMBO J* **22**, 3403-3410 (2003).
58 Chambers, A. L., Pearl, L. H., Oliver, A. W. & Downs, J. A. The BAH domain of Rsc2 is a histone H3 binding domain. *Nucleic Acids Res* **41**, 9168-9182 (2013).
59 Kuo, A. J. *et al.* The BAH domain of ORC1 links H4K20me2 to DNA replication licensing and Meier-Gorlin syndrome. *Nature* **484**, 115-119 (2012).
60 Mazumdar, A. *et al.* Transcriptional repression of oestrogen receptor by metastasis-associated protein 1 corepressor. *Nat Cell Biol* **3**, 30-37 (2001).
61 Zhang, X. Y. *et al.* Metastasis-associated protein 1 (MTA1) is an essential downstream effector of the c-MYC oncoprotein. *Proc Natl Acad Sci U S A* **102**, 13968-13973 (2005).
62 Yoo, Y. G., Kong, G. & Lee, M. O. Metastasis-associated protein 1 enhances stability of hypoxia-inducible factor-1alpha protein by recruiting histone deacetylase 1. *EMBO J* **25**, 1231-1241 (2006).
63 Cui, Y. *et al.* Metastasis-associated protein 2 is a repressor of estrogen receptor alpha whose overexpression leads to estrogen-independent growth of human breast cancer cells. *Mol Endocrinol* **20**, 2020-2035 (2006).
64 Kumar, R. Another tie that binds the MTA family to breast cancer. *Cell* **113**, 142-143 (2003).
65 Manavathi, B. & Kumar, R. Metastasis tumor antigens, an emerging family of multifaceted master coregulators. *J Biol Chem* **282**, 1529-1533 (2007).
66 Kumar, R. *et al.* A naturally occurring MTA1 variant sequesters oestrogen receptor-alpha in the cytoplasm. *Nature* **418**, 654-657 (2002).
67 Fujita, N. *et al.* MTA3, a Mi-2/NuRD complex subunit, regulates an invasive growth pathway in breast cancer. *Cell* **113**, 207-219 (2003).
68 Zhang, H., Singh, R. R., Talukder, A. H. & Kumar, R. Metastatic tumor antigen 3 is a direct corepressor of the Wnt4 pathway. *Genes Dev* **20**, 2943-2948 (2006).
69 Kleene, R., Classen, B., Zdzieblo, J. & Schrader, M. SH3 binding sites of ZG29p mediate an interaction with amylase and are involved in condensation-sorting in the exocrine rat pancreas. *Biochemistry* **39**, 9893-9900 (2000).
70 Watson, P. J., Fairall, L., Santos, G. M. & Schwabe, J. W. Structure of HDAC3 bound to co-repressor and inositol tetraphosphate. *Nature* **481**, 335-340 (2012).
71 Lewis, J. D. *et al.* Purification, sequence, and cellular localization of a novel chromosomal protein that binds to methylated DNA. *Cell* **69**, 905-914 (1992).
72 Meehan, R. R., Lewis, J. D., McKay, S., Kleiner, E. L. & Bird, A. P. Identification of a mammalian protein that binds specifically to DNA containing methylated CpGs. *Cell* **58**, 499-507 (1989).
73 Nan, X., Meehan, R. R. & Bird, A. Dissection of the methyl-CpG binding domain from the chromosomal protein MeCP2. *Nucleic Acids Res* **21**, 4886-4892 (1993).
74 Nan, X. *et al.* Transcriptional repression by the methyl-CpG-binding protein MeCP2 involves a histone deacetylase complex. *Nature* **393**, 386-389 (1998).
75 Hendrich, B. & Bird, A. Identification and characterization of a family of mammalian methyl-CpG binding proteins. *Mol Cell Biol* **18**, 6538-6547 (1998).
76 Lyko, F., Ramsahoye, B. H. & Jaenisch, R. DNA methylation in Drosophila melanogaster. *Nature* **408**, 538-540 (2000).
77 Marhold, J., Kramer, K., Kremmer, E. & Lyko, F. The Drosophila MBD2/3 protein mediates interactions between the MI-2 chromatin complex and CpT/A-methylated DNA. *Development* **131**, 6033-6039 (2004).
78 Ng, H. H. *et al.* MBD2 is a transcriptional repressor belonging to the MeCP1 histone deacetylase complex. *Nat Genet* **23**, 58-61 (1999).
79 Le Guezennec, X. *et al.* MBD2/NuRD and MBD3/NuRD, two distinct complexes with different biochemical and functional properties. *Mol Cell Biol* **26**, 843-851 (2006).
80 Fraga, M. F. *et al.* The affinity of different MBD proteins for a specific methylated locus depends on their intrinsic binding properties. *Nucleic Acids Res* **31**, 1765-1774 (2003).
81 Hendrich, B. & Tweedie, S. The methyl-CpG binding domain and the evolving role of DNA methylation in animals. *Trends Genet* **19**, 269-277 (2003).
82 Saito, M. & Ishikawa, F. The mCpG-binding domain of human MBD3 does not bind to mCpG but interacts with NuRD/Mi2 components HDAC1 and MTA2. *J Biol Chem* **277**, 35434-35439 (2002).
83 Wade, P. A. *et al.* Mi-2 complex couples DNA methylation to chromatin remodelling and histone deacetylation. *Nat Genet* **23**, 62-66 (1999).
84 Feng, Q. & Zhang, Y. The MeCP1 complex represses transcription through preferential binding, remodeling, and deacetylating methylated nucleosomes. *Genes Dev* **15**, 827-832 (2001).

85 Hendrich, B., Guy, J., Ramsahoye, B., Wilson, V. A. & Bird, A. Closely related proteins MBD2 and MBD3 play distinctive but interacting roles in mouse development. *Genes Dev* **15**, 710-723 (2001).
86 Yildirim, O. *et al.* Mbd3/NURD complex regulates expression of 5-hydroxymethylcytosine marked genes in embryonic stem cells. *Cell* **147**, 1498-1510 (2011).
87 Hashimoto, H. *et al.* Recognition and potential mechanisms for replication and erasure of cytosine hydroxymethylation. *Nucleic Acids Res* **40**, 4841-4849 (2012).
88 Baubec, T., Ivanek, R., Lienert, F. & Schubeler, D. Methylation-dependent and -independent genomic targeting principles of the MBD protein family. *Cell* **153**, 480-492 (2013).
89 Shimbo, T. *et al.* MBD3 localizes at promoters, gene bodies and enhancers of active genes. *PLoS Genet* **9**, e1004028 (2013).
90 Cramer, J. M. *et al.* Probing the dynamic distribution of bound states for methylcytosine-binding domains on DNA. *J Biol Chem* **289**, 1294-1302 (2014).
91 Ho, K. L. *et al.* MeCP2 binding to DNA depends upon hydration at methyl-CpG. *Mol Cell* **29**, 525-531 (2008).
92 Ohki, I. *et al.* Solution structure of the methyl-CpG binding domain of human MBD1 in complex with methylated DNA. *Cell* **105**, 487-497 (2001).
93 Ohki, I., Shimotake, N., Fujita, N., Nakao, M. & Shirakawa, M. Solution structure of the methyl-CpG-binding domain of the methylation-dependent transcriptional repressor MBD1. *EMBO J* **18**, 6653-6661 (1999).
94 Otani, J. *et al.* Structural basis of the versatile DNA recognition ability of the methyl-CpG binding domain of methyl-CpG binding domain protein 4. *J Biol Chem* **288**, 6351-6362 (2013).
95 Scarsdale, J. N., Webb, H. D., Ginder, G. D. & Williams, D. C., Jr. Solution structure and dynamic analysis of chicken MBD2 methyl binding domain bound to a target-methylated DNA sequence. *Nucleic Acids Res* **39**, 6741-6752 (2011).
96 Wakefield, R. I. *et al.* The solution structure of the domain from MeCP2 that binds to methylated DNA. *J Mol Biol* **291**, 1055-1065 (1999).
97 Gnanapragasam, M. N. *et al.* p66Alpha-MBD2 coiled-coil interaction and recruitment of Mi-2 are critical for globin gene silencing by the MBD2-NuRD complex. *Proc Natl Acad Sci U S A* **108**, 7487-7492 (2011).
98 Huntriss, J. *et al.* Expression of mRNAs for DNA methyltransferases and methyl-CpG-binding proteins in the human female germ line, preimplantation embryos, and embryonic stem cells. *Mol Reprod Dev* **67**, 323-336 (2004).
99 Kantor, B., Makedonski, K., Shemer, R. & Razin, A. Expression and localization of components of the histone deacetylases multiprotein repressory complexes in the mouse preimplantation embryo. *Gene Expr Patterns* **3**, 697-702 (2003).
100 Cassel, S., Revel, M. O., Kelche, C. & Zwiller, J. Expression of the methyl-CpG-binding protein MeCP2 in rat brain. An ontogenetic study. *Neurobiol Dis* **15**, 206-211 (2004).
101 Urdinguio, R. G. *et al.* Mecp2-null mice provide new neuronal targets for Rett syndrome. *PLoS One* **3**, e3669 (2008).
102 Auriol, E., Billard, L. M., Magdinier, F. & Dante, R. Specific binding of the methyl binding domain protein 2 at the BRCA1-NBR2 locus. *Nucleic Acids Res* **33**, 4243-4254 (2005).
103 Kaji, K. *et al.* The NuRD component Mbd3 is required for pluripotency of embryonic stem cells. *Nat Cell Biol* **8**, 285-292 (2006).
104 Kaji, K., Nichols, J. & Hendrich, B. Mbd3, a component of the NuRD co-repressor complex, is required for development of pluripotent cells. *Development* **134**, 1123-1132 (2007).
105 Rais, Y. *et al.* Deterministic direct reprogramming of somatic cells to pluripotency. *Nature* **502**, 65-70 (2013).
106 Dos Santos, R. L. *et al.* MBD3/NuRD Facilitates Induction of Pluripotency in a Context-Dependent Manner. *Cell Stem Cell* **15**, 102-110 (2014).
107 Cukier, H. N. *et al.* Novel variants identified in methyl-CpG-binding domain genes in autistic individuals. *Neurogenetics* **11**, 291-303 (2010).
108 Brackertz, M., Boeke, J., Zhang, R. & Renkawitz, R. Two highly related p66 proteins comprise a new family of potent transcriptional repressors interacting with MBD2 and MBD3. *J Biol Chem* **277**, 40958-40966 (2002).
109 Brackertz, M., Gong, Z., Leers, J. & Renkawitz, R. p66alpha and p66beta of the Mi-2/NuRD complex mediate MBD2 and histone interaction. *Nucleic Acids Res* **34**, 397-406 (2006).
110 Nicolas, E. *et al.* RbAp48 belongs to the histone deacetylase complex that associates with the retinoblastoma protein. *J Biol Chem* **275**, 9797-9804 (2000).
111 Qian, Y. W. & Lee, E. Y. Dual retinoblastoma-binding proteins with properties related to a negative regulator of ras in yeast. *J Biol Chem* **270**, 25507-25513 (1995).
112 Qian, Y. W. *et al.* A retinoblastoma-binding protein related to a negative regulator of Ras in yeast. *Nature* **364**, 648-652 (1993).
113 Zhang, Y., Iratni, R., Erdjument-Bromage, H., Tempst, P. & Reinberg, D. Histone deacetylases and SAP18, a novel polypeptide, are components of a human Sin3 complex. *Cell* **89**, 357-364 (1997).
114 Knoepfler, P. S. & Eisenman, R. N. Sin meets NuRD and other tails of repression. *Cell* **99**, 447-450 (1999).
115 Ahringer, J. NuRD and SIN3 histone deacetylase complexes in development. *Trends Genet* **16**, 351-356 (2000).
116 Parthun, M. R. Hat1: the emerging cellular roles of a type B histone acetyltransferase. *Oncogene* **26**, 5319-5328 (2007).
117 Benson, L. J. *et al.* Properties of the type B histone acetyltransferase Hat1: H4 tail interaction, site preference, and involvement in DNA repair. *J Biol Chem* **282**, 836-842 (2007).
118 Hoek, M. & Stillman, B. Chromatin assembly factor 1 is essential and couples chromatin assembly to DNA replication in vivo. *Proc Natl Acad Sci U S A* **100**, 12183-12188 (2003).

119 Korenjak, M. *et al.* Native E2F/RBF complexes contain Myb-interacting proteins and repress transcription of developmentally controlled E2F target genes. *Cell* **119**, 181-193 (2004).
120 Kuzmichev, A., Jenuwein, T., Tempst, P. & Reinberg, D. Different EZH2-containing complexes target methylation of histone H1 or nucleosomal histone H3. *Mol Cell* **14**, 183-193 (2004).
121 Martinez-Balbas, M. A., Tsukiyama, T., Gdula, D. & Wu, C. Drosophila NURF-55, a WD repeat protein involved in histone metabolism. *Proc Natl Acad Sci U S A* **95**, 132-137 (1998).
122 Murzina, N. V. *et al.* Structural basis for the recognition of histone H4 by the histone-chaperone RbAp46. *Structure* **16**, 1077-1085 (2008).
123 Xu, C. & Min, J. Structure and function of WD40 domain proteins. *Protein Cell* **2**, 202-214 (2011).
124 Lejon, S. *et al.* Insights into association of the NuRD complex with FOG-1 from the crystal structure of an RbAp48.FOG-1 complex. *J Biol Chem* **286**, 1196-1203 (2011).
125 Alqarni, S. S. *et al.* Insight into the architecture of the NuRD complex: Structure of the RbAp48-MTA1 subcomplex. *J Biol Chem* (2014).
126 Zhang, W. *et al.* Structural plasticity of histones H3-H4 facilitates their allosteric exchange between RbAp48 and ASF1. *Nat Struct Mol Biol* **20**, 29-35 (2013).
127 Guan, L. S., Li, G. C., Chen, C. C., Liu, L. Q. & Wang, Z. Y. Rb-associated protein 46 (RbAp46) suppresses the tumorigenicity of adenovirus-transformed human embryonic kidney 293 cells. *Int J Cancer* **93**, 333-338 (2001).
128 Thakur, A. *et al.* Aberrant expression of X-linked genes RbAp46, Rsk4, and Cldn2 in breast cancer. *Mol Cancer Res* **5**, 171-181 (2007).
129 Zhang, T. F., Yu, S. Q., Deuel, T. F. & Wang, Z. Y. Constitutive expression of Rb associated protein 46 (RbAp46) reverts transformed phenotypes of breast cancer cells. *Anticancer Res* **23**, 3735-3740 (2003).
130 Creekmore, A. L. *et al.* The role of retinoblastoma-associated proteins 46 and 48 in estrogen receptor alpha mediated gene expression. *Mol Cell Endocrinol* **291**, 79-86 (2008).
131 Kong, L. *et al.* RbAp48 is a critical mediator controlling the transforming activity of human papillomavirus type 16 in cervical cancer. *J Biol Chem* **282**, 26381-26391 (2007).
132 Ginger, M. R., Gonzalez-Rimbau, M. F., Gay, J. P. & Rosen, J. M. Persistent changes in gene expression induced by estrogen and progesterone in the rat mammary gland. *Mol Endocrinol* **15**, 1993-2009 (2001).
133 Zheng, L. *et al.* Radiation-inducible protein RbAp48 contributes to radiosensitivity of cervical cancer cells. *Gynecol Oncol* **130**, 601-608 (2013).
134 Pavlopoulos, E. *et al.* Molecular mechanism for age-related memory loss: the histone-binding protein RbAp48. *Sci Transl Med* **5**, 200ra115 (2013).
135 Tsuji, T. *et al.* Cloning, mapping, expression, function, and mutation analyses of the human ortholog of the hamster putative tumor suppressor gene Doc-1. *J Biol Chem* **273**, 6704-6709 (1998).
136 Yuan, Z., Sotsky Kent, T. & Weber, T. K. Differential expression of DOC-1 in microsatellite-unstable human colorectal cancer. *Oncogene* **22**, 6304-6310 (2003).
137 Spruijt, C. G. *et al.* CDK2AP1/DOC-1 is a bona fide subunit of the Mi-2/NuRD complex. *Mol Biosyst* **6**, 1700-1706 (2010).
138 Shintani, S. *et al.* p12(DOC-1) is a novel cyclin-dependent kinase 2-associated protein. *Mol Cell Biol* **20**, 6300-6307 (2000).

ANNEXES

RÉSUMÉ DE LA THÈSE

Dans un organisme, chaque cellule possède le même matériel génétique. Cependant, l'expression sélective de certains gènes et la répression d'autres a permis la spécialisation des cellules. Avec la mise en place de l'expression différentielle des gènes, les cellules ont pu se différencier et les organismes se développer. Pour une cellule donnée, divers processus normaux et pathologiques peuvent subvenir, tel que des réactions au stress (carences nutritives, hypoxie, manque de facteurs de croissance, etc.), des pathologies associées à la dérégulation de l'expression des gènes (cancers, etc.) ou simplement la progression du cycle cellulaire. Cette expression modulée des gènes est possible grâce au remodelage de la chromatine, un processus associé à l'accessibilité de l'ADN aux facteurs de transcription ou à l'ARN polymérase, en particulier.

En 1942, Conrad Waddington proposa le terme « épigénétique », la branche de la biologie qui étudie les relations de cause à effet entre les gènes et leurs produits. En effet, les gènes et plus généralement la chromatine, sont la cible de modifications covalentes, qui peuvent être reconnues par des effecteurs protéiques, permettant le recrutement d'enzymes et autres partenaires impliqués dans le remodelage de la chromatine. En 1998, plusieurs groupes ont décrit simultanément un complexe possédant une activité de remodelage ATP-dépendant de la chromatine, similaire à ySWI/SNF de *Saccharomyces cerevisiae*, and couplé à une activité histone déacétylase. Ce complexe, baptisé NURD, NRD, complexe Mi-2, et finalement, NuRD, pour « *Nucleosome Remodelling and histone Deacetylation »*, est, à ce jour, le seul complexe connu couplant deux activités indépendantes de remodelage de la chromatine. Une raison possible est que l'activité de remodelage ATP-dépendant de la chromatine est requise pour que les sous-unités déacétylases (HDAC) puissent accéder à leur cible. Cette idée est fondée sur l'observation qu'en absence d'ATP, la déacétylation n'est possible que sur l'octamère d'histones et non sur les nucléosomes. Les sites de liaison aux HDACs pourraient ainsi être protégés par l'ADN et donc inaccessibles. Des expériences menées pour déterminer si l'ATP pouvait stimuler ou non l'activité déacétylase n'ont pas montré d'effet significatif sur des octamères d'histones libres. En revanche, quand des nucléosomes ont été testés, la présence d'ATP a permis une augmentation par deux fois de l'activité déacétylase : sans ATP, 30-35% des histones H4 acétylés étaient déacétylés, alors qu'en présence d'ATP, 60-70 % l'étaient[1].

Le complexe NuRD est hautement conservé chez tous les eucaryotes supérieurs, et est exprimé dans une grande majorité de tissues. Il forme un important assemblage macromoléculaire, constitué de plusieurs sous-unités protéiques ; cependant, différents homologues et isoformes existent pour chaque sous-unité, entraînant une cohorte de complexes NuRD coexistant, en fonction du type cellulaire, tissulaire, et du contexte physiologique ou pathologique. De plus, la stœchiométrie des différentes sous-unités reste question ouverte. Récemment, les développements en spectrométrie de masse quantitative « label-free », appliqués à l'analyse du complexe NuRD, ont suggéré que ce dernier était composé d'une protéine CHD3 ou CHD4 (*Chromodomain, Helicase, DNA binding domain*), d'une HDAC1 ou HDAC2 (*Histone Deacetylase*), trois MTA1/2/3 (*Metastasis Associated*), une MBD3 (*Methylated CpG-Binding*), six RbAp46/48 (*Retinoblastoma Associated*

protein), deux p66α ou p66β et deux DOC-1 (*Deleted in Oral Cancer*)[2]. Ces résultats sont cependant en contradiction avec l'analyse structurale du complexe HDAC1/MTA1 qui montre une dimérisation de MTA1, suggérant la présence de deux MTA1/2/3 et deux HDAC1 ou HDAC2 dans NuRD[3]. Les spécificités de chaque isoforme, ainsi que la mise en commun au sein d'un même complexe de leurs spécificités, à l'instar des deux activités opposées de déacétylation et de remodelage ATP-dépendant, assure que NuRD est un acteur majeur dans différents processus biologiques, comme le développement embryonnaire, la différenciation cellulaire, l'hémato- et lymphopoïèse, l'inhibition de la croissance tumorale, ou la répression générale de la transcription. De plus, il interagit directement avec différents partenaires, comme la déméthylase de lysine LSD1/KDM1A[4], Ikaros, Aiolos, Helios[5-7], *B-cell lymphoma 6* (BCL6)[8,9], le récepteur des œstrogènes α (ERα/NR3A1)[10-12] ou Oct4/Sox2/Klf4/c-Myc (OSKM)[13,14]. Ceci démontre donc le rôle très large et général du complexe NuRD, d'autant plus qu'il s'agit de la forme de la plus abondante de déacétylase chez les mammifères.

Le travail que j'ai mené pendant mon doctorat fait partie d'un projet global et à long terme d'étude du complexe NuRD. En effet, 16 ans après sa découverte, son rôle reste toujours mal compris. Bien que ce complexe joue indubitablement un rôle quasi-ubiquitaire dans nos cellules, nous manquons toujours de données biochimiques, génétiques et structurales pour comprendre le rôle précis d'un complexe NuRD donné, *in vitro*, mais également dans son environnement cellulaire. De nombreuses études se focalisent sur les sous-unités isolées du complexe, et les résultats obtenus sont extrapolés à l'ensemble du complexe.

Dans l'équipe du Dr. Bruno Klaholz, à l'IGBMC, nous étudions de grands complexes impliqués dans la régulation de l'expression des gènes, par une approche de biologie structurale intégrative. Diverses techniques comme la cristallographie aux rayons-X, la cryo-microscopie électronique (cryo-EM) et un large panel d'outils biophysiques nous permettent de décrire avec précision les interactions entre partenaires à l'intérieur d'un complexe, comme les récepteurs nucléaires sur leur ADN-cible, les polyribosomes avec leur ARN messager, ou les complexes d'initiation de la traduction, tous impliqués dans la régulation de l'expression des gènes. Ces mêmes méthodes ont été appliquées pour l'étude de la relation structure-fonction du complexe NuRD, de ses sous-unités et de leurs complexes avec le nucléosome, en lien avec les études fonctionnelles menées par notre collaborateur, le Dr. Ali Hamiche à l'IGBMC.

Ce travail de thèse a donc permis la mise en place de ce nouveau projet ambitieux, avec deux idées principales : l'analyse structurale de sous-unités isolées du complexe, et de sous-complexes stables au sein de ce complexe ou avec le nucléosome ; et l'étude du complexe endogène entier, purifié à partir de cellules humaines.

Afin de reconstituer le complexe NuRD *in vitro*, chacune de ses sous-unités a été clonés dans des vecteurs d'expression baculovirus, pour leur production en cellules d'insecte. Onze vecteurs ont ainsi été construits et testés, avec différentes étiquettes d'affinité. La production de protéine en cellules d'insecte requiert une optimisation fine du protocole de culture : lignée cellulaire, titration du virus, temps de culture, etc. Après optimisation de ces paramètres, mes efforts se sont

principalement concentrés sur trois protéines de NuRD : RbAp46, RbAp48 et MBD3. Cette dernière reste à ce jour très peu étudiée et mal comprise. MBD3 appartient à la famille MBD, fixant les îlots CpG méthylés. Chez les mammifères cependant, cette protéine a perdu sa capacité à fixer l'ADN méthylé et fixe préférentiellement l'ADN non-modifié. Ceci est dû à une mutation ponctuelle (Y34F) qui est apparue avec l'émergence de l'embranchement des mammifères. Les protéines RbAp46 et RbAp48, quant à elles, sont des chaperonnes d'histones, retrouvées dans de nombreux complexes associés à la chromatine comme HAT-1, CAF1, Sin3A, Polycomb, EZH2/EED, NURF et NuRD. Elles sont souvent décrites comme une plateforme structurale stable dans ces complexes, mais leur rôle de chaperonne a été mal étudié jusqu'à présent. Leur étude en complexe avec le nucléosome permettra ainsi de lever le voile sur de nouvelles informations quant à leur fonction première.

Alors que ce projet était en cours, les structures cristallographiques de RbAp46 et RbAp48, en complexe avec un court peptide de l'histone H4, ont été déterminées par le groupe de Ernest Laue[15]. Ces structures suggèrent que les chaperonnes RbAp ne peuvent lier que les histones H4 libres, et non les nucléosomes. En utilisant des nucléosomes humains recombinants, reconstitués au laboratoire, nous avons décidé d'entreprendre des premières études de liaison pour vérifier l'hypothèse posée précédemment. Des expériences de retard sur gel ont été menées avec les deux protéines en complexe avec des nucléosomes reconstitués, et, étonnamment, ont montré un résultat positif dans les deux cas. Ces complexes ont été reconstitués biochimiquement et des essais de cristallisation ont été effectués. Des premiers cristaux de RbAp46-Nucléosome ont été obtenus dans une douzaine de conditions différentes. Leur diffraction a été testée à la source synchrotron SLS (Swiss Light Source, Villigen, Suisse). Cependant, leur forte mosaïcité et leur faible pouvoir diffractant n'ont pas permis de déterminer la structure. Les conditions de cristallisation sont en cours d'optimisation et devraient conduire à une première structure de nucléosome en complexe avec une chaperonne d'histone.

En parallèle, MBD3 a été produit et purifié à partir de cellules d'insectes infectées par baculovirus, avec un bon rendement final, mais sous forme d'agrégats solubles uniquement. L'optimisation de cette purification n'a, cependant, pas été poursuivi, après que des études de spectrométrie de masse aient révélé la présence d'un isoforme court de MBD3, dépourvu de son domaine de liaison à l'ADN. L'ADN complémentaire de l'isoforme long de MBD3 étant indisponible dans les banques d'ADN, un gène synthétique a été conçu et synthétisé, avec une séquence optimisée pour le biais du codon à la fois chez la bactérie et le baculovirus. La production de cette protéine a été mise au point et optimisée en bactérie, mais la perte non négligeable de matériel insoluble nous a mené à développer un nouveau protocole de purification en conditions dénaturantes. Avec un rendement de plusieurs dizaines de milligrammes de protéine pure par litre de culture, de nombreuses conditions de repliement *in vitro* ont pu être testées. Malheureusement, en raison de l'instabilité de cette protéine en absence d'agents dénaturants, aucun échantillon de qualité n'a pu être obtenu.

La production en bioréacteur de 20 ou 100-litres a donc été mise au point afin d'obtenir d'avantage de matériel soluble à traiter en condition native. Une protéine pure a ainsi pu être

obtenue, sous forme d'agrégats solubles, et des expériences de Thermofluor® ont été menées. Ainsi, de nouvelles conditions ont pu être définies, dans lesquelles MBD3 semblait plus stable. Ce nouveau protocole de purification a enfin permis d'obtenir entre 200 et 300 microgrammes de protéine pure par litre de culture. La chromatographie d'exclusion stérique a montré la présence d'une espèce majoritaire de 70 à 80 kilo Daltons, suggérant donc un dimère de MBD3, comme mentionné dans la littérature[16]. Cependant, une analyse par ultracentrifugation analytique a révélé que la protéine purifiée était monomérique et partiellement dépliée. Des essais de fixation au nucléosome ont tout de même été menés et l'optimisation des concentrations salines et du ratio a permis d'observer un résultat positif en retard du gel. Ceci suggère donc que seule la partie C-terminale de MBD3, qui n'est pas impliquée dans la liaison à l'ADN, est dépliée. Par ailleurs, ceci expliquerait, au moins en partie, la grande instabilité de cette protéine face à des changements mineurs de température. Des conditions de travail rigoureuses ont en effet dû être mises en place, notamment la manipulation de l'échantillon en-deçà de 4°C uniquement. Des essais de cristallisation de ce complexe MBD3-Nucléosome ont été menés et des cristaux ont pu être obtenus dans différentes conditions, majoritairement différentes des conditions typiquement observées pour la cristallisation du nucléosome seul. Leur diffraction a été testée aux synchrotrons DLS (Diamond Light Source, Oxfordshire, Angleterre) et SLS, et les résultats obtenus encouragent à poursuivre l'optimisation de cette étape de cristallisation. En particulier, la topologie ainsi que le groupe d'espace et les paramètres de maille se sont avérés être différents de ceux des cristaux de nucléosome uniquement.

En considérant que ce projet a été construit de zéro dans notre équipe, il existe un nombre considérable de perspectives envisageables. Tout d'abord, les cristaux des complexes MBD3-Nucléosome et RbAp-Nucléosome sont en phase d'optimisation et devraient permettre d'obtenir des structures à haute résolution de ces complexes. Dans le cadre du projet MBD3, plusieurs clonages ont été réalisés, pour produire le domaine de liaison à l'ADN uniquement, avec différentes limites. Des premiers essais de purification ont été concluants, et des oligos d'ADN ont été utilisés pour mener des premières expériences de liaison. En s'inspirant des conditions déjà publiées pour les complexes de MBD2 et MBD4, j'ai pu observer un résultat positif de liaison à l'ADN, par expérience de retard sur gel. De plus, des essais de complexe de ce domaine de liaison à l'ADN avec des nucléosomes reconstitués ont également été menés, et malgré les nombreuses optimisations de ratio et de concentration saline, aucune fixation n'a pu être observée. Ceci suggère donc un rôle de la partie C-terminale de MBD3 dans la liaison spécifique aux nucléosomes.

Grace à notre expertise en cryo-EM, la protéine MBD3 entière en complexe avec des nucléosomes reconstitués a pu être congelée sur des grilles de microscopie, et des micrographes ont été collectés. Ce premier jeu de données a permis une reconstruction 3-D à basse résolution. Malgré cette faible résolution (25 Å), cette première densité électronique obtenue montre une forme circulaire et aplatie, dans laquelle la structure cristallographique du nucléosome a pu être superposée ; une densité électronique additionnelle était également visible, dans laquelle la structure cristallographique du domaine MBD de MBD2 a pu être superposée. Cette densité additionnelle montre clairement une interaction avec l'ADN sur le côté du nucléosome, mais semble également s'étendre sur la face du nucléosome. Cependant, la basse résolution de cette

reconstruction ne permet pas de conclure quant à l'interaction de MBD3 avec les histones, bien que cette idée soit en accord avec l'hypothèse que la partie C-terminale de MBD3 serait impliquée dans la reconnaissance du nucléosome. Des études fonctionnelles récentes appuient cette idée, puisque MBD3 s'est montré être impliqué dans l'organisation des nucléosomes à proximité des promoteurs et dans les gènes actifs[17]. De nouvelles acquisitions de données par cryo-EM sont en phase de traitement et devraient conduire à une réponse claire dans le futur.

Enfin, les vecteurs baculovirus qui ont été construits pour toutes les sous-unités de NuRD pourront être utilisés pour procéder à des coinfections et produire *in vivo* des sous-complexes stables de NuRD. Ainsi, la protéine MTA2 semble être le parfait point de départ. En effet, des publications récentes de la structure cristallographique de MTA1 mettent en lumière ses interactions avec HDAC1 et RbAp48. Ces données suggèrent donc l'existence d'un sous-complexe stable incluant HDAC1 ou HDAC2, RbAp46 ou RbAp48 et MTA2.

Pour finir, dans le cadre de notre collaboration avec le Dr. Ali Hamiche sur ce projet, son équipe a mis au point une lignée de cellules HeLa pour la production de complexe NuRD endogène et sa purification. Ceci ouvrira ainsi de nouvelles voies pour l'étude du complexe NuRD entier et permettra de répondre à des questions fondamentales quant à la fonction de ce complexe, comme la localisation des diverses sous-unités à l'intérieur du complexe, et leurs surfaces d'interaction, la stœchiométrie des sous-unités, et, le plus important, la structure globale du complexe.

LE COMPLEXE NURD

En 1998, plusieurs groupes ont simultanément décrit un complexe assurant la fonction de remodelage ATP-dépendant de la chromatine, similaire au complexe ySWI/SNF de *Saccharomyces cerevisiae*, et couplé à une fonction de déacétylation des histones. Ce complexe, baptisé NURD, NRD, complexe Mi-2 et finalement NuRD pour « ***Nu****cleosome* ***R****emodelling and histone* ***D****eacetylation* » est à ce jour le seul complexe connu couplant deux activités indépendantes de remodelage de la chromatine[1,457-459]. Une théorie émise propose que l'activité de remodelage ATP-dépendante de la chromatine soit requise pour que les sous-unités responsables de l'activité déacétylase aient accès à leur cible[462]. Cette hypothèse est soutenue par l'observation qu'en absence d'activité de remodelage ATP-dépendante, la déacétylation ne se fait que sur les octamère d'histone et non sur les nucléosomes entiers. Le site de liaison des déacétylases doit donc probablement se trouver protégé par l'ADN et donc être inaccessible. Des expériences ont été menées en ce sens pour déterminer si la présence d'ATP pouvait stimuler ou non l'activité déacétylase. Aucun effet significatif n'a été détecté sur des octamères d'histones libres. En revanche, quand des histones nucléosomiques ont été testés, l'ATP stimulait l'activité déacétylase deux fois plus : en absence d'ATP, 30-35% des histones H4 acétylés étaient déacétylés ; tandis qu'en présente d'ATP, 60-70% des histones H4 acétylés étaient déacétylés[1]. Ces données montrent que l'activité ATP-dépendante de remodelage pourrait faciliter l'activité déacétylase du complexe, probablement en exposant les substrats aux sous-unités du complexe concernées.

Le complexe NuRD est très conservé chez les eucaryotes supérieurs (plantes et animaux), et est exprimé dans la majorité des tissus. Il consiste en un assemblage de différentes sous-unités protéiques (*figure 33*) ; cependant, différents homologues et isoformes existent pour certaines sous-unités, et il existe donc une multitude de complexes NuRD différents, en fonction du contexte cellulaire, tissulaire, physiologique, pathologique, etc… La stœchiométrie du complexe reste cependant une question ouverte à ce jour. Récemment, la mise au point d'une nouvelle méthode de spectrométrie de masse pour l'analyse quantitative de la stœchiométrie de complexes par une approche relative « label-free », appliquée à NuRD, a suggéré que ce complexe était composé d'une protéine CHD3 ou CHD4, une protéine HDAC1 ou HDAC2, trois protéines MTA1/2/3, une protéine MBD3, six protéines RbAp46/48, deux protéines p66α ou p66β et deux protéines DOC-1[2]. Ces données sont cependant en contradiction avec l'analyse structurale du complexe HDAC1/MTA1 résolu en 2013, qui montre une interface de dimérisation dans le domaine ELM2 de MTA1, suggérant la présence dans NuRD de deux protéines MTA1/2/3 et deux protéines HDAC1 ou HDAC2[3].

Par ailleurs, ce complexe interagit directement avec de très nombreux partenaires, comme LSD1, Ikaros, Aiolos, Helios, BCL-6, ERα/NR3A1, OSKM, etc… Ceci confirmerait donc le rôle très général de NuRD, d'autant qu'il s'agit chez les mammifères de l'une des formes les plus abondantes de déacétylase.

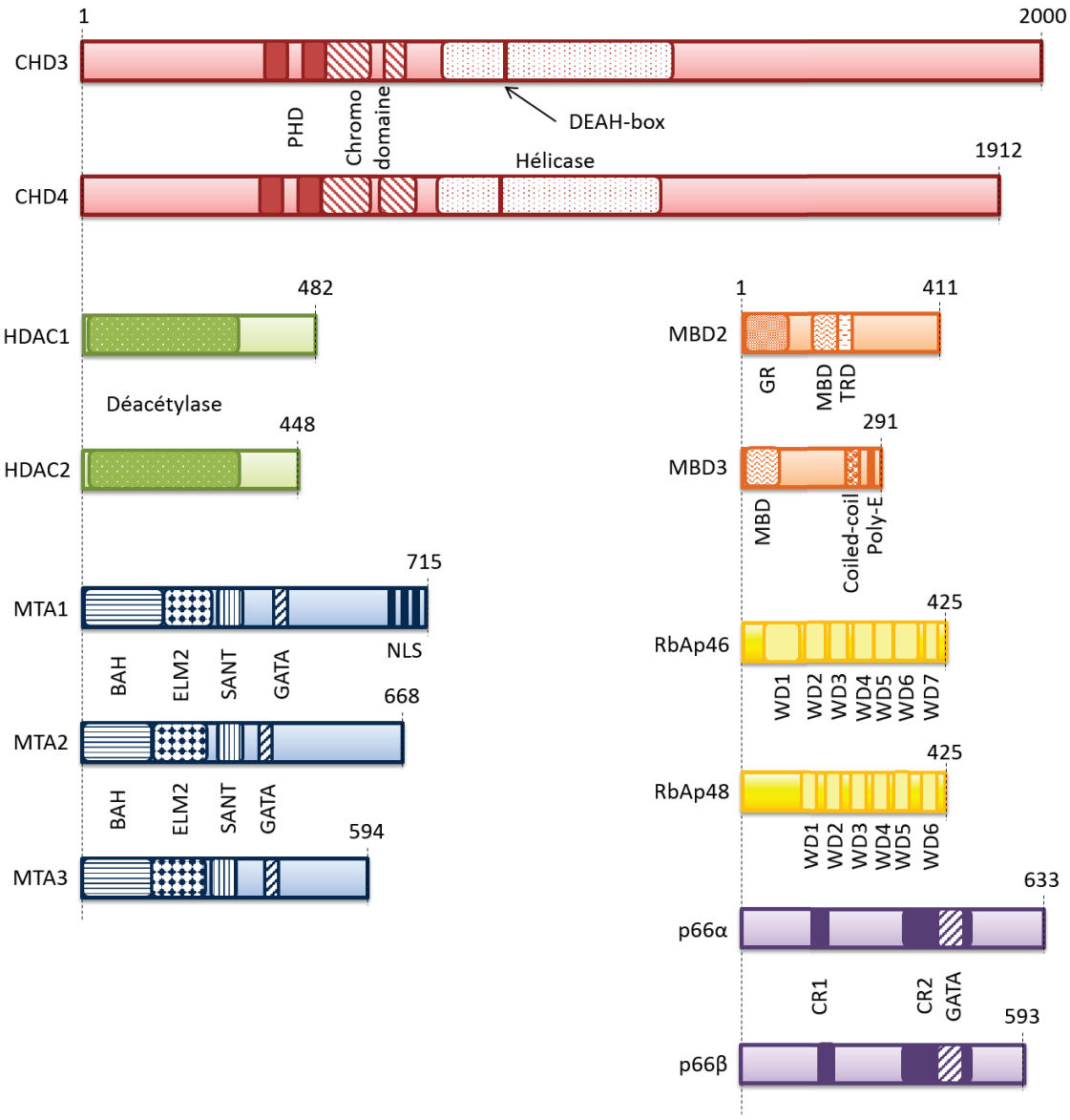


FIGURE 33

Description schématique des différents composants de NuRD

Les différentes sous-unités de NuRD sont représentées schématiquement, avec leurs principaux isoformes. Leur taille est indiquée en nombre d'acides aminés, et les différents domaines caractérisés les constituant sont représentés et annotés.

PHD : Plant HomeoDomain ; Chromodomaine : Chromatin Organization Modifier ; DEAH-box : Asp-Glu-Ala-His-box ; BAH : Bromo Adjacent Homology ; ELM2 : Egl-27 and MTA1 homology ; SANT : Switching-defective protein 3, Adaptor 2, Nuclear receptor co-repressor, Transcription factor IIIB ; GATA : ; NLS : Nuclear localization sequence ; GR : région riche en Glycine-Arginine ; MBD : Methyl-CpG Binding Domain ; TRD : Transcription Repression Domain ; Poly-E : Poly-glutamate ; WD : région riche en Tryptophane-Aspartate ; CR1 : Conserved Region 1 ; CR2 : Conserved Region 2.

(Codes d'accession Uniprot - CHD3 : Q12873 ; CHD4 : Q14839 ; HDAC1 : Q13547 ; HDAC2 : Q92769 ; MTA1 : Q13330 ; MTA2 : O94776 ; MTA3 : Q9BTC8 ; MBD2 : Q9UBB5 ; MBD3 : O95983 ; RbAp46 : Q16576 ; RbAp48 : Q09028 ; p66α : Q86YP4 ; p66β : Q8WXI9 ; DOC-1 : O14519)

A. Détail des composants de NuRD

1) CHD3/4 : le remodelage ATP-dépendant de la chromatine

L'activité ATPase du complexe NuRD réside dans les sous-unités Mi-2. Cette protéine existe sous deux isoformes : Mi-2α (ou CHD3) et Mi-2β (ou CHD4), cette dernière étant la plus abondante dans le complexe NuRD. Il semblerait que les deux isoformes puissent coexister au sein du même complexe, et que trois espèces existent donc : Mi-2α/NuRD, Mi-2α/Mi-2β/NuRD et Mi-2β/NuRD.

La protéine Mi-2 a initialement été identifiée comme un auto-antigène chez les patients atteints de dermatopolymyosite[463,464]. Environ un quart de ces patients est positif aux anticorps anti-Mi-2. Par ailleurs, aucune corrélation entre la présence d'anticorps anti-Mi-2 et le développement de tumeurs n'a pas été prouvée à ce jour ; on peut toutefois noter que dans 20 à 25% des cas, les patients développent un cancer[465] de type ovarien, colorectal, pulmonaire, pancréatique, stomatique ou lymphatique[466].

Les protéines CHD3 et CHD4 sont des grandes ATPases d'environ 220 kilo Daltons, contenant deux doigts PHD conservés, deux chromodomaines en tandem et un domaine hélicase SWI2/SNF-like[467]. Elles font de ce fait partie d'une sous-classe de la famille Snf2[468] et sont très conservées à travers l'ensemble du règne animal et végétal, bien qu'absentes chez les levures. L'activité ATPase des protéines Mi-2 de trois espèces (*Drosophila melanogaster ; Xenopus laevis ; Homo sapiens*) a montré être stimulée par la chromatine mais pas par l'ADN libre ou les histones[458,469,470]. Ceci indique que ces enzymes sont impliquées dans la reconnaissance du nucléosome plutôt que de ses composants isolés. Aucune structure cristallographique de ces protéines, complète ou partielle, n'a été résolue à ce jour ; seules des données en solution par RMN du chromodomaine ainsi que des deux domaines PHD sont disponibles[471,472]. Elles révèlent un mode de liaison bivalent aux histones, reconnaissant deux queues d'histones H3[473] au sein d'un même nucléosome ou de deux nucléosomes adjacents. Les modifications post-traductionnelles de ces queues d'histones régissent l'affinité de liaison de CHD4 : ainsi, la méthylation de H3K9 promeut la fixation de CHD4, alors que la méthylation de H3K4 l'abolit[474].

De plus, la protéine CHD3 existe sous trois isoformes : CHD3.1, CHD3.2 et CHD3.3. L'isoforme 3 résulte de l'utilisation d'un codon d'initiation situé environ 200 nucléotides en amont du codon ATG classique. Les isoformes 1 et 3 possèdent, en position carboxy-terminale, un motif d'interaction SUMO (SIM) qui leur permet d'interagir avec la forme sumoylée de la protéine KAP-1, un composant majeur de l'hétérochromatine. La phosphorylation de KAP-1 par ATM, observée dans le cas de cassures double brin de l'ADN, inhibe l'interaction avec CHD3, résultant en une décompaction de la chromatine[475-477].

Les protéines Mi-2 ont également montré leur rôle crucial dans le développement de différents organismes modèles. Ainsi, chez *Caenorhabditis elegans*, les deux protéines CHD3 et CHD4 sont impliquées dans la voie de signalisation Ras qui permet la différentiation des cellules pendant le développement de l'adulte hermaphrodite[478]. Chez *Arabidopsis thaliana*, PICKLE, l'homologue de CHD3, est impliqué dans la voie de signalisation de l'auxine, nécessaire à la formation des racines

latérales[479]. Chez l'homme, les deux protéines CHD3/4 interagissent directement avec les facteurs de transcription Ikaros, Aïolos et Helios, et permettraient de recruter NuRD spécifiquement au promoteur de certains gènes important dans le développement et la prolifération lymphocytaire, pour en réguler l'expression[5-7]. Parmi les gènes potentiellement ciblés, on peut citer *CD179b* (pour la différenciation des cellules progénitrices pro-B en cellules précurseurs pré-B)*, dntt (pour la recombinaison V-DJ) ou encore CD4* et *CD8a (pour la maturation des thymocytes)*. Ces données suggèrent donc que Mi-2 puisse également avoir un rôle important dans le développement chez les mammifères, mais le manque de modèles génétiques reste aujourd'hui un frein à l'étude de ces protéines.

2) HDAC1/2 : la déacétylation des lysines

Les sous-unités catalytiques assurant la déacétylation sont composées de HDAC1 et HDAC2. Ces protéines d'environ 55 kilo Daltons sont très conservées et présentes de manière ubiquitaire chez tous les eucaryotes. Elles partagent 83% d'identité de séquence, et leur invalidation simultanée dans des lymphocytes T ou les cellules souches embryonnaires provoquent une diminution de moitié de l'activité déacétylase de ces cellules[480]. Elles sont donc les enzymes prédominantes, en termes d'activité, dans les cellules mammifères. Elles font partie de la classe I de déacétylases d'histones et ne présentent aucune préférence pour une séquence d'ADN spécifique. Il a cependant été montré que HDAC1 et HDAC2 pouvaient se lier à des coactivateurs et des corépresseurs pour cibler d'ADN de manière plus spécifique[481].

Les alignements de séquences des HDACs de classe I montrent des différences majeures dans leurs domaines C-terminaux, qui est notamment absent dans HDAC8. Ce domaine est requis dans HDAC1 et HDAC2 pour se lier à des partenaires au sein de complexes protéiques, et est la cible de modifications post-traductionnelles pour réguler leur activité catalytique[482]. Cependant, la première structure cristallographique d'une histone déacétylase, celle de HDAC8 en complexe avec différents inhibiteurs, a permis de poser les bases structurales de la classe I des HDACs[483,484]. Ces protéines sont composées d'un domaine unique α/β, dont un feuillet de huit brins β parallèles, imbriqué à l'intérieur de treize hélices α. L'ensemble de ces feuillets et hélices est relié entre eux par de longues boucles, formant entre-autre le site catalytique de ces enzymes. Ce site actif se présente sous la forme d'un long tunnel, appelé tube lipophile, d'environ 8 Å de profondeur, menant à la machinerie catalytique. Il est occupé par les quatre carbones de la chaine latérale de la lysine acétylée, stabilisé par interactions hydrophobes avec les résidus G151, F152, H180, F208, M274 et F306 (numérotation de HDAC8). Tous ces résidus sont conservés dans tous les HDACs de classe I, à l'exception de M274 qui est une leucine dans les autres HDACs. La fin du tunnel contient un ion zinc, chélaté par cinq liaisons de coordination dans une géométrie bipyramidale trigonale, et stabilisé par l'oxygène carboxylique des résidus D178 et D267, et par l'atome Nδ1 de H180 (*figure 34*). L'oxygène carbonylique du groupement acétyle de l'acétyllysine, ainsi qu'une molécule d'eau, occupent les deux autres sites de coordination. Plus récemment, les structures de HDAC2 en complexe avec des inhibiteurs[485,486], ainsi que celle de HDAC1 en complexe avec les domaines ELM et SANT de MTA1[3], ont montré la même structure globale du cœur catalytique.

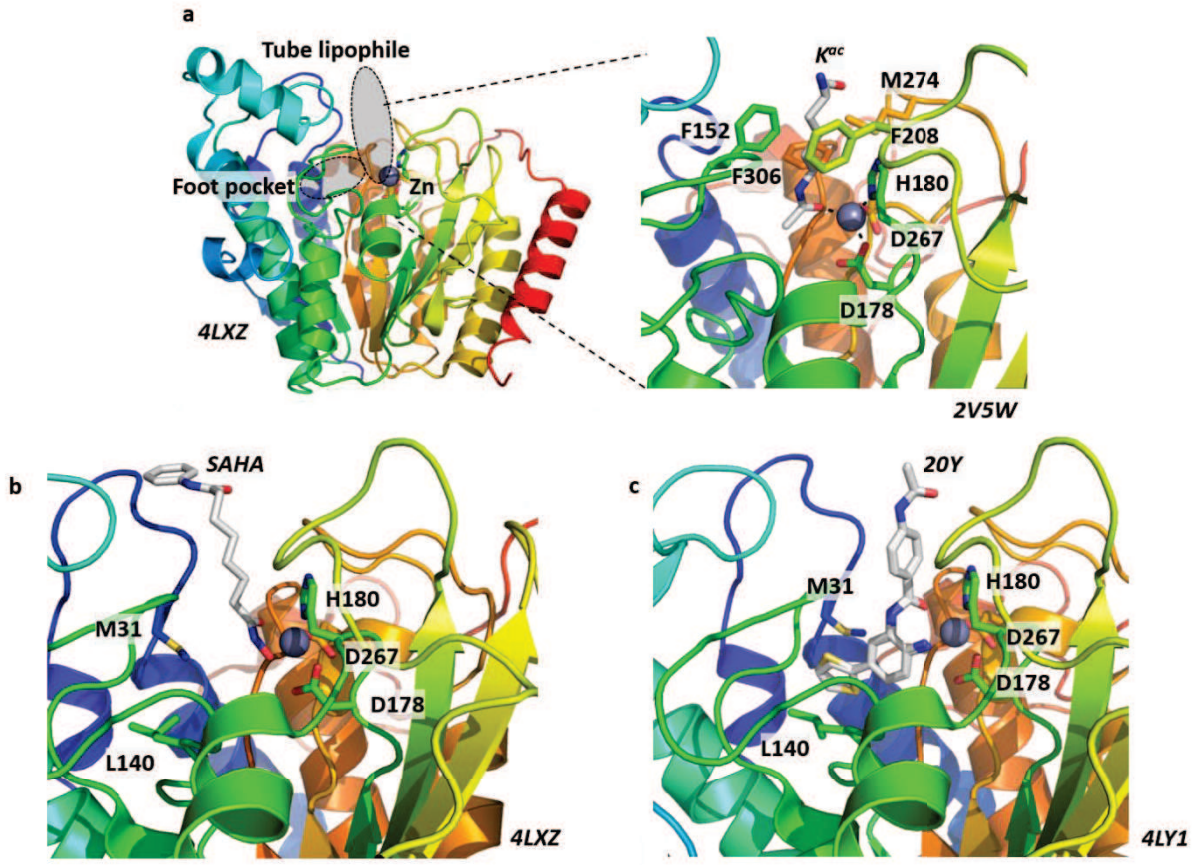


FIGURE 34

Structure cristallographique des HDACs de classe I

a) La structure de HDAC8 en complexe avec une lysine acétylée est ici représentée. Le tube lipophile ainsi que le chausson (*foot pocket*) sont indiqués. Ils contiennent un atome de zinc, représenté par une sphère mauve. Les résidus formant l'interface de liaison avec l'acétyllysine sont représentés (*pdb : 2v5w*)

b) La structure de HDAC2 en complexe avec l'inhibiteur SAHA montre un mode de reconnaissance semblable à celui du substrat naturel, la lysine acétylée (*pdb : 4lxz*)

c) La structure de HDAC2 en complexe avec l'inhibiteur 20Y (4-acetylamino-N-2-amino-5-thiophen-2-ylphenylbenzamide) montre l'ouverture du chausson par rotation des résidus M31 et L140 (*pdb : 4ly1*)

Il a été montré que les inhibiteurs de la classe des hydroxymates, à l'instar de SAHA ou de la trichostatine A, pouvaient interagir globalement de la même manière qu'une acétyllysine, avec une cinétique de liaison rapide et un K_d de l'ordre du nanomolaire dans la plupart des HDACs de classe I et II. Ceci s'explique par l'accès direct du ligand à travers le tube lipophile, lui permettant de chélater l'ion zinc avec son groupement hydroxamique. En revanche, les inhibiteurs de la classe des benzamides, comme l'entinostat ou le mocetinostat, se fixent également dans le tube lipophile, mais leur groupement thiophène se loge dans une poche plus profonde, appelée chausson (*foot pocket* en anglais). Cette poche s'ouvre par rotation et déplacement des résidus M31 et L140 (numérotation de HDAC2). Ils sont conservés dans HDAC1 et HDAC3 mais sont absent dans HDAC8 ainsi que la classe II de HDACs. Enfin, l'amide secondaire central de ces inhibiteurs chélate l'ion zinc, permettant

de bloquer la molécule en place. Ceci s'illustre par une cinétique plus lente en comparaison aux hydroxymates, ainsi qu'une spécificité plus importante pour les HDACs de classe I, notamment HDAC1 et HDAC2.

Les propriétés à la fois génétiques et biochimiques de ces deux enzymes compliquent la compréhension des fonctions spécifiques du complexe NuRD. En effet, bien que la déacétylation soit associée à la répression de l'expression génique, des expériences d'invalidation ont montré que plusieurs gènes étaient réprimés, suggérant que HDAC1 et HDAC2 puissent avoir un rôle dans l'activation de la transcription de certains gènes[487-489]. Des études complémentaires réalisées en traitant des cellules souches embryonnaires à la trichostatine A ont montré à la fois une diminution de l'expression des gènes liés à la pluripotence, et une augmentation de l'expression des gènes de différenciation. Par immunoprécipitation de la chromatine, il a également été montré que ces enzymes étaient présentes à certains sites transcriptionnellement actifs chez l'homme[490], la souris[491] et la levure[492], correspondant à des sites d'hypersensitivité à la DNAse I. En particulier, HDAC1 a été détecté dans les régions promotrices des gènes de pluripotence dans les cellules souches embryonnaires (comme fgf4, mbd3, nanog, oct4, sox2, tbx3 ou zfp42) ainsi que des gènes de lignée trophoblastique dans les cellules souches trophoblastiques (comme bmpr1a, cdkn1c, cdx2, elf5, hand1, msx2 ou tcfap2c)[491].

Un phénomène communément observé lors de l'invalidation de HDAC1 et HDAC2 est la réduction de la prolifération cellulaire[488,493-496]. La perte de ces enzymes cause en effet une surexpression des inhibiteurs de kinases $p21^{WAF1/CIP1}$ [493,496] et $p57^{Kip2}$ [488], bloquant la transition G1/S. Des inhibiteurs de HDACs sont testés dans de nombreux cas de cancers, afin de limiter la progression tumorale[497]. Cependant, la plupart de ces inhibiteurs, à l'instar de SAHA (*SuberoylAnilide Hydroxamic Acid*), la romidepsine ou l'acide valproïque, sont des inhibiteurs à spectre large de toutes les HDACs à activité dépendante du zinc, soit l'ensemble des classes I et II. Des études menées sur la souris ont montré que l'utilisation d'inhibiteurs spécifiques à HDAC1 et HDAC2, comme les benzamides décrites ci-dessus, ont les mêmes effets antiprolifératifs, avec des effets secondaires potentiellement réduits[494,498].

Étant données leur identité biochimique et génétique, il n'est pas étonnant d'observer une redondance des fonctions de HDAC1 et HDAC2. En effet, l'invalidation de l'une ou l'autre de ces enzymes ne provoque aucun phénotype délétère, l'enzyme présente compensant l'absence de son homologue[480,487-489,493,499-502].

3) MTA1/2/3 : le recrutement promoteur-dépendant

Provisoirement baptisées p70, les protéines de la famille MTA (*Metastasis Associated proteins*) ont été caractérisées dans un contexte tumoral. Le premier représentant de cette famille, MTA1, a été isolé après l'observation faite de son expression différentielle dans des modèles cellulaires de croissance métastasique[503]. Mais malgré la surexpression de cette protéine dans les cellules cancéreuses, il aura fallu attendre la découverte du complexe NuRD et l'appartenance des

protéines de la famille MTA à ce complexe pour ouvrir les premières pistes vers la compréhension de leur fonction[1,459].

Des études phylogénétiques ont suggéré que le gène MTA a subit des duplications pour aboutir aux trois loci présents chez les vertébrés (Mta1 sur le chromosome 14q, Mta2 sur le chromosome 11q et Mta3 sur le chromosome 2q)[504], Mta2 étant le plus proche du gène ancestral MTA des invertébrés. Ces trois gènes codent pour les trois protéines MTA1, MTA2 et MTA3, ainsi que pour trois produits d'épissage alternatif : MTA1S, MTA1-ZG29p et MTA3L[505].

Les protéines de la famille MTA ont une masse de 80, 70 et 65 kilo Daltons respectivement, et partagent une homologie de séquence de 68% entre MTA1 et MTA2 ; et 73% entre MTA1 et MTA3. Cette forte homologie est notamment due aux domaines amino-terminaux des protéines MTA, leurs extrémités carboxy-terminales étant divergentes. À l'exception de MTA1-ZG29p, les protéines MTA possèdent toutes des domaines hautement conservés : un domaine BAH (*Bromo Adjacent Homology* ; 70% d'identité entre $MTA1^{BAH}$ et $MTA2^{BAH}$ et 76% d'identité entre $MTA1^{BAH}$ et $MTA3^{BAH}$), un domaine ELM (*Egl-27 and MTA1 homology* ; 76% d'identité entre $MTA1^{ELM}$ et $MTA2^{ELM}$ et 78% d'identité entre $MTA1^{ELM}$ et $MTA3^{ELM}$) et un domaine SANT (87% d'identité entre $MTA1^{SANT}$ et $MTA2^{SANT}$ et 94% d'identité entre $MTA1^{SANT}$ et $MTA3^{SANT}$). Le rôle de ces domaines n'a pas encore été pleinement étudié dans le contexte des protéines MTA, mais l'étude de ces domaines au sein d'autres protéines a permis d'apporter des premiers éléments de réponse. Ainsi, le domaine SANT de Ada2 ou de SMRT est connu pour interagir avec les queues d'histones non-modifiés[433,434,506], et le domaine BAH de Rsc2 est impliqué dans la liaison à l'histone H3[507], tandis que celui de ORC1 reconnait la modification H4K20me2[508]. La protéine MTA1S est produite par épissage alternatif au niveau d'un site cryptique de l'exon 14, résultant en un décalage du cadre de lecture impliquant l'addition de 33 nouveaux acides aminés. L'extrémité carboxy-terminale de cet isoforme est donc unique au sein des protéines MTA, et ne présente aucune homologie de séquence au sein de la GenBank[509]. Pour finir, l'isoforme MTA1-ZG29p est un produit du gène Mta1, ne contenant que les sept derniers exons du gène. Il ne présente donc pas les domaines décrits précédemment, et sa localisation semble restreinte aux granules de zymogènes du pancréas[510].

La régulation de l'expression de ces protéines est encore peu connue à ce jour, bien que des premiers résultats soient disponibles. Ainsi, l'héréguline, un facteur de croissance qui lie les récepteurs transmembranaires HER3 et HER4, est capable d'induire l'expression de MTA1 dans les cancers du sein[11]. Il a aussi été montré que le proto-oncogène c-Myc pouvait se lier au promoteur du gène Mta1 pour activer son expression[511]. Enfin, la protéine MTA1 est surexprimée dans le cas d'hypoxies, et induit la stabilisation du facteur HIF-1 (*Hypoxia Inducible Factor 1*) par recrutement de HDAC et déacétylation, le rendant résistant à la dégradation par le protéasome 26S[512].

En outre, les protéines MTA sont intimement liées au récepteur des œstrogènes (ER)[10,513], dans le cancer du sein et le développement de la glande mammaire[514]. Ainsi, l'isoforme court MTA1S interagit directement avec ER et est impliqué dans sa séquestration dans le cytoplasme[509]. MTA1 bloque également l'activation des gènes par ER en antagonisant l'effet de l'œstradiol[11], tandis que MTA2 peut rendre les cellules mammaires tumorales insensibles aux œstrogènes et au tamoxifène,

en déacétylant le récepteur ER lui-même[10]. Le promoteur du gène Mta3 enfin est directement activé par ER-α[515-517], grâce à la présence d'un demi-élément de réponse ERE et la protéine MTA3 semble être responsable de la répression de l'expression de certains gènes impliqués dans la croissance invasive, tels que SNAI1, Snail[515] ou Wnt4[518]. De ce fait, MTA1 et MTA3 semblent jouer un rôle opposé. Les profils d'expression de ces deux protéines vont dans ce sens : MTA3 est largement exprimée dans les cellules épithéliales saines et son expression décroit au fur et à mesure de la progression cancéreuse, jusqu'à devenir totalement réprimée au stade carcinome ; MTA1 au contraire est peu à peu activée, concomitamment avec la tumorigenèse.

Récemment, une première structure cristallographique a été publiée à 3 Å de résolution. Il s'agit des domaines ELM et SANT de MTA1 (résidus 162-335, soit environ un quart de la protéine), en complexe avec HDAC1[3] (*figure 35*). La structure montre que MTA1 s'enroule autour de HDAC1, avec une surface d'interaction de 5185 Å². Trois régions d'interactions ont pu être décrites : la première correspond au domaine SANT de MTA1, composé de trois hélices α (H1 à H3). L'interface avec HDAC1 forme une poche chargée positivement permettant d'accommoder une molécule d'inositol tétraphosphate (Ins[1,4,5,6]P_4), pour stabiliser cette interaction hautement basique grâce aux résidus K31, R270 et R306, entre-autres. La présence de cette molécule d'inositol tétraphosphate avait déjà été observée dans le complexe HDAC3-SMRT^{SANT}, copurifié à partir de cellules mammifères[520]. Des études complémentaires ont montré que la mutation des résidus de MTA1 impliqués dans la coordination de la molécule d'inositol, résultait en une diminution de l'affinité entre le domaine SANT et HDAC1. Toutefois, MTA1 peut toujours être recruté sur HDAC1 en absence d'inositol grâce à l'interaction de son domaine ELM avec HDAC1, décrit ultérieurement. Les études menées sur le complexe HDAC3-SMRT montrent un lien entre l'âge du complexe, la perte de la molécule d'inositol et la diminution de l'activité déacétylase de HDAC3. De manière intéressante, l'ajout d'Ins[1,4,5,6]P_4 exogène permet de rétablir pleinement l'activité déacétylase. La même observation a pu être faite avec le complexe HDAC1-MTA1, qui présente un K_d d'activation d'environ 5 µM. Ces éléments tendent à confirmer que l'inositol tétraphosphate joue un rôle de régulateur de l'activité des HDACs de classe I.

La seconde région correspond aux trois-quarts carboxy-terminaux du domaine ELM, structuré en quatre hélices α (H1 à H4). Le domaine ELM isolé n'a montré aucune structure secondaire en dichroïsme circulaire, impliquant donc une réorganisation structurale importante lors de la fixation avec HDAC1[3]. Les hélices H1 et H3 sont responsables de cette interaction avec HDAC1, sur une surface de 1278 Å². En parallèle, cette région est également responsable de la dimérisation de deux protéines MTA1, par interaction des hélices H1 et H4, et dans une moindre mesure, H2. Au total, 28 résidus apolaires (quatorze pour chaque monomère) sont impliqués dans cette dimérisation, avec une interface importante de 2332 Å². Ceci laisse à penser que cette interface de dimérisation est physiologiquement justifiée, et que le complexe NuRD possède probablement deux protéines MTA et deux protéines HDAC. Enfin, la troisième région correspond à l'extrémité amino-terminale du domaine ELM. Cette région comprend un motif spécifique conservé (EIRVGxxYQAxI), et est entièrement dépourvue de structure secondaire. Cette longue chaine d'une trentaine de résidus parcourt la surface de HDAC1, le long d'un sillon apolaire.

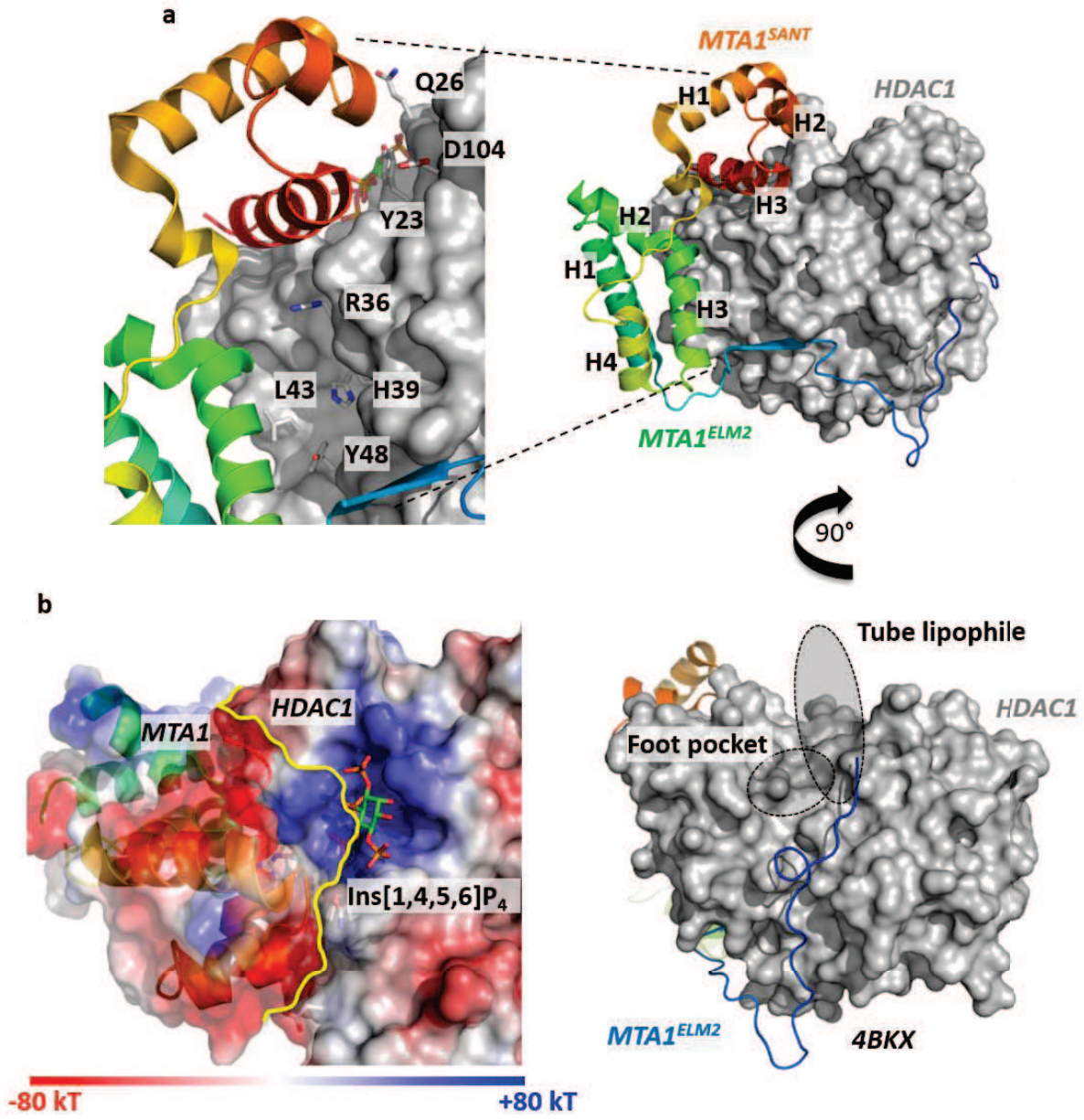


FIGURE 35

Structure cristallographique de HDAC1 en complexe avec MTA1

a) La structure de HDAC1 (en gris) en complexe avec les domaines SANT et ELM2 de MTA1 est ici représentée. On peut observer l'enroulement de MTA1 autour de HDAC1, ainsi que les résidus cruciaux dans l'interface de liaison.

b) Le modèle montre comment une molécule d'inositol tétraphosphate peut s'accommoder dans la poche basique (représentée en bleu), formée à l'interface entre HDAC1 et MTA1. En jaune, l'interface des deux protéines est représentée. Ce modèle a été réalisé en superposant les structures des HDACs issues des complexes HDAC3-SMRT (*pdb : 4a69*) contenant une molécule d'IP4 et HDAC1-MTA1 (*pbd : 4bkx*). Les potentiels électrostatiques de surface sont représentés en rouge (négatif), blanc (neutre) et bleu (positif).

4) MBD2/3 : la liaison à l'ADN

La plus petite sous-unité du complexe NuRD est représentée par une protéine à domaine MBD (*Methyl-CpG Binding Domain-containing protein*)[481,521]. Ce sont ainsi deux membres différents et interchangeables qui peuvent être retrouvés au sein du complexe : MBD2 et MBD3[522]. Ce sont les seules protéines de la famille MBD pour lesquelles des homologues ont été retrouvés chez les invertébrés. Il s'agit donc probablement des descendants directs du gène ancestral de cette famille. Chez les invertébrés, la protéine MBD2/3 est encodée par un gène unique, contrairement aux vertébrés qui présentent un gène mbd2 et un gène mbd3. Aussi, chez les vertébrés, les deux gènes mbd2 et mbd3 ont la même structure génomique, différant uniquement par la taille de leurs introns, et les protéines codées par ces gènes partagent 70% d'identité de séquence. Ces observations corroborent l'idée d'une duplication du gène ancestral mbd2/3 au moment de l'apparition du sous-embranchement des vertébrés.

MBD2 est une protéine capable de lier sélectivement l'ADN méthylé[297], alors que MBD3 a perdu cette capacité chez les mammifères uniquement. En effet, l'apparition de la classe des mammifères s'est accompagnée de deux mutations ponctuelles dans le gène mbd3, entraînant l'incorporation de deux nouveaux acides aminés en position 30 et 34 – une histidine et une phénylalanine, en place d'une lysine et d'une tyrosine, respectivement – abolissant ainsi la sélectivité de cette protéine pour l'ADN méthylé[523-525]. Alors que les premières études sur NuRD, il y a quinze ans, n'attribuaient à MBD2 qu'un rôle transitoire au sein du complexe, servant notamment de recruteur vers l'ADN méthylé avant d'être remplacé par MBD3[481,521], d'autres études ont depuis permis d'asseoir l'existence d'un complexe distinct NuRD/MBD2, biochimiquement et probablement fonctionnellement différent du complexe NuRD/MBD3[522,526]. En ce sens, des expériences de mutagenèse dirigée sur MBD2 n'ont montré des conséquences que limitées au niveau phénotypique, tandis que celles sur MBD3 ont provoqué la mort embryonnaire[319].

Récemment, il a été proposé que MBD3 et dans une moindre mesure, MBD2, étaient capables de lier spécifiquement les îlots CpG hydrométhylés. Notamment, MBD3 semble être colocalisée avec la protéine TET1[527], responsable de l'hydroxyméthylation des 5mC. Des expériences additionnelles ont cependant failli à montrer une interaction entre MBD3 et l'ADN hydroxyméthylé[528]. En revanche, des études ont montré que MBD2 et MBD3 étaient localisés préférentiellement aux sites d'initiations de la transcription, riches en CpG. MBD2 y lie de manière prédominante les CpG méthylés, entraînant une répression de l'expression de ces gènes ; alors que MBD3 se fixe aux CpG non méthylés, et est associée à une activation de la transcription[17,529]. Récemment, des analyses dynamiques par spectroscopie RMN ont suggéré que MBD3 pourrait avoir un rôle de contrebalance, fixant de manière compétitive les îlots CpG non méthylés, empêchant ainsi une répression abusive par MBD2 de ces gènes actifs[530].

À ce jour, plusieurs structures cristallographiques ou en solution de domaines MBDs en complexe avec l'ADN ont été résolues, ce qui a permis de proposer un modèle d'interaction commun à l'ensemble de ces MBPs[310,530-535] (*figure 22*). Notamment, deux structures en solution de MBD2 et une structure en solution de MBD3 ont été résolues, montrant une quasi-identité structurale des

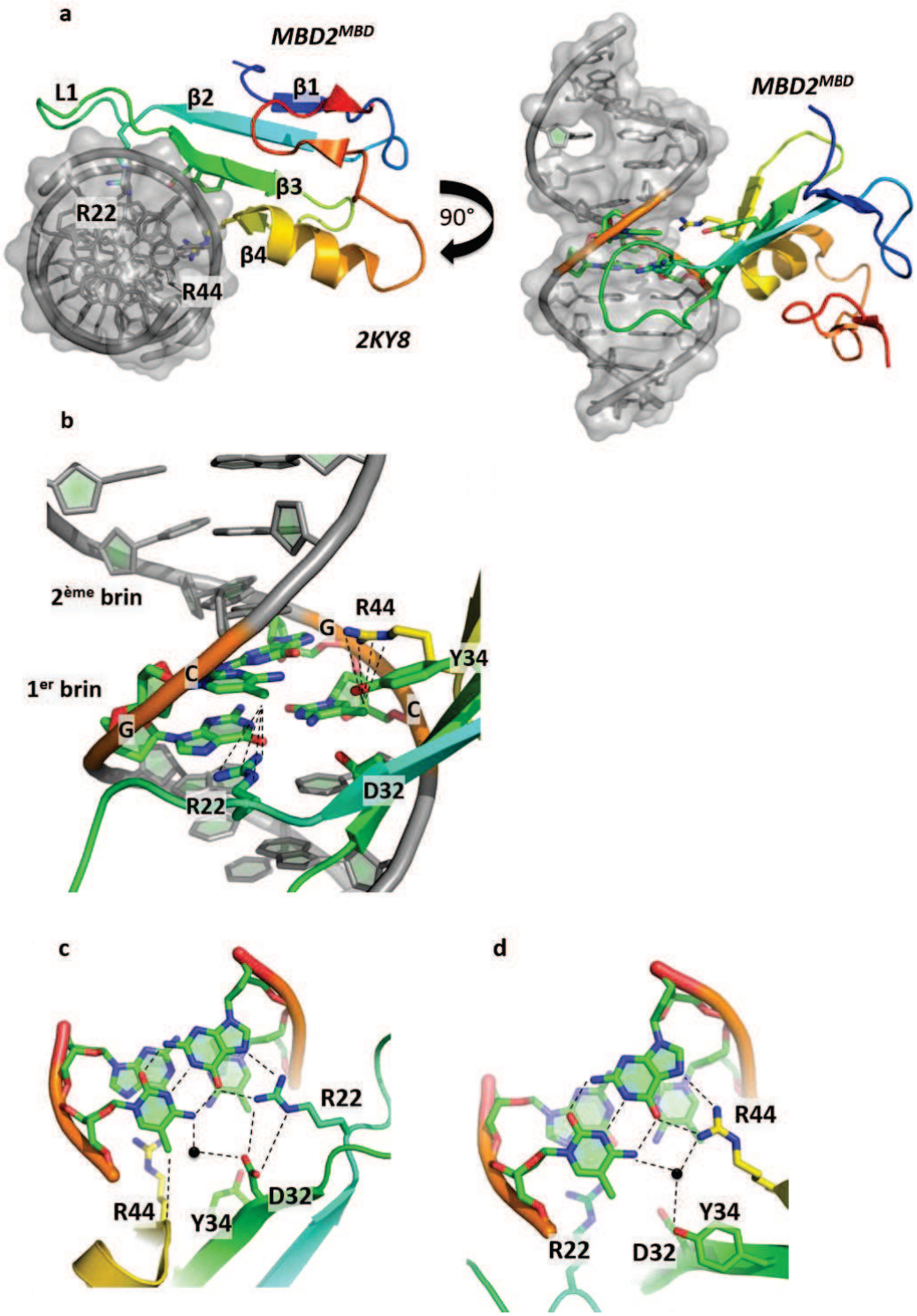


FIGURE 36

La reconnaissance de l'ADN méthylé par MBD2 et xMBD3

a) Le domaine MBD de la protéine MBD2, en complexe avec un ADN contenant un ilot CpG méthylé, est ici représenté.

b) Une vue détaillée de l'interface de liaison montre les résidus cruciaux impliqués dans l'interaction, ainsi que les forces de van der Waals entre les arginines et les cytosines méthylées.

c) Les paires de bases C-G et leurs liaisons hydrogènes spécifiques avec MBD2 sont représentées. Les molécules d'eau impliquées dans des liaisons hydrogènes indirectes sont représentées par des sphères noires (*pdb : 2ky8*)

deux protéines[530,534,536]. Le domaine MBD est caractérisé par un sandwich α/β, composé du côté amino-terminal d'un feuillet de quatre brins β antiparallèles (β_1 : résidus 6-8 dans la séquence de MBD3 ; β_2 : résidus 15-20 ; β_3 : résidus 32-37 ; β_4 : résidus 41-43), et du côté carboxy-terminal, d'une hélice α (résidus 47-53). L'hélice α est maintenue de manière antiparallèle contre le brin β_4 par interactions hydrophobes. De plus, le domaine MBD présente trois boucles L1, L2 et épingle C-terminale. Alors que la boucle L2, qui connecte l'hélice α et l'épingle C-terminale, est bien définie en solution, la longue boucle L1 entre les brins β_2 et β_3, d'une douzaine de résidus, est plus flexible, prérequis nécessaire à la fixation à l'ADN. En effet, sept résidus de cette boucle exhibent des contacts avec l'un des brins de l'ADN, au niveau du sillon majeur. L'autre brin interagit principalement avec des résidus de l'hélice α et de la boucle L2.

La reconnaissance d'un îlot CpG se fait indépendamment pour chaque méthylcytosine, comme l'indique l'absence de symétrie stricte au sein du domaine d'interaction (*figure 36*). Les arginines R22 et R44, conservées dans toutes les MBPs, interagissent avec les guanines d'un îlot CpG. Le groupement guanidinium de R22 fait ainsi une liaison hydrogène avec les atomes O6 et N7 de la guanine sur le premier brin d'ADN, tandis que le groupement guanidinium de R44 montre le même type de liaison sur la guanine du second brin d'ADN. Ces deux arginines se trouvent dans le même plan que les guanines avec lesquelles elles interagissent, stabilisées et maintenues en place par liaison hydrogène directe ou indirecte à travers une molécule d'eau, par les résidus D32 et Y34. De plus, cette orientation plane permet aux deux arginines de ponter par interactions van der Waals faibles les méthylcytosines adjacentes aux guanines avec lesquelles elles interagissent. Enfin, le groupement carbonyle de R44 fait une liaison hydrogène faible de type CO-HC avec le groupement méthyl de la cytosine sur le second brin d'ADN ; tandis que Y34 fait une liaison hydrogène indirecte à travers une molécule d'eau pour reconnaître la cytosine méthylée du premier brin d'ADN. L'intégrité de ces résidus est cruciale pour assurer la liaison à l'ADN méthylé, comme l'ont prouvé plusieurs expériences de mutagenèse. En particulier, Y34 a montré être le résidu-clef dans la reconnaissance de la méthylation de l'ADN, sa mutation en phénylalanine entrainant une perte d'affinité pour les îlots CpG méthylés comme c'est le cas dans la protéine MBD3 de mammifère. En revanche, la protéine MBD3 de *Xenopus laevis* n'ayant pas subit cette mutation évolutive, elle est toujours apte à lier l'ADN méthylé.

Récemment, le rôle central de MBD3 dans la reprogrammation de cellules somatiques et la différentiation cellulaire a été mis en évidence[13,14,537]. Des chercheurs de l'institut Weizmann ont en effet montré que, lors de l'extinction du gène mbd3 par un ARN interférent, le processus de reprogrammation des cellules somatiques murines et humaines en cellules souches pluripotentes était grandement facilité, avec des taux de réussite allant jusqu'à 100% dans certains cas, contre 0,01-5% généralement, et à une vitesse substantiellement améliorée, puisque la reprogrammation ne nécessite qu'une semaine contre un mois avec le protocole habituel. Cette découverte fait suite à la mise en évidence d'une interaction directe entre MBD3 et les protéines OSKM (Oct4, Sox2, Klf4 et Myc), des facteurs de transcription responsables du maintien de la totipotence jusqu'au stade blastocyste. MBD3, en liant ces protéines, pourrait ainsi recruter le complexe NuRD aux gènes de totipotence pour en réprimer l'expression ; ou agir de manière autonome et induire un changement

conformationnel de ces protéines, empêchant leur fixation sur l'ADN. Ces résultats corroborent la létalité embryonnaire observée dans les invalidations de MBD3, l'embryon n'étant alors pas capable de se différencier correctement. Cependant, des résultats contradictoires ont été obtenus à partir de conditions de culture différentes ainsi que d'un choix de cassette de reprogrammation différent. Ceci suggère donc un rôle contexte-dépendant de MBD3 dans le processus de reprogrammation[538].

Dans un autre contexte, l'implication de MeCP2 dans le syndrome de Rett a amené les chercheurs à poser l'hypothèse du rôle des autres MBPs dans les désordres neurologiques. L'ADN de 226 patients autistes et de leurs familles, caucasiens et afro-américains, a ainsi été analysé et des altérations au sein des gènes mbd1-4 ont été détectées chez 198 de ces patients[539]. De manière intéressante, l'une de ces altérations touche l'exon 1 du gène mbd3 : il s'agit d'une mutation ponctuelle d'un nucléotide (G>T en position 1,543,563 dans le locus 19p13.3), entrainant l'incorporation d'un nouvel acide aminé au sein du domaine MBD (R23M). Cette mutation, qui induit une perte de charge négative, a été observée chez deux demi-frères afro-américains, présentant une acquisition tardive et non fonctionnelle du langage. Elle semble héritée de leur grand-mère maternelle, saine, suggérant un effet lié au sexe.

5) RbAp46/48 : la liaison aux histones

Les protéines RbAp46 et RbAp48 (respectivement, Rbbp7 et Rbbp4) ont été initialement identifiées grâce à leurs interactions avec le facteur suppresseur de tumeur Rb[540-542]. Des études ont montré par la suite leur affinité pour les histones, et leur appartenance à différents complexes de déacétylation et de remodelage de la chromatine[504,543-545].

Bien que ces deux protéines partagent 90% d'identité de séquence[541], elles possèdent des activités biochimiques distinctes. Ainsi, RbAp46 s'associe avec d'autres protéines impliquées dans l'acétylation *de novo* des histones H4 synthétisées, sur leurs résidus lysines 5 et 12, notamment HAT1[462,546]. Ce motif d'acétylation est conservé chez tous les eucaryotes, des levures à l'homme[547] ; tandis que RbAp48 est une chaperonne essentielle pour le dépôt d'histones sur l'ADN nouvellement répliqué[548]. Elle est retrouvée notamment dans le complexe d'assemblage CAF-1, aux côtés de p150/CHAF1A et p60/CHAF1B. Cependant, RbAp46 et RbAp48 peuvent être trouvées conjointement au sein de mêmes complexes, par exemple, en association avec HDAC1 et/ou HDAC2 dans les complexes Sin3A ou NuRD, pour promouvoir la répression de gènes, dont ceux régulés par la protéine Rb[481,542,549] ; au sein des complexes de méthylation Polycomb (PRC2 et PRC3), avec la lysine N-méthyltransférase EZH2/EED, pour méthyler H3K27 ou H1K26[550] ; ou encore dans le complexe de remodelage de la chromatine NURF, aux côtés de ISWI (SNF2L chez l'homme), où les protéines RbAp sont appelées NURF55[551].

Les protéines RbAp46 et 48 contiennent un motif de séquence répété WD40. La grande stabilité de ces protéines a permis à ce jour la résolution de sept structures cristallographiques : deux structures de RbAp46 en complexe avec un peptide de l'histone H4, à 2.4 et 2.6 Å de résolution[15] ; une structure de RbAp48 seule à 2.3 Å de résolution[552] ; une structure de RbAp48 en

complexe avec un peptide de FOG-1 à 1.9 Å de résolution[412] ; et trois structures de RbAp48 en complexe avec un peptide de la protéine MTA1 à 2.5 et 2.15 Å de résolution[553]. Sans surprise, les protéines RbAp ont montré une structure similaire aux autres protéines à motif WD40 : un tonneau de feuillets β à sept lames, semblable à un donut, avec une longue hélice α amino-terminale (résidus 9 à 28) qui repose sur la lame 7 du tonneau, et une courte hélice α carboxy-terminale (résidus 405 à 409) qui se place au-dessus et semble allonger l'hélice amino-terminale. Enfin, une particularité est la présence d'une boucle de 17 résidus, chargée négativement, à l'intérieur de la lame 6, et baptisée boucle PP à cause de la présence de deux prolines successives (P362 et P363) (*figure 37*).

Le complexe RbAp46/H4 montre une perte de surface accessible au solvant d'environ 700 Å². Ce peptide H4 correspond aux résidus 25 à 42 de l'isoforme humain, c'est-à-dire la première hélice α et une partie de la queue amino-terminale. Alors que l'interaction dans le cas des autres protéines à motif WD40 décrite jusqu'alors se situait sur une face du tonneau
voire au centre de celui-ci, l'histone H4 se fixe préférentiellement dans une poche sur le côté du tonneau, formée d'une part, par la boucle PP et d'autre part, par la longue hélice amino-terminale. Ainsi, les résidus hydrophobes I34, L37 et A38 de l'hélice α1 de l'histone H4 interagissent avec la région hydrophobe formée par F29, L30, F367, I368 et I407 de RbAp46. Un réseau complexe de ponts salins et de liaisons hydrogène est également observable, entre Q27, K31, R35, R36, R39 et R40 de l'histone H4, et E356, D357, D360, G361, P362, L365, N406, I407 et D410 de RbAp46[15]. L'ensemble des résidus impliqués dans des interactions étant conservé dans RbAp48 et l'homologue de levure p55, le mécanisme d'interaction des trois protéines avec l'histone H4 est similaire. Pour finir, il a été montré qu'afin de promouvoir l'interaction avec RbAp46, l'hélice α1 de H4 doit partiellement s'ouvrir, abolissant de ce fait les interactions avec l'hélice α2 ainsi qu'avec H3, notamment grâce aux résidus I34, L37 et A38. Cette observation pose donc la question de l'interaction des protéines RbAp avec le nucléosome au sein des complexes dont elles font partie, ou d'une flexibilité accrue du nucléosome[15]. Récemment, des expériences de résonnance paramagnétique électronique (PELDOR) ont montré que la protéine RbAp48 pouvait interagir avec un dimère H3-H4 mais pas avec un tétramère $(H3\text{-}H4)_2$[554].

La structure de RbAp48 avec les quinze acides aminés amino-terminaux de FOG-1, un cofacteur impliqué dans la différenciation des érythrocytes et des mégacaryocytes, montre une interface de liaison située sur la face du tonneau, et s'étendant dans la cavité centrale, contrairement au complexe avec H4[412]. Cette interaction est hautement spécifique, puisque huit des treize résidus du peptide FOG-1 sont impliqués dans une liaison hydrogène ou ionique avec RbAp48. Notamment, l'interface est composée pour RbAp48 de nombreux résidus acides (E231, E319, E179, E126, E395, E41), permettant de stabiliser une triade de résidus basiques de FOG-1 (R3, R4 et K5). On peut noter que l'ensemble de ces résidus sont conservés dans RbAp46, suggérant le même mécanisme d'interaction.

Enfin, la structure de RbAp48 avec un court peptide de l'extrémité carboxy-terminale de MTA1 montre une interface de liaison très similaire à celle observée dans le complexe avec H4, suggérant que RbAp46/48 ne peuvent pas interagir simultanément avec MTA1 et l'histone H4[553].

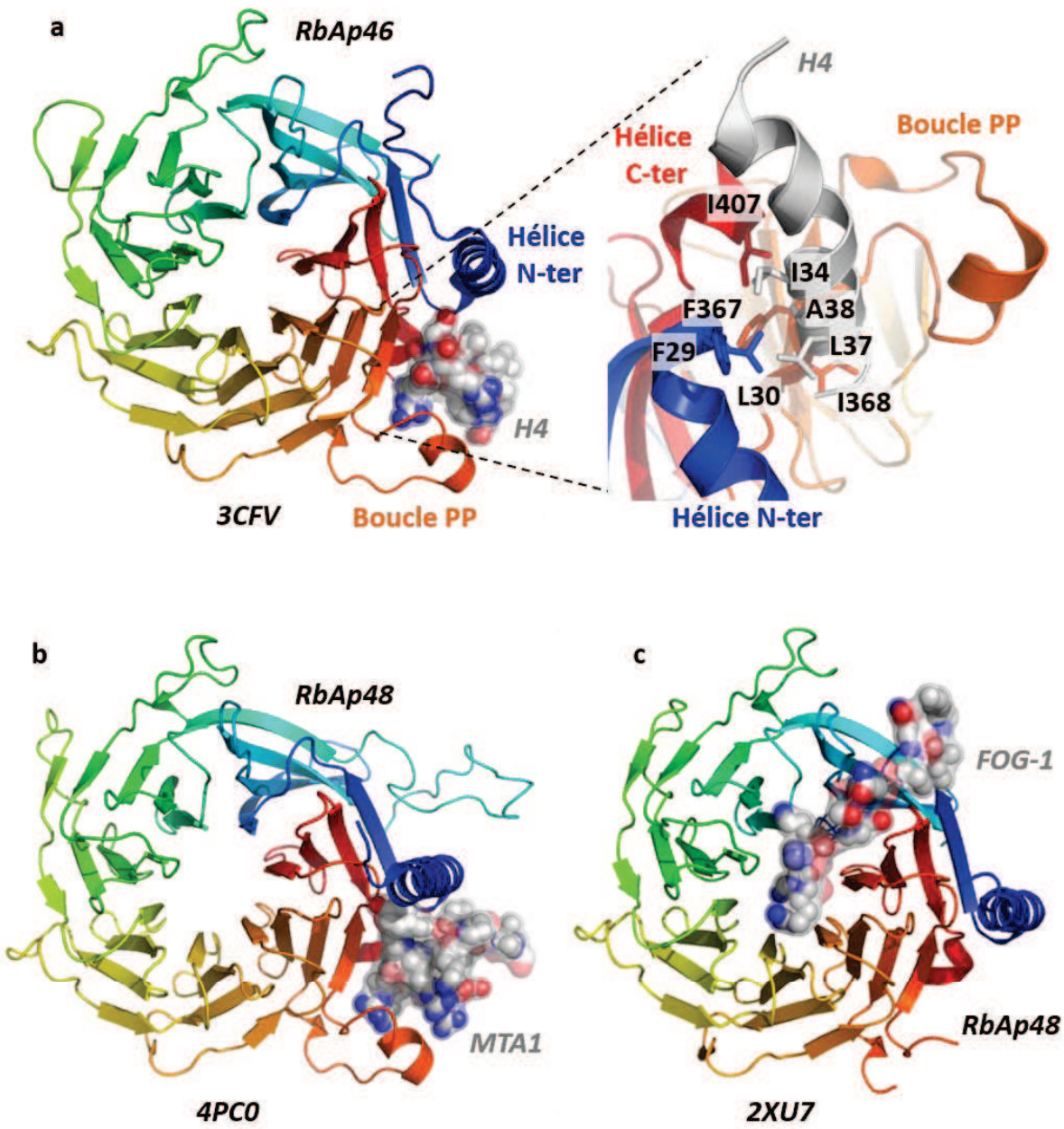


FIGURE 37

Structures cristallographiques de RbAp46 et RbAp48 en complexe

a) La protéine RbAp46 en complexe avec un peptide de l'extrémité amino-terminale de l'histone H4 montre une interface de liaison située sur le côté du tonneau, dans une poche formée par la boucle PP et la longue hélice N-ter. Les résidus hydrophobes cruciaux pour l'interaction sont représentés (*pdb : 3cfv*)

b) La protéine RbAp48 en complexe avec un peptide de l'extrémité carboxy-terminale de la protéine MTA1 montre une interface de liaison similaire à celle de l'histone H4 (*pdb : 4pc0*)

c) La protéine RbAp48 en complexe avec un peptide de la protéine FOG-1 montre une interface de liaison sur la face du tonneau, s'étendant à l'intérieur du canal (*pdb : 2xu7*)

Des dérégulations de ces deux protéines semblent être liées à la tumorigenèse dans plusieurs tissus, notamment mammaire et cervical[555-557]. RbAp46 et RbAp48 ont en effet montré leur interaction directe avec le récepteur nucléaire ERα, et leur influence dans la régulation de l'expression des gènes soumis à ce récepteur[558]. Par exemple, des expériences d'interférence par ARN sur les gènes RbAp46 et RbAp48 dans des cellules MCF-7 ont été menées, et l'activité des gènes Sox9 et cycline-G2, normalement réprimés par ERα en présence d'œstradiol, a été étudié. Il a été conclu que RbAp46 entraînait une activation de la transcription de ces gènes en présence d'œstradiol ; tandis que RbAp48 maintenait leur répression en l'absence du ligand. De plus, l'exposition prolongée de ces cellules cancéreuses à l'œstradiol entrainerait une augmentation d'un facteur deux à trois de l'expression de RbAp46. L'ensemble de ces données laissent donc à penser que RbAp46 pourrait avoir un rôle de modérateur face à l'activité prolongée d'ERα, tandis que RbAp48 assurerait la répression basale de ces gènes en absence de ligand. Pour corroborer cette idée, on peut rappeler que la répression de RbAp48 est impliquée dans la formation de cancers du col de l'utérus[559] ; tandis que l'augmentation de RbAp46 prévient la formation de cancers du sein[555,556,560].

Récemment, RbAp48 s'est montré être une cible thérapeutique de choix dans le traitement du cancer du col de l'utérus[561]. En effet, il a été mis en évidence que l'expression de RbAp48 était favorisée par les irradiations dues à la radiothérapie, et que les lignées SiHa, HeLa et Caski étaient d'autant plus radiosensibles que le taux de RbAp48 était important. Ainsi, la surexpression de RbAp48 induite par adénovirus, couplée à la radiothérapie, a montré un effet antiprolifératif convainquant chez la souris athymique.

Dans un autre contexte, une étude récente, menée sur huit cerveaux humains âgés de 33 à 88 ans, a montré une expression différentielle, réduite avec l'âge, de la protéine RbAp48, spécifiquement dans le gyrus denté, une sous-région de l'hippocampe connue pour sa neurogenèse marquée, et pressentie pour être le siège de la mémoire épisodique[562]. Des études complémentaires menées sur la souris ont confirmé le rôle de RbAp48 dans le processus de mémorisation. Une invalidation de la protéine chez la souris jeune réduit ses performances quant à la mémorisation de nouveaux objets et environnements ; au contraire, la réexpression induite par lentivirus chez les souris âgées a permis une nette amélioration de leurs capacités cognitives. Ces phénomènes semblent étroitement liés au niveau d'acétylation des histones H4 et H2B, par la protéine CBP/p300, un coactivateur favorisant l'expression génétique grâce à son activité acétyltransférase intrinsèque.

6) GATAD2a/b : la potentialisation de la répression

En 2002, des expériences de criblage à double hybride sur la protéine MBD2 ont permis de mettre en évidence une interaction avec deux nouvelles protéines, baptisées p66α et p66β, et plus tard nommées respectivement GATAD2A et GATAD2B (*GATA Zinc Finger Domain Containing 2A/B*)[563]. Ces deux protéines, initialement supposée être deux isoformes d'un même gène, semblent en réalité résulter d'une duplication d'un gène ancestral, survenue lors de l'apparition des mammifères. En effet, un orthologue unique, baptisé p66, est retrouvé chez *Drosophila*

melanogaster, *Caenorhabtidis elegans* et *Xenopus laevis*[521,564]. Le gène humain p66α a ainsi pu être localisé sur le chromosome 19p13.11, tandis que p66β est localisé sur le chromosome 1q23.1.

Ces deux protéines ont montré leurs interactions ainsi que leur colocalisation avec les protéines MBD2 et MBD3[563]. Plus tard, des tests fonctionnels ont permis de mettre en évidence que GATAD2A/B étaient recrutées par l'intermédiaire de deux domaines, sur MBD2 d'une part, via leur domaine CR1 ; mais également sur l'ADN et les histones déacétylés d'autre part, via leur domaine CR2, de type doigt de zinc GATA[564]. De plus, la surexpression des deux protéines p66 induit une augmentation de l'action répressive de MBD2, tandis qu'une invalidation de gène permet une reprise partielle de l'expression des gènes réprimés par MBD2[565].

GATAD2A/B peuvent être la cible de modifications post-traductionnelles, telle que la sumoylation. Ainsi, les résidus K30 et K487 de p66α, et K33 de p66β, sumoylés, favorisent l'interaction de ces protéines avec d'autres partenaires du complexe NuRD, comme HDAC1 ou RbAp46[566].

7) DOC-1 : le suppresseur de tumeur oublié

Récemment, des expériences de copurification effectuées à partir de lignées cellulaires exprimant les protéines MBD2 et MBD3 recombinantes, ont permis de montrer la présence au sein des deux complexes NuRD/MBD2 et NuRD/MBD3 d'une nouvelle sous-unité, une petite protéine de douze kilo Daltons, appelée CDK2AP1 (cdk2-associated protein 1) ou DOC-1 (*Deleted in Oral Cancer-1*)[522]. Comme son nom l'indique, cette protéine, un suppresseur de tumeur putatif interagissant avec CDK2, a montré son inhibition dans le cas de cancers de la bouche et colorectaux[567,568]. Plus tard, des expériences de spectrométrie de masse ont confirmé la présence de DOC-1 au sein du complexe NuRD[2,569].

Le rôle de cette protéine est encore incertain ; cependant, il a été montré que la surexpression de DOC-1 dans des cellules 293T provoquait un arrêt partiel du cycle cellulaire en phase G1/S et un retard de croissance significatif[570]. Ceci est à opposer aux conséquences de la surexpression de MBD2, qui dans ces mêmes cellules 293T, promeut la prolifération cellulaire, suggérant un rôle opposé des deux protéines au sein du complexe.

B. Les fonctions de NuRD : historique et théories actuelles

Lors de la découverte du complexe NuRD à la fin des années 1990, les connaissances de l'époque sur le rôle du remodelage de la chromatine ont amené les chercheurs à définir ce complexe comme un répresseur de la transcription[571]. De plus, on supposait son recrutement uniquement causé par des interactions protéine-ADN (notamment avec MBD2/3) ou des interactions protéine-protéine (avec des répresseurs de la transcription).

Pendant les années qui suivirent la mise en évidence du complexe, et du fait de l'absence de modèles génétiques pour l'étudier, le rôle de NuRD fut décrit sur la base de profils d'expression et de données isolées disponibles pour certaines sous-unités. Ainsi, la sous-unité MTA1 par exemple

était connue pour être surexprimée dans certains cancers du sein. C'est ainsi que le complexe NuRD fut promu au titre de régulateur de la transcription dans les cellules tumorales mammaires[11,572]. Ces études ont également montré que l'expression de MTA1 était amplifiée dans les cellules ERBB2+/HER2+, et que MTA1 interagissait directement avec le récepteur ER pour réprimer la transcription ER-dépendante, notamment du gène BRCA1, provoquant ainsi une croissance cellulaire invasive. C'était la première confirmation évidente que NuRD jouait un rôle de répresseur de la transcription, agissant spécifiquement par recrutement direct[11].

Cette première étude fut suivie de plusieurs autres, décernant chacune à NuRD un rôle de répresseur de la transcription dans le cas d'un processus cellulaire précis : MTA3, responsable de la répression de Snail pour empêcher la croissance invasive dans le cancer du sein[515] ; CHD4, en interagissant avec NAB2, pour coréprimer les transactivateurs EGR (*early growth response*) responsables de la progression du cancer de la prostate[573] ; MBD3, en interagissant avec l'oncoprotéine c-JUN non phosphorylée, pour réprimer son activité transcriptionnelle dans le cas de cancers du côlon[574] ; MTA3, pour réguler le devenir des lymphocytes B, avec interagissant directement avec BCL-6[9] ; CHD4 pour inhiber l'activation du promoteur mb-1 par EBF et Pax5 dans les lymphoblastes[575] ; MTA1 et MTA2, en interagissant avec BCL11b, pour réprimer l'expression du LTR du virus VIH-1 dans les lymphocytes T infectés[576,577] ; MBD2, en partenariat avec GATAD2A, en recrutant directement le complexe NuRD sur des CpG méthylés, pour réprimer l'expression du gène de la globine-β embryonnaire et fœtale[536] ; etc. Parallèlement, des premières expériences biochimiques avec des peptides d'histones ont montré un enrichissement du complexe NuRD au niveau des queues d'histones H3, mais cette interaction est inhibée par la méthylation de H3K4, une marque épigénétique associée à l'activation de la transcription[578-580]. L'ensemble de ces exemples, et d'autres, ont conforté la position prise à cette époque. Le complexe NuRD était un complexe de répression de la transcription, impliqué dans une multitude de voie de signalisation, et couvrant un très large panel de contextes biologiques.

Il faudra attendre 2004, soit six ans après la mise en évidence du complexe NuRD, pour que les premiers modèles génétiques soient mis au point. La protéine CHD4 fut la première cible de ces expériences[581]. Une souris conditionnelle fut mise au point pour étudier les conséquences de l'invalidation de cette protéine dans les lymphocytes T exclusivement, et l'équipe de Georgopoulos fut capable de démontrer l'importance du complexe NuRD à plusieurs étapes du développement de ces cellules. Mais une découverte plus intéressante encore fut celle du rôle de NuRD dans l'activation de l'expression du gène CD4. Il s'agissait alors du premier constat d'un rôle d'activateur pour ce complexe jusqu'alors considéré comme l'un des principaux répresseurs de la transcription. Et quelques années plus tard, le même groupe a montré l'importance de CHD4 et du complexe NuRD dans le maintien de la pluripotence des cellules souches hématopoïétiques. L'invalidation de CHD4 entraînait dans ces cellules une différentiation en érythroblastes, au dépend des lignées lymphoblastes et myéloblastes. Et l'étude des motifs d'expression dans ces cellules, après l'invalidation de CHD4, montra qu'il y avait environ autant de gènes anormalement activés que de gènes anormalement réprimés[582]. Le complexe NuRD avait ainsi définitivement perdu son titre de répresseur général de la transcription.

Ces cinq dernières années, le complexe NuRD a également montré son action dans d'autres processus cellulaires que la régulation de la transcription, tels que la réponse aux dommages de l'ADN ou l'assemblage et le maintien des structures chromosomiques. Par exemple, l'invalidation du complexe NuRD provoque une hypersensibilité des cellules aux dommages de l'ADN et une accumulation de ceux-ci. NuRD semble en effet être recruté aux sites de dommages, soit par interaction de CHD4 avec l'enzyme PARP1[583], soit par interaction de CHD4 avec RNF8, elle-même recruté par MDC1 qui interagit avec H2A.XS139ph[370,584,585].

Dans le cas unique des lymphocytes à prolifération rapide, de fortes accumulations de complexe NuRD ont été mise en évidence, appelées foyers NuRD (*NuRD foci*), localisées dans l'hétérochromatine péricentromérique hyperméthylée des chromosomes 1, 9 et 16 pendant la phase S du cycle cellulaire[586]. Ces foyers NuRD prennent la place des foyers Polycomb PRC1, observés dans tous les autres types cellulaires, suggérant ainsi un rôle unique de NuRD dans la prolifération des lymphocytes. Il est plausible que NuRD soit recruté à ces sites hyperméthylés par la sous-unité MBD2 pour assurer l'assemblage de la chromatine pendant et/ou après la réplication dans ces cellules à prolifération rapide. Aucun mécanisme similaire n'a été décrit à ce jour dans le cas des cellules tumorales à prolifération rapide.

Cette revue des fonctions du complexe NuRD et de ses différentes sous-unités n'avait pas pour but de dresser une liste exhaustive de l'ensemble des processus dans lesquels NuRD est impliqué. Tout d'abord parce que cette liste est en constante évolution. Mais surtout parce que nous ne possédons pas encore le recul nécessaire pour comprendre pleinement l'importance de ce complexe. Ceci s'illustre notamment par l'absence d'unité dans les différentes fonctions énoncées ci-dessus. Certes, il semble évident aujourd'hui que le complexe NuRD joue un rôle quasi-ubiquiste dans nos cellules ; cependant, son étude étant encore relativement récente, nous manquons toujours de données biochimiques, génétiques et structurales pour comprendre le rôle précis d'un complexe NuRD donné, *in vitro* mais également dans son environnement. Nombre d'études ne se focalisent que sur l'une ou l'autre sous-unité, et les résultats obtenus sont extrapolés à l'ensemble du complexe. Par ailleurs, plus de quinze ans après sa découverte, sa composition est toujours incertaine. Quels isoformes composent ce complexe, et dans quelle stœchiométrie ? Ces différents complexes NuRD sont-ils présents de manière systématique dans toutes les cellules, ou est-ce dépendant du type cellulaire, du stade de développement, de la pathologie ? Enfin, l'expression différentielle de certaines sous-unités dans les cellules cancéreuses a poussé à étudier le rôle de NuRD dans un contexte tumoral. Mais qu'en est-il du rôle sauvage de NuRD, dans les cellules saines ?

DISCUSSION ET PERSPECTIVES

Pendant ces quatre années de doctorat, j'ai eu la grande opportunité de mettre en place un nouveau projet, centré sur la régulation de la transcription au niveau chromatinien, qui s'est avéré être original et novateur dans une équipe qui travaillait principalement sur les récepteurs nucléaires et les processus traductionnels.

Le remodelage de la chromatine est un domaine vaste et encore mal compris ; il s'agit pour autant de l'un des principaux processus cellulaires assurant la grande diversité de fonctions observées. Dans notre corps, chaque cellule possède le même matériel génétique, et pourtant, il n'y pas beaucoup de points communs entre une cellule de peau, un neurone, un myocyte ou un lymphocyte. Chacune de ces cellules a acquis des fonctions très différentes, en partant du même mode d'emploi, par une sélection spatio-temporelle stricte de gènes à exprimer ou à réprimer. Le remodelage de la chromatine est impliqué dans ce processus évolutionnaire, qui a permis aux organismes unicellulaires d'évoluer vers des organismes pluricellulaires, avec une complexité accrue, due à la compartimentalisation ainsi qu'à la différenciation cellulaire.

Ce remodelage peut s'effectuer de diverses manières : par exemple, de manière ATP-dépendante, en utilisant l'énergie de l'hydrolyse de l'ATP pour déstabiliser la structure de la chromatine ; ou par des modifications covalentes des composants de la chromatine (histones et/ou ADN), ce qui peut provoquer des changements directs dans le compactage de la chromatine (dans le cas de l'acétylation/déacétylation par exemple), ou indirectement par le recrutement d'enzymes, de facteurs de transcription ou tout autre partenaire pour réguler la transcription.

Le complexe NuRD est l'un des nombreux complexes impliqués dans ce processus de remodelage. Il a été découvert en 1998 et s'est montré, depuis, être un complexe singulier, étant le seul connu à ce jour portant deux activités indépendantes de remodelage de la chromatine, qui plus est *a priori* opposée : une activité de remodelage ATP-dépendant et une activité déacétylase. De plus, il a été montré que ce complexe représente la forme principale de déacétylases d'histones dans nos cellules, et qu'il agit très largement comme un répresseur général de la transcription; cependant, des études récentes semblent montrer un rôle bien plus complexe et multiple. Quoi qu'il en soit, ce complexe n'a pas levé le voile sur toutes ses spécificités à ce jour, et un travail remarquable sera encore nécessaire, en particulier d'un point de vue structural.

C'est dans ce contexte que, en collaboration avec le groupe de Ali Hamiche, nous avons décidé de commencer ce projet ambitieux. J'ai rejoint l'équipe de Bruno Klaholz à cette époque en tant que doctorant, pour mettre en place ce projet avec deux idées majeures. Tout d'abord, étudier chaque sous-unité isolée, ainsi que des sous-complexes stables de NuRD, si possible avec des composants de la chromatine (nucléosomes en particulier). En effet, seulement quelques données structurales étaient disponibles à ce moment et la plupart des sous-unités de NuRD n'avaient pas été étudiées. Beaucoup de travail devait, et doit toujours, être accompli de ce point de vue. Ensuite, l'étude du complexe endogène entier, purifié à partir de cellules humaines. Là encore, malgré les 16

années qui se sont écoulées depuis la découverte de NuRD, des questions fondamentales n'ont toujours pas trouvé de réponse. En particulier, la question de la stœchiométrie de chaque sous-unité est toujours non résolue, sans mentionner la structure globale du complexe qui est totalement inconnue.

* * * *

Mon premier travail a donc été de cloner les gènes de chaque sous-unité du complexe NuRD dans des vecteurs d'expression en vue de leur production recombinante. Dans l'optique d'exprimer plusieurs gènes pour produire des sous-complexes stables, nous avons fait le choix du système d'expression baculovirus. En vue des bons taux de surexpression obtenus pour MBD3, RbAp46 et RbAp48, j'ai sélectionné ces trois protéines comme principal sujet d'étude. En particulier, ces trois sous-unités ont montré une interaction direct avec un composant de la chromatine, que ce soit l'ADN, les histones ou les nucléosomes, que nous reconstituons en routine au laboratoire.

Le cas de MBD3 est particulièrement intéressant, dans le sens où cette protéine est un exemple parfait d'évolution récente. Elle appartient à la famille MBD, qui lie les îlots CpG méthylés, et partage 77 % d'identité de séquence avec son paralogue MBD2. Pourtant, chez les mammifères, l'apparition d'une mutation ponctuelle dans la protéine MBD3 (Y34F) has conduit à une perte de spécificité pour l'ADN méthylé. Cette mutation ne s'observe pas chez les vertébrés inférieurs, suggérant donc un rôle redondant des protéines MBD2 et MBD3 chez ces organismes. Et chez les invertébrés, seule une protéine, nommée MBD2/3, existe. D'un point de vue pathologique, une autre mutation ponctuelle (R23M) a été observée chez deux patients, présentant une acquisition retardée et non-fonctionnelle du langage. J'ai donc commencé à étudier la protéine MBD3 humaine dans le but de comprendre le mécanisme moléculaire de liaison à l'ADN. En plus du gène sauvage de MBD3, les deux mutations ponctuelles décrites précédemment dans le domaine de liaison à l'ADN ont pu être clonées dans des vecteurs d'expression, afin d'aider la compréhension de la spécificité de liaison à l'ADN non-méthylé, et ses implications dans les maladies neurologiques.

L'absence de données biochimiques sur MBD3 a rendu sa purification compliquée. Bien que la surexpression fût importante en cellules d'insecte, sa solubilité limitée ainsi que les problèmes d'agrégation et de précipitation une fois purifiée étaient récurrents. Plusieurs mois d'essais ont ainsi été nécessaires pour obtenir une protéine soluble et en quantité suffisante pour mener une étude structurale. Mais malgré les 3500 conditions de cristallisation testées, aucune d'entre-elles n'a donné de cristaux de protéine. Finalement, une analyse de spectrométrie de masse de cet échantillon a montré que la protéine produite était un isoforme court de MBD3, appelé MBD3Δ, dont il manque le domaine de liaison à l'ADN. J'ai donc décidé d'entreprendre de nouveaux clonages pour produire l'isoforme principal de MBD3.

* * * *

L'ADN complémentaire de MBD3 étant indisponible dans les banques d'ADN, un gène synthétique a été synthétisé et sous-cloné dans différents vecteurs d'expression. Ainsi, un vecteur baculovirus a été choisi pour procéder à des coinfections pour produire des sous-complexes stables. En parallèle, des vecteurs bactériens ont été conçus pour exprimer rapidement et à faible coût de grandes quantités de protéines pour des études isolées. Encore une fois, l'insolubilité et l'agrégation a provoqué un retard dans la mise en place de l'étude structurale. Cependant, des expériences de Thermofluor® ont contribué de manière significative à résoudre les problèmes d'agrégations, et ce qui a pu être caractérisé comme un dimère de MBD3 par chromatographie d'exclusion stérique a été purifié et isolé.

Des essais de caractérisation de ce dimère ont abouti à des résultats ambigus et incertains. Bien que la littérature mentionne la dimérisation de MBD3, aucune donné fiable n'a été publiée. J'ai donc utilisé plusieurs méthodes biophysiques pour résoudre ce problème. Tout d'abord, le profil de chromatographie d'exclusion stérique montre un volume d'élution correspondant à une protéine de 70-80 kilo Daltons, c'est-à-dire, un dimère de MBD3. Des expériences de DLS, à leur tour, ont montré une espèce monodisperse avec une masse moléculaire estimée entre 70 et 80 kilo Daltons. Finalement, l'ultracentrifugation analytique a contredit ces résultats, en montrant une espèce monomérique et partiellement dépliée. Nous en avons ainsi conclu que MBD3 requiert probablement des partenaires pour être stabilisée et se replier correctement, expliquant la très grande sensibilité de la protéine. Ce travail sur une protéine isolé et non-stable a en effet requis des conditions strictes de travail, et l'installation d'une paillasse de travail à 0-2°C.

Des études de liaison aux nucléosomes ont cependant été effectuées sur cette protéine partiellement dépliée, et malgré son instabilité, MBD3 a montré un résultat positif en retard sur gel en complexe avec des nucléosomes reconstitués. Une procédure de repliement a été mise au point pour permettre à MBD3 d'interagir avec les nucléosomes et de se replier correctement. Ceci a été effectué par une lente réduction de la concentration saline par addition d'une solution à bas-sel. Bien que cela ait fonctionné parfaitement pour les études biophysiques comme le retard sur gel, qui ne nécessite que de faibles quantités et concentration de complexe, ce procédé a dû être optimisé davantage pour les essais de cristallisation. Ainsi, en jonglant entre des échantillons hautement concentrés, une lente addition d'une solution bas-sel et des périodes d'incubation, nous avons finalement réussi à obtenir un complexe MBD3-Nucléosome à concentration compatible avec la cristallographie. Plus d'une douzaine de conditions ont donné des cristaux, principalement des aiguilles ou des plaques, mais également certains cristaux en 3-dimensions, morphologiquement compatible avec des tests de diffraction aux rayons-X. Des tests sur plaque à température ambiante ont été effectués au DLS (Oxfordshire, Angleterre), et ont montré un pouvoir diffractant pour certains de ces cristaux, mais malheureusement trop faible et rapidement atténué par les dommages dus aux radiations. Si ces cristaux contenaient effectivement le complexe MBD3-Nucléosome ou le nucléosome seul est donc, à ce jour, toujours incertain. En effet, ce complexe s'est montré instable après quelques jours, ce qui est incompatible avec la croissance cristalline. Cependant, des cristaux ont également pu être obtenus dans une condition optimisée, et après cryoprotection, un jeu de données a pu être collecté au SLS (Villigen, Suisse), avec une diffraction

atteignant 7 Å. Ce premier jeu a permis de résoudre le groupe d'espace ainsi que les paramètres de maille cristalline, qui étaient différents de ceux typiquement observés pour les cristaux de nucléosome seul. Mais la complétude du jeu de données (52.5 %) était trop faible pour pouvoir procéder au remplacement moléculaire et déterminer la structure tridimensionnelle. De futures expériences et optimisations sont donc prévues.

Pour contrecarrer la problématique exposée précédemment, concernant l'instabilité du complexe MBD3-Nucléosome dans le temps, nous avons décidé d'utiliser notre expertise en cryo-EM. Le complexe MBD3-Nucléosome a pu ainsi être congelé directement sur des grilles de microscopie après sa formation, préservant ainsi les interactions observées par retard sur gel. Un premier jeu de données a conduit à une reconstruction tridimensionnelle préliminaire, à basse résolution, environ 25 Å, montrant une forme circulaire et aplatie correspondant au nucléosome, et surmontée d'une densité additionnelle dans laquelle le la structure cristallographique du domaine MBD de MBD2 a pu être superposée. Cette densité additionnelle montre une interaction claire avec le côté du nucléosome, précisément avec l'ADN comme attendu ; mais de manière plus surprenante, cette densité semble également s'étendre sur la face du nucléosome pour atteindre le dimère H3-H4. Les détails de cette interaction nécessitent d'être confirmés. Pour cela, de nouveaux jeux de données ont été collectés sur notre nouveau microscope Titan Krios, après optimisation de la concentration de l'échantillon ainsi que des conditions salines. Remarquablement, les micrographes ont montré une stabilisation du nucléosome en présence de MBD3, alors qu'en absence de ce dernier, les nucléosomes ont tendance à se dissocier dans des conditions salines identiques. Là encore, une structure de résolution moyenne à haute sera nécessaire pour comprendre le mode d'interaction entre MBD3 et le nucléosome. Les nouveaux jeux de données sont en phase de traitement dans cette optique, et devraient nous permettre de décrire avec plus de précision la liaison de MBD3 au nucléosome.

* * * *

Considérant l'état déplié de MBD3, j'ai décidé en parallèle de travailler sur le domaine MBD de liaison à l'ADN, isolé. Après l'avoir cloné et mené des tests d'expression, j'ai pu optimiser un protocole de purification, permettant d'obtenir un rendement important de protéine pure, cependant toujours dépliée en absence d'ADN. En m'inspirant du même protocole que celui utilisé pour la protéine MBD3 complète, j'ai pu étudier les propriétés de liaison de ce domaine MBD, sur les nucléosomes reconstitués d'une part et sur des oligos d'ADN d'autre part. De manière intéressante, le domaine MBD isolé n'a montré aucune affinité pour les nucléosomes, comparé à la protéine entière. Au contraire, les expériences de liaisons sur des oligos d'ADN se sont avérées positifs. La concentration saline, le type de tampon et le pH nécessitent encore d'être optimisés davantage pour caractériser l'interaction du domaine MBD avec des oligos d'ADN non-modifiés et modifié (méthylé, hydroxyméthylé, formylé et carboxylé).

* * * *

Pour résumé mes travaux sur la protéine MBD3, nous pouvons souligner qu'il s'agit d'un travail très ambitieux. Nous avons cependant pu poser certaines hypothèses, basées sur des observations faites et une fine compréhension du comportement de cette protéine. Avec les premiers résultats qui s'annoncent très prochainement, ces questions pourront enfin trouver une réponse. En particulier, nous avons observé que la protéine MBD3 entière se liait aux nucléosomes reconstitués, tandis que le domaine MBD seul ne montrait aucune affinité pour ces mêmes nucléosomes. En considérant que le domaine MBD de MBD3 ne représente que 25 % de la protéine entière, la fonction des 75 % restants de cette dernière est toujours inconnue. Une première reconstruction 3-D de données de cryo-EM suggèrent une interaction entre la partie C-terminale de MBD3 et les histones H3 et/ou H4. Ceci pourrait ainsi expliquer la perte d'affinité du domaine MBD seul pour les nucléosomes. En tant que sous-unité du complexe de remodelage de la chromatine NuRD, MBD3 doit être recrutée directement ou indirectement vers des composants de la chromatine, et en particulier les nucléosomes. Le rôle de MBD3 dans la reconnaissance des nucléosomes doit donc être sérieusement pris en considération, comme un mécanisme permettant de distinguer des îlots CpG présents dans l'ADN nucléosomique ou non. De nouveaux jeux de données sont en phase de traitement pour répondre à cette interrogation.

Parallèlement, les études préliminaires prometteuses du domaine MBD de MBD3, complexé avec des oligos d'ADN dont la séquence provient de régions promotrices, devraient permettre d'obtenir une structure à haute résolution de ce complexe et permettre de décrire les spécificités moléculaires qui régissent la reconnaissance de l'ADN non modifié. De plus, l'étude prochaine des deux mutants F34Y et R23M devrait donner un aperçu détaillé de cette spécificité d'interaction. En particulier, le résidu R23 n'est pas un résidu crucial pour la reconnaissance des îlots CpG mais il est localisé à proximité. L'hypothèse d'une abolition de l'affinité pour les îlots CpG est, cependant, très peu probable, sachant que l'invalidation de MBD3 chez la souris entraîne une létalité embryonnaire précoce. Il est donc possible que cette mutation R23M confère de nouvelles propriétés fonctionnelles à MBD3, sans affecter sa fonction primaire.

* * * *

Mon travail s'est également focalisé sur les protéines RbAp46 et RbAp48. Après la mise au point des vecteurs d'expression, j'ai pu produire et purifier ces deux chaperonnes d'histones à partir de cellules d'insecte. Un protocole facile à mettre en œuvre et commun aux deux protéines a pu être mis en place et des premiers essais de liaison aux nucléosomes recombinants ont été effectués. La grande stabilité de RbAp46 et RbAp48 a permis d'éliminer complétement le sel et des expériences de retard sur gel ont montré la liaison de ces deux protéines aux nucléosomes, avec un ADN naturel α-satellite. Ce résultat inattendu est en contradiction avec deux publications du groupe de Ernest Laue qui suggèrent que ces chaperonnes ne peuvent lier que le dimère H3-H4 au mieux, mais en aucun cas le tétramère, l'octamère ou des nucléosomes complets. Puisque nous avons pu montrer une interaction avec le nucléosome, des premiers essais de cristallisation ont été menés, et des cristaux ont pu être observé dans une douzaine de conditions différentes. Certains d'entre eux

étaient compatibles avec des tests de diffraction aux rayons-X. Cependant, des données de moyenne résolution seulement ont pu être enregistrées, qui n'ont pas permis de déterminer une structure. Ces résultats sont cependant très prometteurs, tout d'abord parce que la cristallisation de nucléosomes en complexe avec des facteurs liés est difficile, mais également parce que ces cristaux ont été obtenus dans des conditions jusqu'alors jamais observées pour la cristallisation de nucléosomes. De nouvelles études de liaison, avec de nouveaux outils biophysiques, seront menées et la cristallisation de ces complexes sera optimisée afin de confirmer expérimentalement la liaison de ces chaperonnes aux nucléosomes.

* * * *

Pour conclure, ce projet ambitieux a consisté en l'étude du complexe NuRD d'un point de vue structural, afin d'obtenir de nouveaux détails sur ses fonctions. Mon travail a permis de mettre en place ce projet à long terme, qui est à présent opérationnel, notamment concernant l'étude de MBD3, RbAp46 et RbAp48 qui devrait révéler de nouveaux détails quant aux fonctions de NuRD. Pour aller plus loin, la stabilisation de la protéine MBD3 entière va requérir la caractérisation de partenaires qui lient cette protéine et induisent son repliement. Cependant, la cristallisation des complexes MBD3-Nucléosome et RbAp46-Nucléosome démontre la faisabilité de l'analyse structurale de nucléosomes reconstitués avec des sous-unités de NuRD.

En parallèle, de nouvelles études seront menées avec les autres sous-unités de NuRD, en particulier CHD4, l'ATPase du complexe, qui reste à ce jour une protéine peu connue. Avec ses 218 kilo Dalton, cette protéine a été relativement bien étudiée du point de vue fonctionnel mais les aspects structuraux restent négligés. Ainsi, à ce jour, seules des structures en solution par RMN des deux domaines PHD et du chromodomaine de cette protéine sont publiées.

La protéine MTA2 est également une protéine majeure, surtout si on considère son implication dans le cancer. Elle est en effet responsable de la désensibilisation des des cellules tumorales de sein contre les œstrogènes et le tamoxifène. Récemment, la publication de deux structures cristallographiques de MTA1 a remis MTA2 sur le devant de la scène. Ces structures correspondent aux domaines ELM2 et SANT de MTA1 (à l'extrémité N-terminale) liés à HDAC1 ; et à un petit peptide de l'extrémité C-terminale de MTA1 lié à RbAp48. Considéré dans leur ensemble, ces deux structures suggèrent ainsi l'existence d'un sous-complexe stable de NuRD, incluant HDAC1, RbAp48 et MTA2. De plus, le domaine BAH de MTA2 interagit avec l'histone H3, et pourrait ainsi être complexé sur des nucléosomes reconstitués. Toutes ces sous-unités sont disponibles pour une expression en système baculovirus, et des coinfections pourraient permettre la formation *in vivo* de ce sous-complexe.

Finalement, la partie la plus excitante de ce projet reste l'étude du complexe NuRD entier. Pour cela, l'équipe de Ali Hamiche a mis au point une lignée stable de cellules HeLa, exprimant une protéine MTA2 avec une étiquette d'affinité. Ainsi, par purification « TAP-tag », le complexe NuRD endogène peut être isolé et étudié. En particulier, des expériences de liaisons transversales des protéines pourraient permettre, par spectrométrie de masse, de déduire le réseau d'interaction entre les différentes sous-unités de NuRD ; et la cryo-EM pourra être utilisée pour révéler la

structure de ce complexe de 1 méga Dalton. Pour cela, des premiers essais ont d'ores et déjà été menés, et le complexe NuRD entier a pu être purifié, au choix, avec ou sans nucléosomes, comme l'ont montré les gels SDS-PAGE. De plus, des études récentes indiquent que le complexe NuRD est très stable, pouvant supporter jusqu'à 1 M de sel sans se dissocier. Mais malgré cela, nos premières observations en cryo-EM ont principalement montré des complexes dissociés et des sous-unités isolées. La préparation de l'échantillon devra donc être optimisée dans un futur proche, pour poursuivre cette étude prometteuse.

REFERENCES

1 Xue, Y. *et al.* NURD, a novel complex with both ATP-dependent chromatin-remodeling and histone deacetylase activities. *Mol Cell* **2**, 851-861 (1998).
2 Smits, A. H., Jansen, P. W., Poser, I., Hyman, A. A. & Vermeulen, M. Stoichiometry of chromatin-associated protein complexes revealed by label-free quantitative mass spectrometry-based proteomics. *Nucleic Acids Res* **41**, e28 (2013).
3 Millard, C. J. *et al.* Class I HDACs share a common mechanism of regulation by inositol phosphates. *Mol Cell* **51**, 57-67 (2013).
4 Wang, Y. *et al.* LSD1 is a subunit of the NuRD complex and targets the metastasis programs in breast cancer. *Cell* **138**, 660-672 (2009).
5 Georgopoulos, K., Winandy, S. & Avitahl, N. The role of the Ikaros gene in lymphocyte development and homeostasis. *Annu Rev Immunol* **15**, 155-176 (1997).
6 Kim, J. *et al.* Ikaros DNA-binding proteins direct formation of chromatin remodeling complexes in lymphocytes. *Immunity* **10**, 345-355 (1999).
7 Sridharan, R. & Smale, S. T. Predominant interaction of both Ikaros and Helios with the NuRD complex in immature thymocytes. *J Biol Chem* **282**, 30227-30238 (2007).
8 Miles, R. R., Crockett, D. K., Lim, M. S. & Elenitoba-Johnson, K. S. Analysis of BCL6-interacting proteins by tandem mass spectrometry. *Mol Cell Proteomics* **4**, 1898-1909 (2005).
9 Fujita, N. *et al.* MTA3 and the Mi-2/NuRD complex regulate cell fate during B lymphocyte differentiation. *Cell* **119**, 75-86 (2004).
10 Cui, Y. *et al.* Metastasis-associated protein 2 is a repressor of estrogen receptor alpha whose overexpression leads to estrogen-independent growth of human breast cancer cells. *Mol Endocrinol* **20**, 2020-2035 (2006).
11 Mazumdar, A. *et al.* Transcriptional repression of oestrogen receptor by metastasis-associated protein 1 corepressor. *Nat Cell Biol* **3**, 30-37 (2001).
12 Okada, M. *et al.* Switching of chromatin-remodelling complexes for oestrogen receptor-alpha. *EMBO Rep* **9**, 563-568 (2008).
13 Kaji, K. *et al.* The NuRD component Mbd3 is required for pluripotency of embryonic stem cells. *Nat Cell Biol* **8**, 285-292 (2006).
14 Kaji, K., Nichols, J. & Hendrich, B. Mbd3, a component of the NuRD co-repressor complex, is required for development of pluripotent cells. *Development* **134**, 1123-1132 (2007).
15 Murzina, N. V. *et al.* Structural basis for the recognition of histone H4 by the histone-chaperone RbAp46. *Structure* **16**, 1077-1085 (2008).
16 Tatematsu, K. I., Yamazaki, T. & Ishikawa, F. MBD2-MBD3 complex binds to hemi-methylated DNA and forms a complex containing DNMT1 at the replication foci in late S phase. *Genes Cells* **5**, 677-688 (2000).
17 Shimbo, T. *et al.* MBD3 localizes at promoters, gene bodies and enhancers of active genes. *PLoS Genet* **9**, e1004028 (2013).
18 Williamson, R. Properties of rapidly labelled deoxyribonucleic acid fragments isolated from the cytoplasm of primary cultures of embryonic mouse liver cells. *J Mol Biol* **51**, 157-168 (1970).
19 Hewish, D. R. & Burgoyne, L. A. Chromatin sub-structure. The digestion of chromatin DNA at regularly spaced sites by a nuclear deoxyribonuclease. *Biochem Biophys Res Commun* **52**, 504-510 (1973).
20 Olins, A. L. & Olins, D. E. Spheroid chromatin units (v bodies). *Science* **183**, 330-332 (1974).
21 Kornberg, R. D. Chromatin structure: a repeating unit of histones and DNA. *Science* **184**, 868-871 (1974).
22 Kornberg, R. D. & Thomas, J. O. Chromatin structure; oligomers of the histones. *Science* **184**, 865-868 (1974).
23 Oudet, P., Gross-Bellard, M. & Chambon, P. Electron microscopic and biochemical evidence that chromatin structure is a repeating unit. *Cell* **4**, 281-300 (1975).

24 Richmond, T. J., Finch, J. T., Rushton, B., Rhodes, D. & Klug, A. Structure of the nucleosome core particle at 7 A resolution. *Nature* **311**, 532-537 (1984).
25 Luger, K., Mader, A. W., Richmond, R. K., Sargent, D. F. & Richmond, T. J. Crystal structure of the nucleosome core particle at 2.8 A resolution. *Nature* **389**, 251-260 (1997).
26 Davey, C. A., Sargent, D. F., Luger, K., Maeder, A. W. & Richmond, T. J. Solvent mediated interactions in the structure of the nucleosome core particle at 1.9 a resolution. *J Mol Biol* **319**, 1097-1113 (2002).
27 Arents, G., Burlingame, R. W., Wang, B. C., Love, W. E. & Moudrianakis, E. N. The nucleosomal core histone octamer at 3.1 A resolution: a tripartite protein assembly and a left-handed superhelix. *Proc Natl Acad Sci U S A* **88**, 10148-10152 (1991).
28 Rizzo, P. J. Those amazing dinoflagellate chromosomes. *Cell Res* **13**, 215-217 (2003).
29 Luger, K. & Richmond, T. J. The histone tails of the nucleosome. *Curr Opin Genet Dev* **8**, 140-146 (1998).
30 Costanzi, C. & Pehrson, J. R. Histone macroH2A1 is concentrated in the inactive X chromosome of female mammals. *Nature* **393**, 599-601 (1998).
31 Chadwick, B. P. & Willard, H. F. A novel chromatin protein, distantly related to histone H2A, is largely excluded from the inactive X chromosome. *J Cell Biol* **152**, 375-384 (2001).
32 Ahmad, K. & Henikoff, S. Histone H3 variants specify modes of chromatin assembly. *Proc Natl Acad Sci U S A* **99 Suppl 4**, 16477-16484 (2002).
33 Marzluff, W. F., Gongidi, P., Woods, K. R., Jin, J. & Maltais, L. J. The human and mouse replication-dependent histone genes. *Genomics* **80**, 487-498 (2002).
34 Marzluff, W. F. Metazoan replication-dependent histone mRNAs: a distinct set of RNA polymerase II transcripts. *Curr Opin Cell Biol* **17**, 274-280 (2005).
35 Dominski, Z. & Marzluff, W. F. Formation of the 3' end of histone mRNA. *Gene* **239**, 1-14 (1999).
36 Whitfield, M. L. *et al.* Stem-loop binding protein, the protein that binds the 3' end of histone mRNA, is cell cycle regulated by both translational and posttranslational mechanisms. *Mol Cell Biol* **20**, 4188-4198 (2000).
37 Gunjan, A. & Verreault, A. A Rad53 kinase-dependent surveillance mechanism that regulates histone protein levels in S. cerevisiae. *Cell* **115**, 537-549 (2003).
38 Pryciak, P. M. & Varmus, H. E. Nucleosomes, DNA-binding proteins, and DNA sequence modulate retroviral integration target site selection. *Cell* **69**, 769-780 (1992).
39 Imbalzano, A. N., Kwon, H., Green, M. R. & Kingston, R. E. Facilitated binding of TATA-binding protein to nucleosomal DNA. *Nature* **370**, 481-485 (1994).
40 Godde, J. S., Nakatani, Y. & Wolffe, A. P. The amino-terminal tails of the core histones and the translational position of the TATA box determine TBP/TFIIA association with nucleosomal DNA. *Nucleic Acids Res* **23**, 4557-4564 (1995).
41 Schieferstein, U. & Thoma, F. Site-specific repair of cyclobutane pyrimidine dimers in a positioned nucleosome by photolyase and T4 endonuclease V in vitro. *EMBO J* **17**, 306-316 (1998).
42 Kimura, H. Histone dynamics in living cells revealed by photobleaching. *DNA Repair (Amst)* **4**, 939-950 (2005).
43 Jamai, A., Imoberdorf, R. M. & Strubin, M. Continuous histone H2B and transcription-dependent histone H3 exchange in yeast cells outside of replication. *Mol Cell* **25**, 345-355 (2007).
44 Bucceri, A., Kapitza, K. & Thoma, F. Rapid accessibility of nucleosomal DNA in yeast on a second time scale. *EMBO J* **25**, 3123-3132 (2006).
45 Luger, K., Rechsteiner, T. J. & Richmond, T. J. Expression and purification of recombinant histones and nucleosome reconstitution. *Methods Mol Biol* **119**, 1-16 (1999).
46 Burton, D. R. *et al.* The interaction of core histones with DNA: equilibrium binding studies. *Nucleic Acids Res* **5**, 3643-3663 (1978).

47 Oohara, I. & Wada, A. Spectroscopic studies on histone-DNA interactions. II. Three transitions in nucleosomes resolved by salt-titration. *J Mol Biol* **196**, 399-411 (1987).
48 Park, Y. J., Dyer, P. N., Tremethick, D. J. & Luger, K. A new fluorescence resonance energy transfer approach demonstrates that the histone variant H2AZ stabilizes the histone octamer within the nucleosome. *J Biol Chem* **279**, 24274-24282 (2004).
49 Hoch, D. A., Stratton, J. J. & Gloss, L. M. Protein-protein Forster resonance energy transfer analysis of nucleosome core particles containing H2A and H2A.Z. *J Mol Biol* **371**, 971-988 (2007).
50 Gansen, A. *et al.* Nucleosome disassembly intermediates characterized by single-molecule FRET. *Proc Natl Acad Sci U S A* **106**, 15308-15313 (2009).
51 Ausio, J., Seger, D. & Eisenberg, H. Nucleosome core particle stability and conformational change. Effect of temperature, particle and NaCl concentrations, and crosslinking of histone H3 sulfhydryl groups. *J Mol Biol* **176**, 77-104 (1984).
52 Brower-Toland, B. D. *et al.* Mechanical disruption of individual nucleosomes reveals a reversible multistage release of DNA. *Proc Natl Acad Sci U S A* **99**, 1960-1965 (2002).
53 Kulaeva, O. I., Gaykalova, D. A. & Studitsky, V. M. Transcription through chromatin by RNA polymerase II: histone displacement and exchange. *Mutat Res* **618**, 116-129 (2007).
54 Hagerman, T. A. *et al.* Chromatin stability at low concentration depends on histone octamer saturation levels. *Biophys J* **96**, 1944-1951 (2009).
55 Buning, R. & van Noort, J. Single-pair FRET experiments on nucleosome conformational dynamics. *Biochimie* **92**, 1729-1740 (2010).
56 Bohm, V. *et al.* Nucleosome accessibility governed by the dimer/tetramer interface. *Nucleic Acids Res* **39**, 3093-3102 (2011).
57 Dong, F. & van Holde, K. E. Nucleosome positioning is determined by the (H3-H4)2 tetramer. *Proc Natl Acad Sci U S A* **88**, 10596-10600 (1991).
58 Kimura, H. & Cook, P. R. Kinetics of core histones in living human cells: little exchange of H3 and H4 and some rapid exchange of H2B. *J Cell Biol* **153**, 1341-1353 (2001).
59 Laskey, R. A., Honda, B. M., Mills, A. D. & Finch, J. T. Nucleosomes are assembled by an acidic protein which binds histones and transfers them to DNA. *Nature* **275**, 416-420 (1978).
60 Williams, S. K. & Tyler, J. K. Transcriptional regulation by chromatin disassembly and reassembly. *Curr Opin Genet Dev* **17**, 88-93 (2007).
61 Fillingham, J. *et al.* Chaperone control of the activity and specificity of the histone H3 acetyltransferase Rtt109. *Mol Cell Biol* **28**, 4342-4353 (2008).
62 Moshkin, Y. M. *et al.* Histone chaperones ASF1 and NAP1 differentially modulate removal of active histone marks by LID-RPD3 complexes during NOTCH silencing. *Mol Cell* **35**, 782-793 (2009).
63 Andrews, A. J., Chen, X., Zevin, A., Stargell, L. A. & Luger, K. The histone chaperone Nap1 promotes nucleosome assembly by eliminating nonnucleosomal histone DNA interactions. *Mol Cell* **37**, 834-842 (2010).
64 Luk, E. *et al.* Chz1, a nuclear chaperone for histone H2AZ. *Mol Cell* **25**, 357-368 (2007).
65 McBryant, S. J. *et al.* Preferential binding of the histone (H3-H4)2 tetramer by NAP1 is mediated by the amino-terminal histone tails. *J Biol Chem* **278**, 44574-44583 (2003).
66 English, C. M., Adkins, M. W., Carson, J. J., Churchill, M. E. & Tyler, J. K. Structural basis for the histone chaperone activity of Asf1. *Cell* **127**, 495-508 (2006).
67 Donham, D. C., 2nd, Scorgie, J. K. & Churchill, M. E. The activity of the histone chaperone yeast Asf1 in the assembly and disassembly of histone H3/H4-DNA complexes. *Nucleic Acids Res* **39**, 5449-5458 (2011).
68 Finch, J. T. & Klug, A. Solenoidal model for superstructure in chromatin. *Proc Natl Acad Sci U S A* **73**, 1897-1901 (1976).
69 Woodcock, C. L., Frado, L. L. & Rattner, J. B. The higher-order structure of chromatin: evidence for a helical ribbon arrangement. *J Cell Biol* **99**, 42-52 (1984).

70 Dorigo, B. *et al.* Nucleosome arrays reveal the two-start organization of the chromatin fiber. *Science* **306**, 1571-1573 (2004).
71 Schalch, T., Duda, S., Sargent, D. F. & Richmond, T. J. X-ray structure of a tetranucleosome and its implications for the chromatin fibre. *Nature* **436**, 138-141 (2005).
72 Robinson, P. J., Fairall, L., Huynh, V. A. & Rhodes, D. EM measurements define the dimensions of the "30-nm" chromatin fiber: evidence for a compact, interdigitated structure. *Proc Natl Acad Sci U S A* **103**, 6506-6511 (2006).
73 Robinson, P. J. & Rhodes, D. Structure of the '30 nm' chromatin fibre: a key role for the linker histone. *Curr Opin Struct Biol* **16**, 336-343 (2006).
74 Routh, A., Sandin, S. & Rhodes, D. Nucleosome repeat length and linker histone stoichiometry determine chromatin fiber structure. *Proc Natl Acad Sci U S A* **105**, 8872-8877 (2008).
75 Grigoryev, S. A., Arya, G., Correll, S., Woodcock, C. L. & Schlick, T. Evidence for heteromorphic chromatin fibers from analysis of nucleosome interactions. *Proc Natl Acad Sci U S A* **106**, 13317-13322 (2009).
76 Colot, V. & Rossignol, J. L. Eukaryotic DNA methylation as an evolutionary device. *Bioessays* **21**, 402-411 (1999).
77 Mayer, W., Niveleau, A., Walter, J., Fundele, R. & Haaf, T. Demethylation of the zygotic paternal genome. *Nature* **403**, 501-502 (2000).
78 Haaf, T. Methylation dynamics in the early mammalian embryo: implications of genome reprogramming defects for development. *Curr Top Microbiol Immunol* **310**, 13-22 (2006).
79 Watanabe, D., Suetake, I., Tada, T. & Tajima, S. Stage- and cell-specific expression of Dnmt3a and Dnmt3b during embryogenesis. *Mech Dev* **118**, 187-190 (2002).
80 Latham, T., Gilbert, N. & Ramsahoye, B. DNA methylation in mouse embryonic stem cells and development. *Cell Tissue Res* **331**, 31-55 (2008).
81 Heard, E. & Disteche, C. M. Dosage compensation in mammals: fine-tuning the expression of the X chromosome. *Genes Dev* **20**, 1848-1867 (2006).
82 Senner, C. E. & Brockdorff, N. Xist gene regulation at the onset of X inactivation. *Curr Opin Genet Dev* **19**, 122-126 (2009).
83 Ideraabdullah, F. Y., Vigneau, S. & Bartolomei, M. S. Genomic imprinting mechanisms in mammals. *Mutat Res* **647**, 77-85 (2008).
84 Reik, W. & Walter, J. Genomic imprinting: parental influence on the genome. *Nat Rev Genet* **2**, 21-32 (2001).
85 Chen, R. Z., Pettersson, U., Beard, C., Jackson-Grusby, L. & Jaenisch, R. DNA hypomethylation leads to elevated mutation rates. *Nature* **395**, 89-93 (1998).
86 Feinberg, A. P., Ohlsson, R. & Henikoff, S. The epigenetic progenitor origin of human cancer. *Nat Rev Genet* **7**, 21-33 (2006).
87 Kafri, T. *et al.* Developmental pattern of gene-specific DNA methylation in the mouse embryo and germ line. *Genes Dev* **6**, 705-714 (1992).
88 Santos, F., Hendrich, B., Reik, W. & Dean, W. Dynamic reprogramming of DNA methylation in the early mouse embryo. *Dev Biol* **241**, 172-182 (2002).
89 Farthing, C. R. *et al.* Global mapping of DNA methylation in mouse promoters reveals epigenetic reprogramming of pluripotency genes. *PLoS Genet* **4**, e1000116 (2008).
90 Kim, J. B. *et al.* Direct reprogramming of human neural stem cells by OCT4. *Nature* **461**, 649-643 (2009).
91 Bhutani, N. *et al.* Reprogramming towards pluripotency requires AID-dependent DNA demethylation. *Nature* **463**, 1042-1047 (2010).
92 Shen, L. *et al.* Genome-wide profiling of DNA methylation reveals a class of normally methylated CpG island promoters. *PLoS Genet* **3**, 2023-2036 (2007).
93 Illingworth, R. *et al.* A novel CpG island set identifies tissue-specific methylation at developmental gene loci. *PLoS Biol* **6**, e22 (2008).

94 Mohn, F. *et al.* Lineage-specific polycomb targets and de novo DNA methylation define restriction and potential of neuronal progenitors. *Mol Cell* **30**, 755-766 (2008).
95 Smiraglia, D. J. *et al.* Excessive CpG island hypermethylation in cancer cell lines versus primary human malignancies. *Hum Mol Genet* **10**, 1413-1419 (2001).
96 Narayan, A. *et al.* Hypomethylation of pericentromeric DNA in breast adenocarcinomas. *Int J Cancer* **77**, 833-838 (1998).
97 Rodriguez, J. *et al.* Genome-wide tracking of unmethylated DNA Alu repeats in normal and cancer cells. *Nucleic Acids Res* **36**, 770-784 (2008).
98 Badal, V. *et al.* CpG methylation of human papillomavirus type 16 DNA in cervical cancer cell lines and in clinical specimens: genomic hypomethylation correlates with carcinogenic progression. *J Virol* **77**, 6227-6234 (2003).
99 Schulz, W. A. *et al.* Genomewide DNA hypomethylation is associated with alterations on chromosome 8 in prostate carcinoma. *Genes Chromosomes Cancer* **35**, 58-65 (2002).
100 Lengauer, C., Kinzler, K. W. & Vogelstein, B. DNA methylation and genetic instability in colorectal cancer cells. *Proc Natl Acad Sci U S A* **94**, 2545-2550 (1997).
101 Gaudet, F. *et al.* Induction of tumors in mice by genomic hypomethylation. *Science* **300**, 489-492 (2003).
102 Eden, A., Gaudet, F., Waghmare, A. & Jaenisch, R. Chromosomal instability and tumors promoted by DNA hypomethylation. *Science* **300**, 455 (2003).
103 Chen, W. Y. *et al.* Heterozygous disruption of Hic1 predisposes mice to a gender-dependent spectrum of malignant tumors. *Nat Genet* **33**, 197-202 (2003).
104 Yoshiura, K. *et al.* Silencing of the E-cadherin invasion-suppressor gene by CpG methylation in human carcinomas. *Proc Natl Acad Sci U S A* **92**, 7416-7419 (1995).
105 Issa, J. P. CpG island methylator phenotype in cancer. *Nat Rev Cancer* **4**, 988-993 (2004).
106 Ruas, M. & Peters, G. The p16INK4a/CDKN2A tumor suppressor and its relatives. *Biochim Biophys Acta* **1378**, F115-177 (1998).
107 Esteller, M. *et al.* Hypermethylation-associated inactivation of p14(ARF) is independent of p16(INK4a) methylation and p53 mutational status. *Cancer Res* **60**, 129-133 (2000).
108 Masson, D. *et al.* Loss of expression of TIMP3 in clear cell renal cell carcinoma. *Eur J Cancer* **46**, 1430-1437 (2010).
109 Rhee, I. *et al.* DNMT1 and DNMT3b cooperate to silence genes in human cancer cells. *Nature* **416**, 552-556 (2002).
110 Rohrs, S. *et al.* Hypomethylation and expression of BEX2, IGSF4 and TIMP3 indicative of MLL translocations in acute myeloid leukemia. *Mol Cancer* **8**, 86 (2009).
111 Chen, T. & Li, E. Structure and function of eukaryotic DNA methyltransferases. *Curr Top Dev Biol* **60**, 55-89 (2004).
112 Bestor, T. H. Cloning of a mammalian DNA methyltransferase. *Gene* **74**, 9-12 (1988).
113 Bestor, T. Structure of mammalian DNA methyltransferase as deduced from the inferred amino acid sequence and direct studies of the protein. *Biochem Soc Trans* **16**, 944-947 (1988).
114 Bestor, T., Laudano, A., Mattaliano, R. & Ingram, V. Cloning and sequencing of a cDNA encoding DNA methyltransferase of mouse cells. The carboxyl-terminal domain of the mammalian enzymes is related to bacterial restriction methyltransferases. *J Mol Biol* **203**, 971-983 (1988).
115 Yen, R. W. *et al.* Isolation and characterization of the cDNA encoding human DNA methyltransferase. *Nucleic Acids Res* **20**, 2287-2291 (1992).
116 Pradhan, S., Bacolla, A., Wells, R. D. & Roberts, R. J. Recombinant human DNA (cytosine-5) methyltransferase. I. Expression, purification, and comparison of de novo and maintenance methylation. *J Biol Chem* **274**, 33002-33010 (1999).
117 Li, E., Bestor, T. H. & Jaenisch, R. Targeted mutation of the DNA methyltransferase gene results in embryonic lethality. *Cell* **69**, 915-926 (1992).

118 Bestor, T. H. The DNA methyltransferases of mammals. *Hum Mol Genet* **9**, 2395-2402 (2000).
119 Okano, M., Xie, S. & Li, E. Cloning and characterization of a family of novel mammalian DNA (cytosine-5) methyltransferases. *Nat Genet* **19**, 219-220 (1998).
120 Xie, S. *et al.* Cloning, expression and chromosome locations of the human DNMT3 gene family. *Gene* **236**, 87-95 (1999).
121 Pradhan, S. & Esteve, P. O. Mammalian DNA (cytosine-5) methyltransferases and their expression. *Clin Immunol* **109**, 6-16 (2003).
122 Okano, M., Bell, D. W., Haber, D. A. & Li, E. DNA methyltransferases Dnmt3a and Dnmt3b are essential for de novo methylation and mammalian development. *Cell* **99**, 247-257 (1999).
123 Zhu, J. K. Active DNA demethylation mediated by DNA glycosylases. *Annu Rev Genet* **43**, 143-166 (2009).
124 Rideout, W. M., 3rd, Coetzee, G. A., Olumi, A. F. & Jones, P. A. 5-Methylcytosine as an endogenous mutagen in the human LDL receptor and p53 genes. *Science* **249**, 1288-1290 (1990).
125 Kriaucionis, S. & Heintz, N. The nuclear DNA base 5-hydroxymethylcytosine is present in Purkinje neurons and the brain. *Science* **324**, 929-930 (2009).
126 Tahiliani, M. *et al.* Conversion of 5-methylcytosine to 5-hydroxymethylcytosine in mammalian DNA by MLL partner TET1. *Science* **324**, 930-935 (2009).
127 Lorsbach, R. B. *et al.* TET1, a member of a novel protein family, is fused to MLL in acute myeloid leukemia containing the t(10;11)(q22;q23). *Leukemia* **17**, 637-641 (2003).
128 Ono, R. *et al.* LCX, leukemia-associated protein with a CXXC domain, is fused to MLL in acute myeloid leukemia with trilineage dysplasia having t(10;11)(q22;q23). *Cancer Res* **62**, 4075-4080 (2002).
129 Valinluck, V. & Sowers, L. C. Endogenous cytosine damage products alter the site selectivity of human DNA maintenance methyltransferase DNMT1. *Cancer Res* **67**, 946-950 (2007).
130 Cannon, S. V., Cummings, A. & Teebor, G. W. 5-Hydroxymethylcytosine DNA glycosylase activity in mammalian tissue. *Biochem Biophys Res Commun* **151**, 1173-1179 (1988).
131 Boorstein, R. J. *et al.* Definitive identification of mammalian 5-hydroxymethyluracil DNA N-glycosylase activity as SMUG1. *J Biol Chem* **276**, 41991-41997 (2001).
132 Alegria, A. H. Hydroxymethylation of pyrimidine mononucleotides with formaldehyde. *Biochim Biophys Acta* **149**, 317-324 (1967).
133 Liutkeviciute, Z., Lukinavicius, G., Masevicius, V., Daujotyte, D. & Klimasauskas, S. Cytosine-5-methyltransferases add aldehydes to DNA. *Nat Chem Biol* **5**, 400-402 (2009).
134 Privat, E. & Sowers, L. C. Photochemical deamination and demethylation of 5-methylcytosine. *Chem Res Toxicol* **9**, 745-750 (1996).
135 He, Y. F. *et al.* Tet-mediated formation of 5-carboxylcytosine and its excision by TDG in mammalian DNA. *Science* **333**, 1303-1307 (2011).
136 Maiti, A. & Drohat, A. C. Thymine DNA glycosylase can rapidly excise 5-formylcytosine and 5-carboxylcytosine: potential implications for active demethylation of CpG sites. *J Biol Chem* **286**, 35334-35338 (2011).
137 Zhang, L. *et al.* Thymine DNA glycosylase specifically recognizes 5-carboxylcytosine-modified DNA. *Nat Chem Biol* **8**, 328-330 (2012).
138 Schiesser, S. *et al.* Mechanism and stem-cell activity of 5-carboxycytosine decarboxylation determined by isotope tracing. *Angew Chem Int Ed Engl* **51**, 6516-6520 (2012).
139 Liutkeviciute, Z. *et al.* Direct decarboxylation of 5-carboxylcytosine by DNA C5-methyltransferases. *J Am Chem Soc* **136**, 5884-5887 (2014).
140 Allfrey, V. G. & Mirsky, A. E. Structural Modifications of Histones and their Possible Role in the Regulation of RNA Synthesis. *Science* **144**, 559 (1964).
141 Allfrey, V. G., Faulkner, R. & Mirsky, A. E. Acetylation and Methylation of Histones and Their Possible Role in the Regulation of Rna Synthesis. *Proc Natl Acad Sci U S A* **51**, 786-794 (1964).

142 Kuo, M. H. *et al.* Transcription-linked acetylation by Gcn5p of histones H3 and H4 at specific lysines. *Nature* **383**, 269-272 (1996).
143 Taunton, J., Hassig, C. A. & Schreiber, S. L. A mammalian histone deacetylase related to the yeast transcriptional regulator Rpd3p. *Science* **272**, 408-411 (1996).
144 Kuo, M. H., Zhou, J., Jambeck, P., Churchill, M. E. & Allis, C. D. Histone acetyltransferase activity of yeast Gcn5p is required for the activation of target genes in vivo. *Genes Dev* **12**, 627-639 (1998).
145 Wang, L., Liu, L. & Berger, S. L. Critical residues for histone acetylation by Gcn5, functioning in Ada and SAGA complexes, are also required for transcriptional function in vivo. *Genes Dev* **12**, 640-653 (1998).
146 Reifsnyder, C., Lowell, J., Clarke, A. & Pillus, L. Yeast SAS silencing genes and human genes associated with AML and HIV-1 Tat interactions are homologous with acetyltransferases. *Nat Genet* **14**, 42-49 (1996).
147 Clarke, A. S., Lowell, J. E., Jacobson, S. J. & Pillus, L. Esa1p is an essential histone acetyltransferase required for cell cycle progression. *Mol Cell Biol* **19**, 2515-2526 (1999).
148 Smith, E. R. *et al.* ESA1 is a histone acetyltransferase that is essential for growth in yeast. *Proc Natl Acad Sci U S A* **95**, 3561-3565 (1998).
149 Akhtar, A. & Becker, P. B. Activation of transcription through histone H4 acetylation by MOF, an acetyltransferase essential for dosage compensation in Drosophila. *Mol Cell* **5**, 367-375 (2000).
150 Borrow, J. *et al.* The translocation t(8;16)(p11;p13) of acute myeloid leukaemia fuses a putative acetyltransferase to the CREB-binding protein. *Nat Genet* **14**, 33-41 (1996).
151 Carapeti, M., Aguiar, R. C., Watmore, A. E., Goldman, J. M. & Cross, N. C. Consistent fusion of MOZ and TIF2 in AML with inv(8)(p11q13). *Cancer Genet Cytogenet* **113**, 70-72 (1999).
152 Shikama, N., Lyon, J. & LaThangue, N. B. The p300/CBP family: Integrating signals with transcription factors and chromatin. *Trends in Cell Biology* **7**, 230-236 (1997).
153 Mizzen, C. A. *et al.* The TAF(II)250 subunit of TFIID has histone acetyltransferase activity. *Cell* **87**, 1261-1270 (1996).
154 Jacobson, R. H., Ladurner, A. G., King, D. S. & Tjian, R. Structure and function of a human TAFII250 double bromodomain module. *Science* **288**, 1422-1425 (2000).
155 Kawasaki, H. *et al.* ATF-2 has intrinsic histone acetyltransferase activity which is modulated by phosphorylation. *Nature* **405**, 195-200 (2000).
156 Burley, S. K. DNA-Binding Motifs from Eukaryotic Transcription Factors. *Curr Opin Struc Biol* **4**, 3-11 (1994).
157 Clements, A. & Marmorstein, R. Insights into structure and function of GCN5/PCAF and yEsa 1 histone acetyltransferase domains. *Method Enzymol* **371**, 545-+ (2003).
158 Angus-Hill, M. L., Dutnall, R. N., Tafrov, S. T., Sternglanz, R. & Ramakrishnan, V. Crystal structure of the histone acetyltransferase Hpa2: A tetrameric member of the Gcn5-related N-acetyltransferase superfamily. *Journal of Molecular Biology* **294**, 1311-1325 (1999).
159 Itazaki, H. *et al.* Isolation and Structural Elucidation of New Cyclotetrapeptides, Trapoxin-a and Trapoxin-B, Having Detransformation Activities as Antitumor Agents. *J Antibiot* **43**, 1524-1532 (1990).
160 Kijima, M., Yoshida, M., Sugita, K., Horinouchi, S. & Beppu, T. Trapoxin, an Antitumor Cyclic Tetrapeptide, Is an Irreversible Inhibitor of Mammalian Histone Deacetylase. *Journal of Biological Chemistry* **268**, 22429-22435 (1993).
161 Dangond, F. & Gullans, S. R. Differential expression of human histone deacetylase mRNAs in response to immune cell apoptosis induction by trichostatin A and butyrate. *Biochem Bioph Res Co* **247**, 833-837 (1998).
162 Grozinger, C. M., Hassig, C. A. & Schreiber, S. L. Three proteins define a class of human histone deacetylases related to yeast Hda1p. *P Natl Acad Sci USA* **96**, 4868-4873 (1999).

163 Fischle, W. *et al.* A new family of human histone deacetylases related to Saccharomyces cerevisiae HDA1p. *Journal of Biological Chemistry* **274**, 11713-11720 (1999).
164 Zhou, X. B., Marks, P. A., Rifkind, R. A. & Richon, V. M. Cloning and characterization of a histone deacetylase, HDAC9. *P Natl Acad Sci USA* **98**, 10572-10577 (2001).
165 Fischer, D. D. *et al.* Isolation and characterization of a novel class II histone deacetylase, HDAC10. *Journal of Biological Chemistry* **277**, 6656-6666 (2002).
166 Vaziri, H. *et al.* hSIR2(SIRT1) functions as an NAD-dependent p53 deacetylase. *Cell* **107**, 149-159 (2001).
167 Langley, E. *et al.* Human SIR2 deacetylates p53 and antagonizes PML/p53-induced cellular senescence. *EMBO J* **21**, 2383-2396 (2002).
168 Paik, W. K., Paik, D. C. & Kim, S. Historical review: the field of protein methylation. *Trends Biochem Sci* **32**, 146-152 (2007).
169 Pahlich, S., Zakaryan, R. P. & Gehring, H. Protein arginine methylation: Cellular functions and methods of analysis. *Biochim Biophys Acta* **1764**, 1890-1903 (2006).
170 McBride, A. E. & Silver, P. A. State of the arg: protein methylation at arginine comes of age. *Cell* **106**, 5-8 (2001).
171 Bedford, M. T. & Richard, S. Arginine methylation an emerging regulator of protein function. *Mol Cell* **18**, 263-272 (2005).
172 Abramovich, C., Yakobson, B., Chebath, J. & Revel, M. A protein-arginine methyltransferase binds to the intracytoplasmic domain of the IFNAR1 chain in the type I interferon receptor. *EMBO J* **16**, 260-266 (1997).
173 Lin, W. J., Gary, J. D., Yang, M. C., Clarke, S. & Herschman, H. R. The mammalian immediate-early TIS21 protein and the leukemia-associated BTG1 protein interact with a protein-arginine N-methyltransferase. *J Biol Chem* **271**, 15034-15044 (1996).
174 Scott, H. S. *et al.* Identification and characterization of two putative human arginine methyltransferases (HRMT1L1 and HRMT1L2). *Genomics* **48**, 330-340 (1998).
175 Tang, J. *et al.* PRMT1 is the predominant type I protein arginine methyltransferase in mammalian cells. *J Biol Chem* **275**, 7723-7730 (2000).
176 Paik, W. K. & Kim, S. Protein methylase I. Purification and properties of the enzyme. *J Biol Chem* **243**, 2108-2114 (1968).
177 Kwak, Y. T. *et al.* Methylation of SPT5 regulates its interaction with RNA polymerase II and transcriptional elongation properties. *Mol Cell* **11**, 1055-1066 (2003).
178 Katsanis, N., Yaspo, M. L. & Fisher, E. M. Identification and mapping of a novel human gene, HRMT1L1, homologous to the rat protein arginine N-methyltransferase 1 (PRMT1) gene. *Mamm Genome* **8**, 526-529 (1997).
179 Qi, C. *et al.* Identification of protein arginine methyltransferase 2 as a coactivator for estrogen receptor alpha. *J Biol Chem* **277**, 28624-28630 (2002).
180 Meyer, R., Wolf, S. S. & Obendorf, M. PRMT2, a member of the protein arginine methyltransferase family, is a coactivator of the androgen receptor. *J Steroid Biochem Mol Biol* **107**, 1-14 (2007).
181 Tang, J., Gary, J. D., Clarke, S. & Herschman, H. R. PRMT 3, a type I protein arginine N-methyltransferase that differs from PRMT1 in its oligomerization, subcellular localization, substrate specificity, and regulation. *J Biol Chem* **273**, 16935-16945 (1998).
182 Bachand, F. & Silver, P. A. PRMT3 is a ribosomal protein methyltransferase that affects the cellular levels of ribosomal subunits. *EMBO J* **23**, 2641-2650 (2004).
183 Chen, D. *et al.* Regulation of transcription by a protein methyltransferase. *Science* **284**, 2174-2177 (1999).
184 Stallcup, M. R. *et al.* Co-operation between protein-acetylating and protein-methylating co-activators in transcriptional activation. *Biochem Soc Trans* **28**, 415-418 (2000).
185 Schurter, B. T. *et al.* Methylation of histone H3 by coactivator-associated arginine methyltransferase 1. *Biochemistry* **40**, 5747-5756 (2001).

186 Fabbrizio, E. *et al.* Negative regulation of transcription by the type II arginine methyltransferase PRMT5. *EMBO Rep* **3**, 641-645 (2002).
187 Richard, S., Morel, M. & Cleroux, P. Arginine methylation regulates IL-2 gene expression: a role for protein arginine methyltransferase 5 (PRMT5). *Biochem J* **388**, 379-386 (2005).
188 Pal, S., Vishwanath, S. N., Erdjument-Bromage, H., Tempst, P. & Sif, S. Human SWI/SNF-associated PRMT5 methylates histone H3 arginine 8 and negatively regulates expression of ST7 and NM23 tumor suppressor genes. *Mol Cell Biol* **24**, 9630-9645 (2004).
189 Lacroix, M. *et al.* The histone-binding protein COPR5 is required for nuclear functions of the protein arginine methyltransferase PRMT5. *EMBO Rep* **9**, 452-458 (2008).
190 Frankel, A. *et al.* The novel human protein arginine N-methyltransferase PRMT6 is a nuclear enzyme displaying unique substrate specificity. *J Biol Chem* **277**, 3537-3543 (2002).
191 Guccione, E. *et al.* Methylation of histone H3R2 by PRMT6 and H3K4 by an MLL complex are mutually exclusive. *Nature* **449**, 933-937 (2007).
192 Boulanger, M. C. *et al.* Methylation of Tat by PRMT6 regulates human immunodeficiency virus type 1 gene expression. *J Virol* **79**, 124-131 (2005).
193 Invernizzi, C. F. *et al.* Arginine methylation of the HIV-1 nucleocapsid protein results in its diminished function. *AIDS* **21**, 795-805 (2007).
194 Xie, B., Invernizzi, C. F., Richard, S. & Wainberg, M. A. Arginine methylation of the human immunodeficiency virus type 1 Tat protein by PRMT6 negatively affects Tat Interactions with both cyclin T1 and the Tat transactivation region. *J Virol* **81**, 4226-4234 (2007).
195 Miranda, T. B., Miranda, M., Frankel, A. & Clarke, S. PRMT7 is a member of the protein arginine methyltransferase family with a distinct substrate specificity. *J Biol Chem* **279**, 22902-22907 (2004).
196 Lee, J. H. *et al.* PRMT7, a new protein arginine methyltransferase that synthesizes symmetric dimethylarginine. *J Biol Chem* **280**, 3656-3664 (2005).
197 Zurita-Lopez, C. I., Sandberg, T., Kelly, R. & Clarke, S. G. Human protein arginine methyltransferase 7 (PRMT7) is a type III enzyme forming omega-NG-monomethylated arginine residues. *J Biol Chem* **287**, 7859-7870 (2012).
198 Lee, J., Sayegh, J., Daniel, J., Clarke, S. & Bedford, M. T. PRMT8, a new membrane-bound tissue-specific member of the protein arginine methyltransferase family. *J Biol Chem* **280**, 32890-32896 (2005).
199 Sayegh, J., Webb, K., Cheng, D., Bedford, M. T. & Clarke, S. G. Regulation of protein arginine methyltransferase 8 (PRMT8) activity by its N-terminal domain. *J Biol Chem* **282**, 36444-36453 (2007).
200 Kim, J. D., Kako, K., Kakiuchi, M., Park, G. G. & Fukamizu, A. EWS is a substrate of type I protein arginine methyltransferase, PRMT8. *Int J Mol Med* **22**, 309-315 (2008).
201 Pahlich, S., Zakaryan, R. P. & Gehring, H. Identification of proteins interacting with protein arginine methyltransferase 8: the Ewing sarcoma (EWS) protein binds independent of its methylation state. *Proteins* **72**, 1125-1137 (2008).
202 Cook, J. R. *et al.* FBXO11/PRMT9, a new protein arginine methyltransferase, symmetrically dimethylates arginine residues. *Biochem Biophys Res Commun* **342**, 472-481 (2006).
203 Krause, C. D. *et al.* Protein arginine methyltransferases: evolution and assessment of their pharmacological and therapeutic potential. *Pharmacol Ther* **113**, 50-87 (2007).
204 Pal, S. & Sif, S. Interplay between chromatin remodelers and protein arginine methyltransferases. *J Cell Physiol* **213**, 306-315 (2007).
205 Blatch, G. L. & Lassle, M. The tetratricopeptide repeat: a structural motif mediating protein-protein interactions. *Bioessays* **21**, 932-939 (1999).
206 Rea, S. *et al.* Regulation of chromatin structure by site-specific histone H3 methyltransferases. *Nature* **406**, 593-599 (2000).
207 van Leeuwen, F., Gafken, P. R. & Gottschling, D. E. Dot1p modulates silencing in yeast by methylation of the nucleosome core. *Cell* **109**, 745-756 (2002).

208 Shi, Y. *et al.* Histone demethylation mediated by the nuclear amine oxidase homolog LSD1. *Cell* **119**, 941-953 (2004).
209 Allis, C. D. *et al.* New nomenclature for chromatin-modifying enzymes. *Cell* **131**, 633-636 (2007).
210 Shi, Y. & Whetstine, J. R. Dynamic regulation of histone lysine methylation by demethylases. *Mol Cell* **25**, 1-14 (2007).
211 Karytinos, A. *et al.* A novel mammalian flavin-dependent histone demethylase. *J Biol Chem* **284**, 17775-17782 (2009).
212 Metzger, E. *et al.* LSD1 demethylates repressive histone marks to promote androgen-receptor-dependent transcription. *Nature* **437**, 436-439 (2005).
213 Wissmann, M. *et al.* Cooperative demethylation by JMJD2C and LSD1 promotes androgen receptor-dependent gene expression. *Nat Cell Biol* **9**, 347-353 (2007).
214 Shi, Y. J. *et al.* Regulation of LSD1 histone demethylase activity by its associated factors. *Mol Cell* **19**, 857-864 (2005).
215 Lee, M. G., Wynder, C., Cooch, N. & Shiekhattar, R. An essential role for CoREST in nucleosomal histone 3 lysine 4 demethylation. *Nature* **437**, 432-435 (2005).
216 Yang, M. *et al.* Structural basis for CoREST-dependent demethylation of nucleosomes by the human LSD1 histone demethylase. *Mol Cell* **23**, 377-387 (2006).
217 Metzger, E. *et al.* Phosphorylation of histone H3T6 by PKCbeta(I) controls demethylation at histone H3K4. *Nature* **464**, 792-796 (2010).
218 Fang, R. *et al.* Human LSD2/KDM1b/AOF1 regulates gene transcription by modulating intragenic H3K4me2 methylation. *Mol Cell* **39**, 222-233 (2010).
219 Frescas, D., Guardavaccaro, D., Bassermann, F., Koyama-Nasu, R. & Pagano, M. JHDM1B/FBXL10 is a nucleolar protein that represses transcription of ribosomal RNA genes. *Nature* **450**, 309-313 (2007).
220 He, J., Kallin, E. M., Tsukada, Y. & Zhang, Y. The H3K36 demethylase Jhdm1b/Kdm2b regulates cell proliferation and senescence through p15(Ink4b). *Nat Struct Mol Biol* **15**, 1169-1175 (2008).
221 Lu, T. *et al.* Validation-based insertional mutagenesis identifies lysine demethylase FBXL11 as a negative regulator of NFkappaB. *Proc Natl Acad Sci U S A* **106**, 16339-16344 (2009).
222 Tanaka, Y. *et al.* JmjC enzyme KDM2A is a regulator of rRNA transcription in response to starvation. *EMBO J* **29**, 1510-1522 (2010).
223 Yamane, K. *et al.* JHDM2A, a JmjC-containing H3K9 demethylase, facilitates transcription activation by androgen receptor. *Cell* **125**, 483-495 (2006).
224 Okada, Y., Scott, G., Ray, M. K., Mishina, Y. & Zhang, Y. Histone demethylase JHDM2A is critical for Tnp1 and Prm1 transcription and spermatogenesis. *Nature* **450**, 119-123 (2007).
225 Ma, D. K., Chiang, C. H., Ponnusamy, K., Ming, G. L. & Song, H. G9a and Jhdm2a regulate embryonic stem cell fusion-induced reprogramming of adult neural stem cells. *Stem Cells* **26**, 2131-2141 (2008).
226 Klose, R. J. *et al.* The transcriptional repressor JHDM3A demethylates trimethyl histone H3 lysine 9 and lysine 36. *Nature* **442**, 312-316 (2006).
227 Fodor, B. D. *et al.* Jmjd2b antagonizes H3K9 trimethylation at pericentric heterochromatin in mammalian cells. *Genes Dev* **20**, 1557-1562 (2006).
228 Cloos, P. A. *et al.* The putative oncogene GASC1 demethylates tri- and dimethylated lysine 9 on histone H3. *Nature* **442**, 307-311 (2006).
229 Ponnaluri, V. K., Vavilala, D. T., Putty, S., Gutheil, W. G. & Mukherji, M. Identification of non-histone substrates for JMJD2A-C histone demethylases. *Biochem Biophys Res Commun* **390**, 280-284 (2009).
230 Trojer, P. *et al.* Dynamic Histone H1 Isotype 4 Methylation and Demethylation by Histone Lysine Methyltransferase G9a/KMT1C and the Jumonji Domain-containing JMJD2/KDM4 Proteins. *J Biol Chem* **284**, 8395-8405 (2009).

231 Loh, Y. H., Zhang, W., Chen, X., George, J. & Ng, H. H. Jmjd1a and Jmjd2c histone H3 Lys 9 demethylases regulate self-renewal in embryonic stem cells. *Genes Dev* **21**, 2545-2557 (2007).
232 Zhang, D., Yoon, H. G. & Wong, J. JMJD2A is a novel N-CoR-interacting protein and is involved in repression of the human transcription factor achaete scute-like homologue 2 (ASCL2/Hash2). *Mol Cell Biol* **25**, 6404-6414 (2005).
233 Gray, S. G. *et al.* Functional characterization of JMJD2A, a histone deacetylase- and retinoblastoma-binding protein. *J Biol Chem* **280**, 28507-28518 (2005).
234 Klose, R. J. *et al.* The retinoblastoma binding protein RBP2 is an H3K4 demethylase. *Cell* **128**, 889-900 (2007).
235 Iwase, S. *et al.* The X-linked mental retardation gene SMCX/JARID1C defines a family of histone H3 lysine 4 demethylases. *Cell* **128**, 1077-1088 (2007).
236 Yamane, K. *et al.* PLU-1 is an H3K4 demethylase involved in transcriptional repression and breast cancer cell proliferation. *Mol Cell* **25**, 801-812 (2007).
237 Tu, S. *et al.* The ARID domain of the H3K4 demethylase RBP2 binds to a DNA CCGCCC motif. *Nat Struct Mol Biol* **15**, 419-421 (2008).
238 van Oevelen, C. *et al.* A role for mammalian Sin3 in permanent gene silencing. *Mol Cell* **32**, 359-370 (2008).
239 Dey, B. K. *et al.* The histone demethylase KDM5b/JARID1b plays a role in cell fate decisions by blocking terminal differentiation. *Mol Cell Biol* **28**, 5312-5327 (2008).
240 Tan, K. *et al.* Human PLU-1 Has transcriptional repression properties and interacts with the developmental transcription factors BF-1 and PAX9. *J Biol Chem* **278**, 20507-20513 (2003).
241 Jensen, L. R. *et al.* Mutations in the JARID1C gene, which is involved in transcriptional regulation and chromatin remodeling, cause X-linked mental retardation. *Am J Hum Genet* **76**, 227-236 (2005).
242 Lee, M. G., Norman, J., Shilatifard, A. & Shiekhattar, R. Physical and functional association of a trimethyl H3K4 demethylase and Ring6a/MBLR, a polycomb-like protein. *Cell* **128**, 877-887 (2007).
243 Akimoto, C., Kitagawa, H., Matsumoto, T. & Kato, S. Spermatogenesis-specific association of SMCY and MSH5. *Genes Cells* **13**, 623-633 (2008).
244 Hong, S. *et al.* Identification of JmjC domain-containing UTX and JMJD3 as histone H3 lysine 27 demethylases. *Proc Natl Acad Sci U S A* **104**, 18439-18444 (2007).
245 Lan, F. *et al.* A histone H3 lysine 27 demethylase regulates animal posterior development. *Nature* **449**, 689-694 (2007).
246 Agger, K. *et al.* UTX and JMJD3 are histone H3K27 demethylases involved in HOX gene regulation and development. *Nature* **449**, 731-734 (2007).
247 Lee, M. G. *et al.* Demethylation of H3K27 regulates polycomb recruitment and H2A ubiquitination. *Science* **318**, 447-450 (2007).
248 De Santa, F. *et al.* The histone H3 lysine-27 demethylase Jmjd3 links inflammation to inhibition of polycomb-mediated gene silencing. *Cell* **130**, 1083-1094 (2007).
249 Satoh, T. *et al.* The Jmjd3-Irf4 axis regulates M2 macrophage polarization and host responses against helminth infection. *Nat Immunol* **11**, 936-944 (2010).
250 Wang, L., Jin, Q., Lee, J. E., Su, I. H. & Ge, K. Histone H3K27 methyltransferase Ezh2 represses Wnt genes to facilitate adipogenesis. *Proc Natl Acad Sci U S A* **107**, 7317-7322 (2010).
251 Burgold, T. *et al.* The histone H3 lysine 27-specific demethylase Jmjd3 is required for neural commitment. *PLoS One* **3**, e3034 (2008).
252 Akizu, N., Estaras, C., Guerrero, L., Marti, E. & Martinez-Balbas, M. A. H3K27me3 regulates BMP activity in developing spinal cord. *Development* **137**, 2915-2925 (2010).
253 Tsukada, Y., Ishitani, T. & Nakayama, K. I. KDM7 is a dual demethylase for histone H3 Lys 9 and Lys 27 and functions in brain development. *Genes Dev* **24**, 432-437 (2010).

254 Fortschegger, K. & Shiekhattar, R. Plant homeodomain fingers form a helping hand for transcription. *Epigenetics* **6**, 4-8 (2011).
255 Liu, T. *et al.* Broad chromosomal domains of histone modification patterns in C. elegans. *Genome Res* **21**, 227-236 (2011).
256 Qi, H. H. *et al.* Histone H4K20/H3K9 demethylase PHF8 regulates zebrafish brain and craniofacial development. *Nature* **466**, 503-507 (2010).
257 Huang, C. *et al.* The dual histone demethylase KDM7A promotes neural induction in early chick embryos. *Dev Dyn* **239**, 3350-3357 (2010).
258 Wen, H. *et al.* Recognition of histone H3K4 trimethylation by the plant homeodomain of PHF2 modulates histone demethylation. *J Biol Chem* **285**, 9322-9326 (2010).
259 Feng, W., Yonezawa, M., Ye, J., Jenuwein, T. & Grummt, I. PHF8 activates transcription of rRNA genes through H3K4me3 binding and H3K9me1/2 demethylation. *Nat Struct Mol Biol* **17**, 445-450 (2010).
260 Zhu, Z. *et al.* PHF8 is a histone H3K9me2 demethylase regulating rRNA synthesis. *Cell Res* **20**, 794-801 (2010).
261 Hsia, D. A. *et al.* KDM8, a H3K36me2 histone demethylase that acts in the cyclin A1 coding region to regulate cancer cell proliferation. *Proc Natl Acad Sci U S A* **107**, 9671-9676 (2010).
262 Chang, B., Chen, Y., Zhao, Y. & Bruick, R. K. JMJD6 is a histone arginine demethylase. *Science* **318**, 444-447 (2007).
263 Oki, M., Aihara, H. & Ito, T. Role of histone phosphorylation in chromatin dynamics and its implications in diseases. *Subcell Biochem* **41**, 319-336 (2007).
264 Koshland, D. & Strunnikov, A. Mitotic chromosome condensation. *Annu Rev Cell Dev Biol* **12**, 305-333 (1996).
265 Mahadevan, L. C., Willis, A. C. & Barratt, M. J. Rapid histone H3 phosphorylation in response to growth factors, phorbol esters, okadaic acid, and protein synthesis inhibitors. *Cell* **65**, 775-783 (1991).
266 Dawson, M. A. *et al.* JAK2 phosphorylates histone H3Y41 and excludes HP1alpha from chromatin. *Nature* **461**, 819-822 (2009).
267 Wang, Y. *et al.* Human PAD4 regulates histone arginine methylation levels via demethylimination. *Science* **306**, 279-283 (2004).
268 Bannister, A. J. & Kouzarides, T. Reversing histone methylation. *Nature* **436**, 1103-1106 (2005).
269 Goldknopf, I. L. *et al.* Isolation and characterization of protein A24, a "histone-like" non-histone chromosomal protein. *J Biol Chem* **250**, 7182-7187 (1975).
270 Jason, L. J., Moore, S. C., Lewis, J. D., Lindsey, G. & Ausio, J. Histone ubiquitination: a tagging tail unfolds? *Bioessays* **24**, 166-174 (2002).
271 Wang, H. *et al.* Role of histone H2A ubiquitination in Polycomb silencing. *Nature* **431**, 873-878 (2004).
272 Cao, R., Tsukada, Y. & Zhang, Y. Role of Bmi-1 and Ring1A in H2A ubiquitylation and Hox gene silencing. *Mol Cell* **20**, 845-854 (2005).
273 Nickel, B. E. & Davie, J. R. Structure of polyubiquitinated histone H2A. *Biochemistry* **28**, 964-968 (1989).
274 Gearhart, M. D., Corcoran, C. M., Wamstad, J. A. & Bardwell, V. J. Polycomb group and SCF ubiquitin ligases are found in a novel BCOR complex that is recruited to BCL6 targets. *Mol Cell Biol* **26**, 6880-6889 (2006).
275 Chen, A., Kleiman, F. E., Manley, J. L., Ouchi, T. & Pan, Z. Q. Autoubiquitination of the BRCA1*BARD1 RING ubiquitin ligase. *J Biol Chem* **277**, 22085-22092 (2002).
276 Zhu, Q. *et al.* BRCA1 tumour suppression occurs via heterochromatin-mediated silencing. *Nature* **477**, 179-184 (2011).
277 Thorne, A. W., Sautiere, P., Briand, G. & Crane-Robinson, C. The structure of ubiquitinated histone H2B. *EMBO J* **6**, 1005-1010 (1987).

278 Kim, J., Hake, S. B. & Roeder, R. G. The human homolog of yeast BRE1 functions as a transcriptional coactivator through direct activator interactions. *Mol Cell* **20**, 759-770 (2005).
279 Prenzel, T. *et al.* Estrogen-dependent gene transcription in human breast cancer cells relies upon proteasome-dependent monoubiquitination of histone H2B. *Cancer Res* **71**, 5739-5753 (2011).
280 Shiio, Y. & Eisenman, R. N. Histone sumoylation is associated with transcriptional repression. *Proc Natl Acad Sci U S A* **100**, 13225-13230 (2003).
281 Nathan, D., Sterner, D. E. & Berger, S. L. Histone modifications: Now summoning sumoylation. *Proc Natl Acad Sci U S A* **100**, 13118-13120 (2003).
282 Hassa, P. O., Haenni, S. S., Elser, M. & Hottiger, M. O. Nuclear ADP-ribosylation reactions in mammalian cells: where are we today and where are we going? *Microbiol Mol Biol Rev* **70**, 789-829 (2006).
283 Nelson, C. J., Santos-Rosa, H. & Kouzarides, T. Proline isomerization of histone H3 regulates lysine methylation and gene expression. *Cell* **126**, 905-916 (2006).
284 Chen, Y. R. & Clark, A. C. Substitutions of prolines examine their role in kinetic trap formation of the caspase recruitment domain (CARD) of RICK. *Protein Sci* **15**, 395-409 (2006).
285 Tan, M. *et al.* Identification of 67 histone marks and histone lysine crotonylation as a new type of histone modification. *Cell* **146**, 1016-1028 (2011).
286 Namekawa, S. H. *et al.* Postmeiotic sex chromatin in the male germline of mice. *Curr Biol* **16**, 660-667 (2006).
287 Turner, J. M. Meiotic sex chromosome inactivation. *Development* **134**, 1823-1831 (2007).
288 Mueller, J. L. *et al.* The mouse X chromosome is enriched for multicopy testis genes showing postmeiotic expression. *Nat Genet* **40**, 794-799 (2008).
289 Iguchi-Ariga, S. M. & Schaffner, W. CpG methylation of the cAMP-responsive enhancer/promoter sequence TGACGTCA abolishes specific factor binding as well as transcriptional activation. *Genes Dev* **3**, 612-619 (1989).
290 Prendergast, G. C. & Ziff, E. B. Methylation-sensitive sequence-specific DNA binding by the c-Myc basic region. *Science* **251**, 186-189 (1991).
291 Campanero, M. R., Armstrong, M. I. & Flemington, E. K. CpG methylation as a mechanism for the regulation of E2F activity. *Proc Natl Acad Sci U S A* **97**, 6481-6486 (2000).
292 Santoro, R. & Grummt, I. Molecular mechanisms mediating methylation-dependent silencing of ribosomal gene transcription. *Mol Cell* **8**, 719-725 (2001).
293 Meehan, R. R., Lewis, J. D., McKay, S., Kleiner, E. L. & Bird, A. P. Identification of a mammalian protein that binds specifically to DNA containing methylated CpGs. *Cell* **58**, 499-507 (1989).
294 Lewis, J. D. *et al.* Purification, sequence, and cellular localization of a novel chromosomal protein that binds to methylated DNA. *Cell* **69**, 905-914 (1992).
295 Nan, X., Meehan, R. R. & Bird, A. Dissection of the methyl-CpG binding domain from the chromosomal protein MeCP2. *Nucleic Acids Res* **21**, 4886-4892 (1993).
296 Nan, X. *et al.* Transcriptional repression by the methyl-CpG-binding protein MeCP2 involves a histone deacetylase complex. *Nature* **393**, 386-389 (1998).
297 Hendrich, B. & Bird, A. Identification and characterization of a family of mammalian methyl-CpG binding proteins. *Mol Cell Biol* **18**, 6538-6547 (1998).
298 Marhold, J., Kramer, K., Kremmer, E. & Lyko, F. The Drosophila MBD2/3 protein mediates interactions between the MI-2 chromatin complex and CpT/A-methylated DNA. *Development* **131**, 6033-6039 (2004).
299 Lyko, F., Ramsahoye, B. H. & Jaenisch, R. DNA methylation in Drosophila melanogaster. *Nature* **408**, 538-540 (2000).
300 Ng, H. H. *et al.* MBD2 is a transcriptional repressor belonging to the MeCP1 histone deacetylase complex. *Nat Genet* **23**, 58-61 (1999).

301 Kantor, B., Makedonski, K., Shemer, R. & Razin, A. Expression and localization of components of the histone deacetylases multiprotein repressory complexes in the mouse preimplantation embryo. *Gene Expr Patterns* **3**, 697-702 (2003).
302 Huntriss, J. *et al.* Expression of mRNAs for DNA methyltransferases and methyl-CpG-binding proteins in the human female germ line, preimplantation embryos, and embryonic stem cells. *Mol Reprod Dev* **67**, 323-336 (2004).
303 Cassel, S., Revel, M. O., Kelche, C. & Zwiller, J. Expression of the methyl-CpG-binding protein MeCP2 in rat brain. An ontogenetic study. *Neurobiol Dis* **15**, 206-211 (2004).
304 Urdinguio, R. G. *et al.* Mecp2-null mice provide new neuronal targets for Rett syndrome. *PLoS One* **3**, e3669 (2008).
305 Auriol, E., Billard, L. M., Magdinier, F. & Dante, R. Specific binding of the methyl binding domain protein 2 at the BRCA1-NBR2 locus. *Nucleic Acids Res* **33**, 4243-4254 (2005).
306 Amir, R. E. *et al.* Rett syndrome is caused by mutations in X-linked MECP2, encoding methyl-CpG-binding protein 2. *Nat Genet* **23**, 185-188 (1999).
307 Chahrour, M. *et al.* MeCP2, a key contributor to neurological disease, activates and represses transcription. *Science* **320**, 1224-1229 (2008).
308 Skene, P. J. *et al.* Neuronal MeCP2 is expressed at near histone-octamer levels and globally alters the chromatin state. *Mol Cell* **37**, 457-468 (2010).
309 Yang, C., van der Woerd, M. J., Muthurajan, U. M., Hansen, J. C. & Luger, K. Biophysical analysis and small-angle X-ray scattering-derived structures of MeCP2-nucleosome complexes. *Nucleic Acids Res* **39**, 4122-4135 (2011).
310 Ohki, I. *et al.* Solution structure of the methyl-CpG binding domain of human MBD1 in complex with methylated DNA. *Cell* **105**, 487-497 (2001).
311 Jorgensen, H. F., Ben-Porath, I. & Bird, A. P. Mbd1 is recruited to both methylated and nonmethylated CpGs via distinct DNA binding domains. *Mol Cell Biol* **24**, 3387-3395 (2004).
312 Clouaire, T., de Las Heras, J. I., Merusi, C. & Stancheva, I. Recruitment of MBD1 to target genes requires sequence-specific interaction of the MBD domain with methylated DNA. *Nucleic Acids Res* **38**, 4620-4634 (2010).
313 Liu, C. *et al.* Epigenetic regulation of miR-184 by MBD1 governs neural stem cell proliferation and differentiation. *Cell Stem Cell* **6**, 433-444 (2010).
314 Berger, J., Sansom, O., Clarke, A. & Bird, A. MBD2 is required for correct spatial gene expression in the gut. *Mol Cell Biol* **27**, 4049-4057 (2007).
315 Chatagnon, A., Ballestar, E., Esteller, M. & Dante, R. A role for methyl-CpG binding domain protein 2 in the modulation of the estrogen response of pS2/TFF1 gene. *PLoS One* **5**, e9665 (2010).
316 Chatagnon, A. *et al.* Specific association between the methyl-CpG-binding domain protein 2 and the hypermethylated region of the human telomerase reverse transcriptase promoter in cancer cells. *Carcinogenesis* **30**, 28-34 (2009).
317 Hutchins, A. S. *et al.* Gene silencing quantitatively controls the function of a developmental trans-activator. *Mol Cell* **10**, 81-91 (2002).
318 Guy, J., Hendrich, B., Holmes, M., Martin, J. E. & Bird, A. A mouse Mecp2-null mutation causes neurological symptoms that mimic Rett syndrome. *Nat Genet* **27**, 322-326 (2001).
319 Hendrich, B., Guy, J., Ramsahoye, B., Wilson, V. A. & Bird, A. Closely related proteins MBD2 and MBD3 play distinctive but interacting roles in mouse development. *Genes Dev* **15**, 710-723 (2001).
320 Zhao, X. *et al.* Mice lacking methyl-CpG binding protein 1 have deficits in adult neurogenesis and hippocampal function. *Proc Natl Acad Sci U S A* **100**, 6777-6782 (2003).
321 Hendrich, B., Hardeland, U., Ng, H. H., Jiricny, J. & Bird, A. The thymine glycosylase MBD4 can bind to the product of deamination at methylated CpG sites. *Nature* **401**, 301-304 (1999).

322 Petronzelli, F. *et al.* Investigation of the substrate spectrum of the human mismatch-specific DNA N-glycosylase MED1 (MBD4): fundamental role of the catalytic domain. *J Cell Physiol* **185**, 473-480 (2000).
323 Petronzelli, F. *et al.* Biphasic kinetics of the human DNA repair protein MED1 (MBD4), a mismatch-specific DNA N-glycosylase. *J Biol Chem* **275**, 32422-32429 (2000).
324 Daniel, J. M. & Reynolds, A. B. The catenin p120(ctn) interacts with Kaiso, a novel BTB/POZ domain zinc finger transcription factor. *Mol Cell Biol* **19**, 3614-3623 (1999).
325 Prokhortchouk, A. *et al.* The p120 catenin partner Kaiso is a DNA methylation-dependent transcriptional repressor. *Genes Dev* **15**, 1613-1618 (2001).
326 Filion, G. J. *et al.* A family of human zinc finger proteins that bind methylated DNA and repress transcription. *Mol Cell Biol* **26**, 169-181 (2006).
327 Daniel, J. M., Spring, C. M., Crawford, H. C., Reynolds, A. B. & Baig, A. The p120(ctn)-binding partner Kaiso is a bi-modal DNA-binding protein that recognizes both a sequence-specific consensus and methylated CpG dinucleotides. *Nucleic Acids Res* **30**, 2911-2919 (2002).
328 Ruzov, A. *et al.* Kaiso is a genome-wide repressor of transcription that is essential for amphibian development. *Development* **131**, 6185-6194 (2004).
329 Sasai, N., Nakao, M. & Defossez, P. A. Sequence-specific recognition of methylated DNA by human zinc-finger proteins. *Nucleic Acids Res* **38**, 5015-5022 (2010).
330 Citterio, E. *et al.* Np95 is a histone-binding protein endowed with ubiquitin ligase activity. *Mol Cell Biol* **24**, 2526-2535 (2004).
331 Rottach, A. *et al.* The multi-domain protein Np95 connects DNA methylation and histone modification. *Nucleic Acids Res* **38**, 1796-1804 (2010).
332 Unoki, M., Nishidate, T. & Nakamura, Y. ICBP90, an E2F-1 target, recruits HDAC1 and binds to methyl-CpG through its SRA domain. *Oncogene* **23**, 7601-7610 (2004).
333 Bostick, M. *et al.* UHRF1 plays a role in maintaining DNA methylation in mammalian cells. *Science* **317**, 1760-1764 (2007).
334 Sharif, J. *et al.* The SRA protein Np95 mediates epigenetic inheritance by recruiting Dnmt1 to methylated DNA. *Nature* **450**, 908-912 (2007).
335 Arita, K., Ariyoshi, M., Tochio, H., Nakamura, Y. & Shirakawa, M. Recognition of hemi-methylated DNA by the SRA protein UHRF1 by a base-flipping mechanism. *Nature* **455**, 818-821 (2008).
336 Avvakumov, G. V. *et al.* Structural basis for recognition of hemi-methylated DNA by the SRA domain of human UHRF1. *Nature* **455**, 822-825 (2008).
337 Hashimoto, H. *et al.* The SRA domain of UHRF1 flips 5-methylcytosine out of the DNA helix. *Nature* **455**, 826-829 (2008).
338 Fujimori, A. *et al.* Cloning and mapping of Np95 gene which encodes a novel nuclear protein associated with cell proliferation. *Mamm Genome* **9**, 1032-1035 (1998).
339 Bronner, C. *et al.* The antiapoptotic protein ICBP90 is a target for protein kinase 2. *Ann N Y Acad Sci* **1030**, 355-360 (2004).
340 Trotzier, M. A. *et al.* Phosphorylation of ICBP90 by protein kinase A enhances topoisomerase IIalpha expression. *Biochem Biophys Res Commun* **319**, 590-595 (2004).
341 Arima, Y. *et al.* Down-regulation of nuclear protein ICBP90 by p53/p21Cip1/WAF1-dependent DNA-damage checkpoint signals contributes to cell cycle arrest at G1/S transition. *Genes Cells* **9**, 131-142 (2004).
342 Mousli, M. *et al.* ICBP90 belongs to a new family of proteins with an expression that is deregulated in cancer cells. *Br J Cancer* **89**, 120-127 (2003).
343 Jeanblanc, M. *et al.* The retinoblastoma gene and its product are targeted by ICBP90: a key mechanism in the G1/S transition during the cell cycle. *Oncogene* **24**, 7337-7345 (2005).
344 Achour, M. *et al.* The interaction of the SRA domain of ICBP90 with a novel domain of DNMT1 is involved in the regulation of VEGF gene expression. *Oncogene* **27**, 2187-2197 (2008).

345 Li, Y., Mori, T., Hata, H., Homma, Y. & Kochi, H. NIRF induces G1 arrest and associates with Cdk2. *Biochem Biophys Res Commun* **319**, 464-468 (2004).
346 Pichler, G. *et al.* Cooperative DNA and histone binding by Uhrf2 links the two major repressive epigenetic pathways. *J Cell Biochem* **112**, 2585-2593 (2011).
347 Dhalluin, C. *et al.* Structure and ligand of a histone acetyltransferase bromodomain. *Nature* **399**, 491-496 (1999).
348 Bannister, A. J. *et al.* Selective recognition of methylated lysine 9 on histone H3 by the HP1 chromo domain. *Nature* **410**, 120-124 (2001).
349 Sanchez, R. & Zhou, M. M. The role of human bromodomains in chromatin biology and gene transcription. *Curr Opin Drug Discov Devel* **12**, 659-665 (2009).
350 Charlop-Powers, Z., Zeng, L., Zhang, Q. & Zhou, M. M. Structural insights into selective histone H3 recognition by the human Polybromo bromodomain 2. *Cell Res* **20**, 529-538 (2010).
351 Zhang, Q. *et al.* Biochemical profiling of histone binding selectivity of the yeast bromodomain family. *PLoS One* **5**, e8903 (2010).
352 Zeng, L., Zhang, Q., Gerona-Navarro, G., Moshkina, N. & Zhou, M. M. Structural basis of site-specific histone recognition by the bromodomains of human coactivators PCAF and CBP/p300. *Structure* **16**, 643-652 (2008).
353 Li, H. *et al.* Molecular basis for site-specific read-out of histone H3K4me3 by the BPTF PHD finger of NURF. *Nature* **442**, 91-95 (2006).
354 Palacios, A. *et al.* Solution structure and NMR characterization of the binding to methylated histone tails of the plant homeodomain finger of the tumour suppressor ING4. *FEBS Lett* **580**, 6903-6908 (2006).
355 Pena, P. V. *et al.* Molecular mechanism of histone H3K4me3 recognition by plant homeodomain of ING2. *Nature* **442**, 100-103 (2006).
356 Shi, X. *et al.* ING2 PHD domain links histone H3 lysine 4 methylation to active gene repression. *Nature* **442**, 96-99 (2006).
357 Wysocka, J. *et al.* A PHD finger of NURF couples histone H3 lysine 4 trimethylation with chromatin remodelling. *Nature* **442**, 86-90 (2006).
358 Taverna, S. D. *et al.* Yng1 PHD finger binding to H3 trimethylated at K4 promotes NuA3 HAT activity at K14 of H3 and transcription at a subset of targeted ORFs. *Mol Cell* **24**, 785-796 (2006).
359 Palacios, A. *et al.* Molecular basis of histone H3K4me3 recognition by ING4. *J Biol Chem* **283**, 15956-15964 (2008).
360 Champagne, K. S. *et al.* The crystal structure of the ING5 PHD finger in complex with an H3K4me3 histone peptide. *Proteins* **72**, 1371-1376 (2008).
361 Pena, P. V. *et al.* Histone H3K4me3 binding is required for the DNA repair and apoptotic activities of ING1 tumor suppressor. *J Mol Biol* **380**, 303-312 (2008).
362 Hung, T. *et al.* ING4 mediates crosstalk between histone H3 K4 trimethylation and H3 acetylation to attenuate cellular transformation. *Mol Cell* **33**, 248-256 (2009).
363 Wang, G. G. *et al.* Haematopoietic malignancies caused by dysregulation of a chromatin-binding PHD finger. *Nature* **459**, 847-851 (2009).
364 Lan, F. *et al.* Recognition of unmethylated histone H3 lysine 4 links BHC80 to LSD1-mediated gene repression. *Nature* **448**, 718-722 (2007).
365 Zeng, L. *et al.* Mechanism and regulation of acetylated histone binding by the tandem PHD finger of DPF3b. *Nature* **466**, 258-262 (2010).
366 Lange, M. *et al.* Regulation of muscle development by DPF3, a novel histone acetylation and methylation reader of the BAF chromatin remodeling complex. *Genes Dev* **22**, 2370-2384 (2008).
367 Macdonald, N. *et al.* Molecular basis for the recognition of phosphorylated and phosphoacetylated histone h3 by 14-3-3. *Mol Cell* **20**, 199-211 (2005).

368 Bork, P. *et al.* A superfamily of conserved domains in DNA damage-responsive cell cycle checkpoint proteins. *FASEB J* **11**, 68-76 (1997).
369 Lee, M. S., Edwards, R. A., Thede, G. L. & Glover, J. N. Structure of the BRCT repeat domain of MDC1 and its specificity for the free COOH-terminal end of the gamma-H2AX histone tail. *J Biol Chem* **280**, 32053-32056 (2005).
370 Stucki, M. *et al.* MDC1 directly binds phosphorylated histone H2AX to regulate cellular responses to DNA double-strand breaks. *Cell* **123**, 1213-1226 (2005).
371 Maurer-Stroh, S. *et al.* The Tudor domain 'Royal Family': Tudor, plant Agenet, Chromo, PWWP and MBT domains. *Trends Biochem Sci* **28**, 69-74 (2003).
372 Bernstein, E. *et al.* Mouse polycomb proteins bind differentially to methylated histone H3 and RNA and are enriched in facultative heterochromatin. *Mol Cell Biol* **26**, 2560-2569 (2006).
373 Ball, L. J. *et al.* Structure of the chromatin binding (chromo) domain from mouse modifier protein 1. *EMBO J* **16**, 2473-2481 (1997).
374 Horita, D. A., Ivanova, A. V., Altieri, A. S., Klar, A. J. & Byrd, R. A. Solution structure, domain features, and structural implications of mutants of the chromo domain from the fission yeast histone methyltransferase Clr4. *J Mol Biol* **307**, 861-870 (2001).
375 Murzin, A. G. OB(oligonucleotide/oligosaccharide binding)-fold: common structural and functional solution for non-homologous sequences. *EMBO J* **12**, 861-867 (1993).
376 Lachner, M., O'Carroll, D., Rea, S., Mechtler, K. & Jenuwein, T. Methylation of histone H3 lysine 9 creates a binding site for HP1 proteins. *Nature* **410**, 116-120 (2001).
377 Nakayama, J., Rice, J. C., Strahl, B. D., Allis, C. D. & Grewal, S. I. Role of histone H3 lysine 9 methylation in epigenetic control of heterochromatin assembly. *Science* **292**, 110-113 (2001).
378 Schotta, G. *et al.* A silencing pathway to induce H3-K9 and H4-K20 trimethylation at constitutive heterochromatin. *Genes Dev* **18**, 1251-1262 (2004).
379 Ebert, A. *et al.* Su(var) genes regulate the balance between euchromatin and heterochromatin in Drosophila. *Genes Dev* **18**, 2973-2983 (2004).
380 Jacobs, S. A. & Khorasanizadeh, S. Structure of HP1 chromodomain bound to a lysine 9-methylated histone H3 tail. *Science* **295**, 2080-2083 (2002).
381 Hughes, R. M., Wiggins, K. R., Khorasanizadeh, S. & Waters, M. L. Recognition of trimethyllysine by a chromodomain is not driven by the hydrophobic effect. *Proc Natl Acad Sci U S A* **104**, 11184-11188 (2007).
382 Epstein, H., James, T. C. & Singh, P. B. Cloning and expression of Drosophila HP1 homologs from a mealybug, Planococcus citri. *J Cell Sci* **101 (Pt 2)**, 463-474 (1992).
383 Aasland, R. & Stewart, A. F. The chromo shadow domain, a second chromo domain in heterochromatin-binding protein 1, HP1. *Nucleic Acids Res* **23**, 3168-3173 (1995).
384 Brasher, S. V. *et al.* The structure of mouse HP1 suggests a unique mode of single peptide recognition by the shadow chromo domain dimer. *EMBO J* **19**, 1587-1597 (2000).
385 Cowieson, N. P., Partridge, J. F., Allshire, R. C. & McLaughlin, P. J. Dimerisation of a chromo shadow domain and distinctions from the chromodomain as revealed by structural analysis. *Curr Biol* **10**, 517-525 (2000).
386 Huang, Y., Myers, M. P. & Xu, R. M. Crystal structure of the HP1-EMSY complex reveals an unusual mode of HP1 binding. *Structure* **14**, 703-712 (2006).
387 Mendez, D. L. *et al.* The HP1a disordered C terminus and chromo shadow domain cooperate to select target peptide partners. *Chembiochem* **12**, 1084-1096 (2011).
388 Sun, F. L., Cuaycong, M. H. & Elgin, S. C. Long-range nucleosome ordering is associated with gene silencing in Drosophila melanogaster pericentric heterochromatin. *Mol Cell Biol* **21**, 2867-2879 (2001).
389 Smothers, J. F. & Henikoff, S. The HP1 chromo shadow domain binds a consensus peptide pentamer. *Curr Biol* **10**, 27-30 (2000).

390 Thiru, A. *et al.* Structural basis of HP1/PXVXL motif peptide interactions and HP1 localisation to heterochromatin. *EMBO J* **23**, 489-499 (2004).
391 Fischle, W. *et al.* Molecular basis for the discrimination of repressive methyl-lysine marks in histone H3 by Polycomb and HP1 chromodomains. *Genes Dev* **17**, 1870-1881 (2003).
392 Min, J., Zhang, Y. & Xu, R. M. Structural basis for specific binding of Polycomb chromodomain to histone H3 methylated at Lys 27. *Genes Dev* **17**, 1823-1828 (2003).
393 Flanagan, J. F. *et al.* Double chromodomains cooperate to recognize the methylated histone H3 tail. *Nature* **438**, 1181-1185 (2005).
394 Bardsley, A., McDonald, K. & Boswell, R. E. Distribution of tudor protein in the Drosophila embryo suggests separation of functions based on site of localization. *Development* **119**, 207-219 (1993).
395 Shimojo, H. *et al.* Novel structural and functional mode of a knot essential for RNA binding activity of the Esa1 presumed chromodomain. *J Mol Biol* **378**, 987-1001 (2008).
396 Kim, J. *et al.* Tudor, MBT and chromo domains gauge the degree of lysine methylation. *EMBO Rep* **7**, 397-403 (2006).
397 Taverna, S. D., Li, H., Ruthenburg, A. J., Allis, C. D. & Patel, D. J. How chromatin-binding modules interpret histone modifications: lessons from professional pocket pickers. *Nat Struct Mol Biol* **14**, 1025-1040 (2007).
398 Trojer, P. *et al.* L3MBTL1, a histone-methylation-dependent chromatin lock. *Cell* **129**, 915-928 (2007).
399 Qiu, C., Sawada, K., Zhang, X. & Cheng, X. The PWWP domain of mammalian DNA methyltransferase Dnmt3b defines a new family of DNA-binding folds. *Nat Struct Biol* **9**, 217-224 (2002).
400 Slater, L. M., Allen, M. D. & Bycroft, M. Structural variation in PWWP domains. *J Mol Biol* **330**, 571-576 (2003).
401 Lukasik, S. M. *et al.* High resolution structure of the HDGF PWWP domain: a potential DNA binding domain. *Protein Sci* **15**, 314-323 (2006).
402 Nameki, N. *et al.* Solution structure of the PWWP domain of the hepatoma-derived growth factor family. *Protein Sci* **14**, 756-764 (2005).
403 Laue, K. *et al.* The multidomain protein Brpf1 binds histones and is required for Hox gene expression and segmental identity. *Development* **135**, 1935-1946 (2008).
404 Wysocka, J. *et al.* WDR5 associates with histone H3 methylated at K4 and is essential for H3 K4 methylation and vertebrate development. *Cell* **121**, 859-872 (2005).
405 Couture, J. F., Collazo, E. & Trievel, R. C. Molecular recognition of histone H3 by the WD40 protein WDR5. *Nat Struct Mol Biol* **13**, 698-703 (2006).
406 Han, Z. *et al.* Structural basis for the specific recognition of methylated histone H3 lysine 4 by the WD-40 protein WDR5. *Mol Cell* **22**, 137-144 (2006).
407 Ruthenburg, A. J. *et al.* Histone H3 recognition and presentation by the WDR5 module of the MLL1 complex. *Nat Struct Mol Biol* **13**, 704-712 (2006).
408 Schuetz, A. *et al.* Structural basis for molecular recognition and presentation of histone H3 by WDR5. *EMBO J* **25**, 4245-4252 (2006).
409 Margueron, R. *et al.* Role of the polycomb protein EED in the propagation of repressive histone marks. *Nature* **461**, 762-767 (2009).
410 Xu, C. *et al.* Binding of different histone marks differentially regulates the activity and specificity of polycomb repressive complex 2 (PRC2). *Proc Natl Acad Sci U S A* **107**, 19266-19271 (2010).
411 Song, J. J., Garlick, J. D. & Kingston, R. E. Structural basis of histone H4 recognition by p55. *Genes Dev* **22**, 1313-1318 (2008).
412 Lejon, S. *et al.* Insights into association of the NuRD complex with FOG-1 from the crystal structure of an RbAp48.FOG-1 complex. *J Biol Chem* **286**, 1196-1203 (2011).

413 Hicke, L., Schubert, H. L. & Hill, C. P. Ubiquitin-binding domains. *Nat Rev Mol Cell Biol* **6**, 610-621 (2005).
414 Becker, P. B. & Horz, W. ATP-dependent nucleosome remodeling. *Annu Rev Biochem* **71**, 247-273 (2002).
415 Widom, J. Structure, dynamics, and function of chromatin in vitro. *Annu Rev Biophys Biomol Struct* **27**, 285-327 (1998).
416 Lorch, Y., Davis, B. & Kornberg, R. D. Chromatin remodeling by DNA bending, not twisting. *Proc Natl Acad Sci U S A* **102**, 1329-1332 (2005).
417 Aoyagi, S., Wade, P. A. & Hayes, J. J. Nucleosome sliding induced by the xMi-2 complex does not occur exclusively via a simple twist-diffusion mechanism. *J Biol Chem* **278**, 30562-30568 (2003).
418 Saha, A., Wittmeyer, J. & Cairns, B. R. Chromatin remodeling through directional DNA translocation from an internal nucleosomal site. *Nat Struct Mol Biol* **12**, 747-755 (2005).
419 Cote, J., Quinn, J., Workman, J. L. & Peterson, C. L. Stimulation of GAL4 derivative binding to nucleosomal DNA by the yeast SWI/SNF complex. *Science* **265**, 53-60 (1994).
420 Peterson, C. L., Dingwall, A. & Scott, M. P. Five SWI/SNF gene products are components of a large multisubunit complex required for transcriptional enhancement. *Proc Natl Acad Sci U S A* **91**, 2905-2908 (1994).
421 Cairns, B. R., Kim, Y. J., Sayre, M. H., Laurent, B. C. & Kornberg, R. D. A multisubunit complex containing the SWI1/ADR6, SWI2/SNF2, SWI3, SNF5, and SNF6 gene products isolated from yeast. *Proc Natl Acad Sci U S A* **91**, 1950-1954 (1994).
422 Hirschhorn, J. N., Brown, S. A., Clark, C. D. & Winston, F. Evidence that SNF2/SWI2 and SNF5 activate transcription in yeast by altering chromatin structure. *Genes Dev* **6**, 2288-2298 (1992).
423 Kadam, S. *et al.* Functional selectivity of recombinant mammalian SWI/SNF subunits. *Genes Dev* **14**, 2441-2451 (2000).
424 Tamkun, J. W. *et al.* brahma: a regulator of Drosophila homeotic genes structurally related to the yeast transcriptional activator SNF2/SWI2. *Cell* **68**, 561-572 (1992).
425 Dingwall, A. K. *et al.* The Drosophila snr1 and brm proteins are related to yeast SWI/SNF proteins and are components of a large protein complex. *Mol Biol Cell* **6**, 777-791 (1995).
426 Papoulas, O. *et al.* The Drosophila trithorax group proteins BRM, ASH1 and ASH2 are subunits of distinct protein complexes. *Development* **125**, 3955-3966 (1998).
427 Armstrong, J. A. *et al.* The Drosophila BRM complex facilitates global transcription by RNA polymerase II. *EMBO J* **21**, 5245-5254 (2002).
428 Reyes, J. C. *et al.* Altered control of cellular proliferation in the absence of mammalian brahma (SNF2alpha). *EMBO J* **17**, 6979-6991 (1998).
429 Bultman, S. *et al.* A Brg1 null mutation in the mouse reveals functional differences among mammalian SWI/SNF complexes. *Mol Cell* **6**, 1287-1295 (2000).
430 Reisman, D. N., Sciarrotta, J., Bouldin, T. W., Weissman, B. E. & Funkhouser, W. K. The expression of the SWI/SNF ATPase subunits BRG1 and BRM in normal human tissues. *Appl Immunohistochem Mol Morphol* **13**, 66-74 (2005).
431 Elfring, L. K., Deuring, R., McCallum, C. M., Peterson, C. L. & Tamkun, J. W. Identification and characterization of Drosophila relatives of the yeast transcriptional activator SNF2/SWI2. *Mol Cell Biol* **14**, 2225-2234 (1994).
432 Aasland, R., Stewart, A. F. & Gibson, T. The SANT domain: a putative DNA-binding domain in the SWI-SNF and ADA complexes, the transcriptional co-repressor N-CoR and TFIIIB. *Trends Biochem Sci* **21**, 87-88 (1996).
433 Boyer, L. A. *et al.* Essential role for the SANT domain in the functioning of multiple chromatin remodeling enzymes. *Mol Cell* **10**, 935-942 (2002).
434 Grune, T. *et al.* Crystal structure and functional analysis of a nucleosome recognition module of the remodeling factor ISWI. *Mol Cell* **12**, 449-460 (2003).

435 Tsukiyama, T., Daniel, C., Tamkun, J. & Wu, C. ISWI, a member of the SWI2/SNF2 ATPase family, encodes the 140 kDa subunit of the nucleosome remodeling factor. *Cell* **83**, 1021-1026 (1995).
436 Ito, T., Bulger, M., Pazin, M. J., Kobayashi, R. & Kadonaga, J. T. ACF, an ISWI-containing and ATP-utilizing chromatin assembly and remodeling factor. *Cell* **90**, 145-155 (1997).
437 Varga-Weisz, P. D. *et al.* Chromatin-remodelling factor CHRAC contains the ATPases ISWI and topoisomerase II. *Nature* **388**, 598-602 (1997).
438 Deuring, R. *et al.* The ISWI chromatin-remodeling protein is required for gene expression and the maintenance of higher order chromatin structure in vivo. *Mol Cell* **5**, 355-365 (2000).
439 Xiao, H. *et al.* Dual functions of largest NURF subunit NURF301 in nucleosome sliding and transcription factor interactions. *Mol Cell* **8**, 531-543 (2001).
440 Ito, T. *et al.* ACF consists of two subunits, Acf1 and ISWI, that function cooperatively in the ATP-dependent catalysis of chromatin assembly. *Genes Dev* **13**, 1529-1539 (1999).
441 Fyodorov, D. V., Blower, M. D., Karpen, G. H. & Kadonaga, J. T. Acf1 confers unique activities to ACF/CHRAC and promotes the formation rather than disruption of chromatin in vivo. *Genes Dev* **18**, 170-183 (2004).
442 Corona, D. F. *et al.* Two histone fold proteins, CHRAC-14 and CHRAC-16, are developmentally regulated subunits of chromatin accessibility complex (CHRAC). *EMBO J* **19**, 3049-3059 (2000).
443 Kukimoto, I., Elderkin, S., Grimaldi, M., Oelgeschlager, T. & Varga-Weisz, P. D. The histone-fold protein complex CHRAC-15/17 enhances nucleosome sliding and assembly mediated by ACF. *Mol Cell* **13**, 265-277 (2004).
444 Tran, H. G., Steger, D. J., Iyer, V. R. & Johnson, A. D. The chromo domain protein chd1p from budding yeast is an ATP-dependent chromatin-modifying factor. *EMBO J* **19**, 2323-2331 (2000).
445 Marfella, C. G. & Imbalzano, A. N. The Chd family of chromatin remodelers. *Mutat Res* **618**, 30-40 (2007).
446 Kelley, D. E., Stokes, D. G. & Perry, R. P. CHD1 interacts with SSRP1 and depends on both its chromodomain and its ATPase/helicase-like domain for proper association with chromatin. *Chromosoma* **108**, 10-25 (1999).
447 Stokes, D. G., Tartof, K. D. & Perry, R. P. CHD1 is concentrated in interbands and puffed regions of Drosophila polytene chromosomes. *Proc Natl Acad Sci U S A* **93**, 7137-7142 (1996).
448 Martin, D. M. Chromatin remodeling in development and disease: focus on CHD7. *PLoS Genet* **6**, e1001010 (2010).
449 Bergman, J. E. *et al.* CHD7 mutations and CHARGE syndrome: the clinical implications of an expanding phenotype. *J Med Genet* **48**, 334-342 (2011).
450 Shur, I., Socher, R. & Benayahu, D. In vivo association of CReMM/CHD9 with promoters in osteogenic cells. *J Cell Physiol* **207**, 374-378 (2006).
451 Ebbert, R., Birkmann, A. & Schuller, H. J. The product of the SNF2/SWI2 paralogue INO80 of Saccharomyces cerevisiae required for efficient expression of various yeast structural genes is part of a high-molecular-weight protein complex. *Mol Microbiol* **32**, 741-751 (1999).
452 Jin, J. *et al.* A mammalian chromatin remodeling complex with similarities to the yeast INO80 complex. *J Biol Chem* **280**, 41207-41212 (2005).
453 Shen, X., Mizuguchi, G., Hamiche, A. & Wu, C. A chromatin remodelling complex involved in transcription and DNA processing. *Nature* **406**, 541-544 (2000).
454 Kobor, M. S. *et al.* A protein complex containing the conserved Swi2/Snf2-related ATPase Swr1p deposits histone variant H2A.Z into euchromatin. *PLoS Biol* **2**, E131 (2004).
455 Krogan, N. J. *et al.* Regulation of chromosome stability by the histone H2A variant Htz1, the Swr1 chromatin remodeling complex, and the histone acetyltransferase NuA4. *Proc Natl Acad Sci U S A* **101**, 13513-13518 (2004).

456 Mizuguchi, G. *et al.* ATP-driven exchange of histone H2AZ variant catalyzed by SWR1 chromatin remodeling complex. *Science* **303**, 343-348 (2004).
457 Tong, J. K., Hassig, C. A., Schnitzler, G. R., Kingston, R. E. & Schreiber, S. L. Chromatin deacetylation by an ATP-dependent nucleosome remodelling complex. *Nature* **395**, 917-921 (1998).
458 Wade, P. A., Jones, P. L., Vermaak, D. & Wolffe, A. P. A multiple subunit Mi-2 histone deacetylase from Xenopus laevis cofractionates with an associated Snf2 superfamily ATPase. *Curr Biol* **8**, 843-846 (1998).
459 Zhang, Y., LeRoy, G., Seelig, H. P., Lane, W. S. & Reinberg, D. The dermatomyositis-specific autoantigen Mi2 is a component of a complex containing histone deacetylase and nucleosome remodeling activities. *Cell* **95**, 279-289 (1998).
460 Auger, A. *et al.* Eaf1 is the platform for NuA4 molecular assembly that evolutionarily links chromatin acetylation to ATP-dependent exchange of histone H2A variants. *Mol Cell Biol* **28**, 2257-2270 (2008).
461 Doyon, Y. & Cote, J. The highly conserved and multifunctional NuA4 HAT complex. *Curr Opin Genet Dev* **14**, 147-154 (2004).
462 Verreault, A., Kaufman, P. D., Kobayashi, R. & Stillman, B. Nucleosomal DNA regulates the core-histone-binding subunit of the human Hat1 acetyltransferase. *Curr Biol* **8**, 96-108 (1998).
463 Seelig, H. P. *et al.* The major dermatomyositis-specific Mi-2 autoantigen is a presumed helicase involved in transcriptional activation. *Arthritis Rheum* **38**, 1389-1399 (1995).
464 Ge, Q., Nilasena, D. S., O'Brien, C. A., Frank, M. B. & Targoff, I. N. Molecular analysis of a major antigenic region of the 240-kD protein of Mi-2 autoantigen. *J Clin Invest* **96**, 1730-1737 (1995).
465 Callen, J. P. & Wortmann, R. L. Dermatomyositis. *Clin Dermatol* **24**, 363-373 (2006).
466 Hill, C. L. *et al.* Frequency of specific cancer types in dermatomyositis and polymyositis: a population-based study. *Lancet* **357**, 96-100 (2001).
467 Woodage, T., Basrai, M. A., Baxevanis, A. D., Hieter, P. & Collins, F. S. Characterization of the CHD family of proteins. *Proc Natl Acad Sci U S A* **94**, 11472-11477 (1997).
468 Eisen, J. A., Sweder, K. S. & Hanawalt, P. C. Evolution of the SNF2 family of proteins: subfamilies with distinct sequences and functions. *Nucleic Acids Res* **23**, 2715-2723 (1995).
469 Brehm, A. *et al.* dMi-2 and ISWI chromatin remodelling factors have distinct nucleosome binding and mobilization properties. *EMBO J* **19**, 4332-4341 (2000).
470 Wang, H. B. & Zhang, Y. Mi2, an auto-antigen for dermatomyositis, is an ATP-dependent nucleosome remodeling factor. *Nucleic Acids Res* **29**, 2517-2521 (2001).
471 Mansfield, R. E. *et al.* Plant homeodomain (PHD) fingers of CHD4 are histone H3-binding modules with preference for unmodified H3K4 and methylated H3K9. *J Biol Chem* **286**, 11779-11791 (2011).
472 Kwan, A. H. *et al.* Engineering a protein scaffold from a PHD finger. *Structure* **11**, 803-813 (2003).
473 Musselman, C. A. *et al.* Bivalent recognition of nucleosomes by the tandem PHD fingers of the CHD4 ATPase is required for CHD4-mediated repression. *Proc Natl Acad Sci U S A* **109**, 787-792 (2012).
474 Musselman, C. A. *et al.* Binding of the CHD4 PHD2 finger to histone H3 is modulated by covalent modifications. *Biochem J* **423**, 179-187 (2009).
475 Ivanov, A. V. *et al.* PHD domain-mediated E3 ligase activity directs intramolecular sumoylation of an adjacent bromodomain required for gene silencing. *Mol Cell* **28**, 823-837 (2007).
476 Goodarzi, A. A., Kurka, T. & Jeggo, P. A. KAP-1 phosphorylation regulates CHD3 nucleosome remodeling during the DNA double-strand break response. *Nat Struct Mol Biol* **18**, 831-839 (2011).

477 Lee, D. H. *et al.* Phosphoproteomic analysis reveals that PP4 dephosphorylates KAP-1 impacting the DNA damage response. *EMBO J* **31**, 2403-2415 (2012).
478 von Zelewsky, T. *et al.* The C. elegans Mi-2 chromatin-remodelling proteins function in vulval cell fate determination. *Development* **127**, 5277-5284 (2000).
479 Fukaki, H., Taniguchi, N. & Tasaka, M. PICKLE is required for SOLITARY-ROOT/IAA14-mediated repression of ARF7 and ARF19 activity during Arabidopsis lateral root initiation. *Plant J* **48**, 380-389 (2006).
480 Dovey, O. M. *et al.* Histone deacetylase 1 and 2 are essential for normal T-cell development and genomic stability in mice. *Blood* **121**, 1335-1344 (2013).
481 Zhang, Y. *et al.* Analysis of the NuRD subunits reveals a histone deacetylase core complex and a connection with DNA methylation. *Genes Dev* **13**, 1924-1935 (1999).
482 Segre, C. V. & Chiocca, S. Regulating the regulators: the post-translational code of class I HDAC1 and HDAC2. *J Biomed Biotechnol* **2011**, 690848 (2011).
483 Somoza, J. R. *et al.* Structural snapshots of human HDAC8 provide insights into the class I histone deacetylases. *Structure* **12**, 1325-1334 (2004).
484 Vannini, A. *et al.* Crystal structure of a eukaryotic zinc-dependent histone deacetylase, human HDAC8, complexed with a hydroxamic acid inhibitor. *Proc Natl Acad Sci U S A* **101**, 15064-15069 (2004).
485 Bressi, J. C. *et al.* Exploration of the HDAC2 foot pocket: Synthesis and SAR of substituted N-(2-aminophenyl)benzamides. *Bioorg Med Chem Lett* **20**, 3142-3145 (2010).
486 Lauffer, B. E. *et al.* Histone deacetylase (HDAC) inhibitor kinetic rate constants correlate with cellular histone acetylation but not transcription and cell viability. *J Biol Chem* **288**, 26926-26943 (2013).
487 Montgomery, R. L. *et al.* Histone deacetylases 1 and 2 redundantly regulate cardiac morphogenesis, growth, and contractility. *Genes Dev* **21**, 1790-1802 (2007).
488 Yamaguchi, J. *et al.* Histone deacetylase inhibitor (SAHA) and repression of EZH2 synergistically inhibit proliferation of gallbladder carcinoma. *Cancer Sci* **101**, 355-362 (2010).
489 Zupkovitz, G. *et al.* Negative and positive regulation of gene expression by mouse histone deacetylase 1. *Mol Cell Biol* **26**, 7913-7928 (2006).
490 Wang, Z. *et al.* Genome-wide mapping of HATs and HDACs reveals distinct functions in active and inactive genes. *Cell* **138**, 1019-1031 (2009).
491 Kidder, B. L. & Palmer, S. HDAC1 regulates pluripotency and lineage specific transcriptional networks in embryonic and trophoblast stem cells. *Nucleic Acids Res* **40**, 2925-2939 (2012).
492 Kurdistani, S. K., Robyr, D., Tavazoie, S. & Grunstein, M. Genome-wide binding map of the histone deacetylase Rpd3 in yeast. *Nat Genet* **31**, 248-254 (2002).
493 Lagger, G. *et al.* Essential function of histone deacetylase 1 in proliferation control and CDK inhibitor repression. *EMBO J* **21**, 2672-2681 (2002).
494 Senese, S. *et al.* Role for histone deacetylase 1 in human tumor cell proliferation. *Mol Cell Biol* **27**, 4784-4795 (2007).
495 Wilting, R. H. *et al.* Overlapping functions of Hdac1 and Hdac2 in cell cycle regulation and haematopoiesis. *EMBO J* **29**, 2586-2597 (2010).
496 Zupkovitz, G. *et al.* The cyclin-dependent kinase inhibitor p21 is a crucial target for histone deacetylase 1 as a regulator of cellular proliferation. *Mol Cell Biol* **30**, 1171-1181 (2010).
497 Marks, P. A. & Xu, W. S. Histone deacetylase inhibitors: Potential in cancer therapy. *J Cell Biochem* **107**, 600-608 (2009).
498 Rosato, R. R., Almenara, J. A. & Grant, S. The histone deacetylase inhibitor MS-275 promotes differentiation or apoptosis in human leukemia cells through a process regulated by generation of reactive oxygen species and induction of p21CIP1/WAF1 1. *Cancer Res* **63**, 3637-3645 (2003).
499 Dovey, O. M., Foster, C. T. & Cowley, S. M. Histone deacetylase 1 (HDAC1), but not HDAC2, controls embryonic stem cell differentiation. *Proc Natl Acad Sci U S A* **107**, 8242-8247 (2010).

500 Dovey, O. M., Foster, C. T. & Cowley, S. M. Emphasizing the positive: A role for histone deacetylases in transcriptional activation. *Cell Cycle* **9**, 2700-2701 (2010).
501 LeBoeuf, M. *et al.* Hdac1 and Hdac2 act redundantly to control p63 and p53 functions in epidermal progenitor cells. *Dev Cell* **19**, 807-818 (2010).
502 Ma, P., Pan, H., Montgomery, R. L., Olson, E. N. & Schultz, R. M. Compensatory functions of histone deacetylase 1 (HDAC1) and HDAC2 regulate transcription and apoptosis during mouse oocyte development. *Proc Natl Acad Sci U S A* **109**, E481-489 (2012).
503 Pencil, S. D., Toh, Y. & Nicolson, G. L. Candidate metastasis-associated genes of the rat 13762NF mammary adenocarcinoma. *Breast Cancer Res Treat* **25**, 165-174 (1993).
504 Bowen, N. J., Fujita, N., Kajita, M. & Wade, P. A. Mi-2/NuRD: multiple complexes for many purposes. *Biochim Biophys Acta* **1677**, 52-57 (2004).
505 Yaguchi, M. *et al.* Identification and characterization of the variants of metastasis-associated protein 1 generated following alternative splicing. *Biochim Biophys Acta* **1732**, 8-14 (2005).
506 Yu, J., Li, Y., Ishizuka, T., Guenther, M. G. & Lazar, M. A. A SANT motif in the SMRT corepressor interprets the histone code and promotes histone deacetylation. *EMBO J* **22**, 3403-3410 (2003).
507 Chambers, A. L., Pearl, L. H., Oliver, A. W. & Downs, J. A. The BAH domain of Rsc2 is a histone H3 binding domain. *Nucleic Acids Res* **41**, 9168-9182 (2013).
508 Kuo, A. J. *et al.* The BAH domain of ORC1 links H4K20me2 to DNA replication licensing and Meier-Gorlin syndrome. *Nature* **484**, 115-119 (2012).
509 Kumar, R. *et al.* A naturally occurring MTA1 variant sequesters oestrogen receptor-alpha in the cytoplasm. *Nature* **418**, 654-657 (2002).
510 Kleene, R., Classen, B., Zdzieblo, J. & Schrader, M. SH3 binding sites of ZG29p mediate an interaction with amylase and are involved in condensation-sorting in the exocrine rat pancreas. *Biochemistry* **39**, 9893-9900 (2000).
511 Zhang, X. Y. *et al.* Metastasis-associated protein 1 (MTA1) is an essential downstream effector of the c-MYC oncoprotein. *Proc Natl Acad Sci U S A* **102**, 13968-13973 (2005).
512 Yoo, Y. G., Kong, G. & Lee, M. O. Metastasis-associated protein 1 enhances stability of hypoxia-inducible factor-1alpha protein by recruiting histone deacetylase 1. *EMBO J* **25**, 1231-1241 (2006).
513 Kumar, R. Another tie that binds the MTA family to breast cancer. *Cell* **113**, 142-143 (2003).
514 Manavathi, B. & Kumar, R. Metastasis tumor antigens, an emerging family of multifaceted master coregulators. *J Biol Chem* **282**, 1529-1533 (2007).
515 Fujita, N. *et al.* MTA3, a Mi-2/NuRD complex subunit, regulates an invasive growth pathway in breast cancer. *Cell* **113**, 207-219 (2003).
516 Fujita, N., Kajita, M., Taysavang, P. & Wade, P. A. Hormonal regulation of metastasis-associated protein 3 transcription in breast cancer cells. *Mol Endocrinol* **18**, 2937-2949 (2004).
517 Mishra, S. K. *et al.* Upstream determinants of estrogen receptor-alpha regulation of metastatic tumor antigen 3 pathway. *J Biol Chem* **279**, 32709-32715 (2004).
518 Zhang, H., Singh, R. R., Talukder, A. H. & Kumar, R. Metastatic tumor antigen 3 is a direct corepressor of the Wnt4 pathway. *Genes Dev* **20**, 2943-2948 (2006).
519 Millard, C. J., Fairall, L. & Schwabe, J. W. Towards an understanding of the structure and function of MTA1. *Cancer Metastasis Rev* (2014).
520 Watson, P. J., Fairall, L., Santos, G. M. & Schwabe, J. W. Structure of HDAC3 bound to co-repressor and inositol tetraphosphate. *Nature* **481**, 335-340 (2012).
521 Wade, P. A. *et al.* Mi-2 complex couples DNA methylation to chromatin remodelling and histone deacetylation. *Nat Genet* **23**, 62-66 (1999).
522 Le Guezennec, X. *et al.* MBD2/NuRD and MBD3/NuRD, two distinct complexes with different biochemical and functional properties. *Mol Cell Biol* **26**, 843-851 (2006).

523 Saito, M. & Ishikawa, F. The mCpG-binding domain of human MBD3 does not bind to mCpG but interacts with NuRD/Mi2 components HDAC1 and MTA2. *J Biol Chem* **277**, 35434-35439 (2002).
524 Fraga, M. F. *et al.* The affinity of different MBD proteins for a specific methylated locus depends on their intrinsic binding properties. *Nucleic Acids Res* **31**, 1765-1774 (2003).
525 Hendrich, B. & Tweedie, S. The methyl-CpG binding domain and the evolving role of DNA methylation in animals. *Trends Genet* **19**, 269-277 (2003).
526 Feng, Q. & Zhang, Y. The MeCP1 complex represses transcription through preferential binding, remodeling, and deacetylating methylated nucleosomes. *Genes Dev* **15**, 827-832 (2001).
527 Yildirim, O. *et al.* Mbd3/NURD complex regulates expression of 5-hydroxymethylcytosine marked genes in embryonic stem cells. *Cell* **147**, 1498-1510 (2011).
528 Hashimoto, H. *et al.* Recognition and potential mechanisms for replication and erasure of cytosine hydroxymethylation. *Nucleic Acids Res* **40**, 4841-4849 (2012).
529 Baubec, T., Ivanek, R., Lienert, F. & Schubeler, D. Methylation-dependent and -independent genomic targeting principles of the MBD protein family. *Cell* **153**, 480-492 (2013).
530 Cramer, J. M. *et al.* Probing the dynamic distribution of bound states for methylcytosine-binding domains on DNA. *J Biol Chem* **289**, 1294-1302 (2014).
531 Ho, K. L. *et al.* MeCP2 binding to DNA depends upon hydration at methyl-CpG. *Mol Cell* **29**, 525-531 (2008).
532 Ohki, I., Shimotake, N., Fujita, N., Nakao, M. & Shirakawa, M. Solution structure of the methyl-CpG-binding domain of the methylation-dependent transcriptional repressor MBD1. *EMBO J* **18**, 6653-6661 (1999).
533 Otani, J. *et al.* Structural basis of the versatile DNA recognition ability of the methyl-CpG binding domain of methyl-CpG binding domain protein 4. *J Biol Chem* **288**, 6351-6362 (2013).
534 Scarsdale, J. N., Webb, H. D., Ginder, G. D. & Williams, D. C., Jr. Solution structure and dynamic analysis of chicken MBD2 methyl binding domain bound to a target-methylated DNA sequence. *Nucleic Acids Res* **39**, 6741-6752 (2011).
535 Wakefield, R. I. *et al.* The solution structure of the domain from MeCP2 that binds to methylated DNA. *J Mol Biol* **291**, 1055-1065 (1999).
536 Gnanapragasam, M. N. *et al.* p66Alpha-MBD2 coiled-coil interaction and recruitment of Mi-2 are critical for globin gene silencing by the MBD2-NuRD complex. *Proc Natl Acad Sci U S A* **108**, 7487-7492 (2011).
537 Rais, Y. *et al.* Deterministic direct reprogramming of somatic cells to pluripotency. *Nature* **502**, 65-70 (2013).
538 Dos Santos, R. L. *et al.* MBD3/NuRD Facilitates Induction of Pluripotency in a Context-Dependent Manner. *Cell Stem Cell* **15**, 102-110 (2014).
539 Cukier, H. N. *et al.* Novel variants identified in methyl-CpG-binding domain genes in autistic individuals. *Neurogenetics* **11**, 291-303 (2010).
540 Qian, Y. W. *et al.* A retinoblastoma-binding protein related to a negative regulator of Ras in yeast. *Nature* **364**, 648-652 (1993).
541 Qian, Y. W. & Lee, E. Y. Dual retinoblastoma-binding proteins with properties related to a negative regulator of ras in yeast. *J Biol Chem* **270**, 25507-25513 (1995).
542 Nicolas, E. *et al.* RbAp48 belongs to the histone deacetylase complex that associates with the retinoblastoma protein. *J Biol Chem* **275**, 9797-9804 (2000).
543 Zhang, Y., Iratni, R., Erdjument-Bromage, H., Tempst, P. & Reinberg, D. Histone deacetylases and SAP18, a novel polypeptide, are components of a human Sin3 complex. *Cell* **89**, 357-364 (1997).
544 Knoepfler, P. S. & Eisenman, R. N. Sin meets NuRD and other tails of repression. *Cell* **99**, 447-450 (1999).

545 Ahringer, J. NuRD and SIN3 histone deacetylase complexes in development. *Trends Genet* **16**, 351-356 (2000).
546 Parthun, M. R. Hat1: the emerging cellular roles of a type B histone acetyltransferase. *Oncogene* **26**, 5319-5328 (2007).
547 Benson, L. J. *et al.* Properties of the type B histone acetyltransferase Hat1: H4 tail interaction, site preference, and involvement in DNA repair. *J Biol Chem* **282**, 836-842 (2007).
548 Hoek, M. & Stillman, B. Chromatin assembly factor 1 is essential and couples chromatin assembly to DNA replication in vivo. *Proc Natl Acad Sci U S A* **100**, 12183-12188 (2003).
549 Korenjak, M. *et al.* Native E2F/RBF complexes contain Myb-interacting proteins and repress transcription of developmentally controlled E2F target genes. *Cell* **119**, 181-193 (2004).
550 Kuzmichev, A., Jenuwein, T., Tempst, P. & Reinberg, D. Different EZH2-containing complexes target methylation of histone H1 or nucleosomal histone H3. *Mol Cell* **14**, 183-193 (2004).
551 Martinez-Balbas, M. A., Tsukiyama, T., Gdula, D. & Wu, C. Drosophila NURF-55, a WD repeat protein involved in histone metabolism. *Proc Natl Acad Sci U S A* **95**, 132-137 (1998).
552 Xu, C. & Min, J. Structure and function of WD40 domain proteins. *Protein Cell* **2**, 202-214 (2011).
553 Alqarni, S. S. *et al.* Insight into the architecture of the NuRD complex: Structure of the RbAp48-MTA1 sub-complex. *J Biol Chem* (2014).
554 Zhang, W. *et al.* Structural plasticity of histones H3-H4 facilitates their allosteric exchange between RbAp48 and ASF1. *Nat Struct Mol Biol* **20**, 29-35 (2013).
555 Guan, L. S., Li, G. C., Chen, C. C., Liu, L. Q. & Wang, Z. Y. Rb-associated protein 46 (RbAp46) suppresses the tumorigenicity of adenovirus-transformed human embryonic kidney 293 cells. *Int J Cancer* **93**, 333-338 (2001).
556 Zhang, T. F., Yu, S. Q., Deuel, T. F. & Wang, Z. Y. Constitutive expression of Rb associated protein 46 (RbAp46) reverts transformed phenotypes of breast cancer cells. *Anticancer Res* **23**, 3735-3740 (2003).
557 Thakur, A. *et al.* Aberrant expression of X-linked genes RbAp46, Rsk4, and Cldn2 in breast cancer. *Mol Cancer Res* **5**, 171-181 (2007).
558 Creekmore, A. L. *et al.* The role of retinoblastoma-associated proteins 46 and 48 in estrogen receptor alpha mediated gene expression. *Mol Cell Endocrinol* **291**, 79-86 (2008).
559 Kong, L. *et al.* RbAp48 is a critical mediator controlling the transforming activity of human papillomavirus type 16 in cervical cancer. *J Biol Chem* **282**, 26381-26391 (2007).
560 Ginger, M. R., Gonzalez-Rimbau, M. F., Gay, J. P. & Rosen, J. M. Persistent changes in gene expression induced by estrogen and progesterone in the rat mammary gland. *Mol Endocrinol* **15**, 1993-2009 (2001).
561 Zheng, L. *et al.* Radiation-inducible protein RbAp48 contributes to radiosensitivity of cervical cancer cells. *Gynecol Oncol* **130**, 601-608 (2013).
562 Pavlopoulos, E. *et al.* Molecular mechanism for age-related memory loss: the histone-binding protein RbAp48. *Sci Transl Med* **5**, 200ra115 (2013).
563 Brackertz, M., Boeke, J., Zhang, R. & Renkawitz, R. Two highly related p66 proteins comprise a new family of potent transcriptional repressors interacting with MBD2 and MBD3. *J Biol Chem* **277**, 40958-40966 (2002).
564 Feng, Q. *et al.* Identification and functional characterization of the p66/p68 components of the MeCP1 complex. *Mol Cell Biol* **22**, 536-546 (2002).
565 Brackertz, M., Gong, Z., Leers, J. & Renkawitz, R. p66alpha and p66beta of the Mi-2/NuRD complex mediate MBD2 and histone interaction. *Nucleic Acids Res* **34**, 397-406 (2006).
566 Gong, Z., Brackertz, M. & Renkawitz, R. SUMO modification enhances p66-mediated transcriptional repression of the Mi-2/NuRD complex. *Mol Cell Biol* **26**, 4519-4528 (2006).

567 Tsuji, T. *et al.* Cloning, mapping, expression, function, and mutation analyses of the human ortholog of the hamster putative tumor suppressor gene Doc-1. *J Biol Chem* **273**, 6704-6709 (1998).
568 Yuan, Z., Sotsky Kent, T. & Weber, T. K. Differential expression of DOC-1 in microsatellite-unstable human colorectal cancer. *Oncogene* **22**, 6304-6310 (2003).
569 Spruijt, C. G. *et al.* CDK2AP1/DOC-1 is a bona fide subunit of the Mi-2/NuRD complex. *Mol Biosyst* **6**, 1700-1706 (2010).
570 Shintani, S. *et al.* p12(DOC-1) is a novel cyclin-dependent kinase 2-associated protein. *Mol Cell Biol* **20**, 6300-6307 (2000).
571 Tyler, J. K. & Kadonaga, J. T. The "dark side" of chromatin remodeling: repressive effects on transcription. *Cell* **99**, 443-446 (1999).
572 Toh, Y., Pencil, S. D. & Nicolson, G. L. A novel candidate metastasis-associated gene, mta1, differentially expressed in highly metastatic mammary adenocarcinoma cell lines. cDNA cloning, expression, and protein analyses. *J Biol Chem* **269**, 22958-22963 (1994).
573 Srinivasan, R., Mager, G. M., Ward, R. M., Mayer, J. & Svaren, J. NAB2 represses transcription by interacting with the CHD4 subunit of the nucleosome remodeling and deacetylase (NuRD) complex. *J Biol Chem* **281**, 15129-15137 (2006).
574 Aguilera, C. *et al.* c-Jun N-terminal phosphorylation antagonises recruitment of the Mbd3/NuRD repressor complex. *Nature* **469**, 231-235 (2011).
575 Gao, H. *et al.* Opposing effects of SWI/SNF and Mi-2/NuRD chromatin remodeling complexes on epigenetic reprogramming by EBF and Pax5. *Proc Natl Acad Sci U S A* **106**, 11258-11263 (2009).
576 Cismasiu, V. B. *et al.* BCL11B functionally associates with the NuRD complex in T lymphocytes to repress targeted promoter. *Oncogene* **24**, 6753-6764 (2005).
577 Cismasiu, V. B. *et al.* BCL11B is a general transcriptional repressor of the HIV-1 long terminal repeat in T lymphocytes through recruitment of the NuRD complex. *Virology* **380**, 173-181 (2008).
578 Nishioka, K. *et al.* Set9, a novel histone H3 methyltransferase that facilitates transcription by precluding histone tail modifications required for heterochromatin formation. *Genes Dev* **16**, 479-489 (2002).
579 Nishioka, K. *et al.* PR-Set7 is a nucleosome-specific methyltransferase that modifies lysine 20 of histone H4 and is associated with silent chromatin. *Mol Cell* **9**, 1201-1213 (2002).
580 Zegerman, P., Canas, B., Pappin, D. & Kouzarides, T. Histone H3 lysine 4 methylation disrupts binding of nucleosome remodeling and deacetylase (NuRD) repressor complex. *J Biol Chem* **277**, 11621-11624 (2002).
581 Williams, C. J. *et al.* The chromatin remodeler Mi-2beta is required for CD4 expression and T cell development. *Immunity* **20**, 719-733 (2004).
582 Yoshida, T. *et al.* The role of the chromatin remodeler Mi-2beta in hematopoietic stem cell self-renewal and multilineage differentiation. *Genes Dev* **22**, 1174-1189 (2008).
583 Polo, S. E., Kaidi, A., Baskcomb, L., Galanty, Y. & Jackson, S. P. Regulation of DNA-damage responses and cell-cycle progression by the chromatin remodelling factor CHD4. *EMBO J* **29**, 3130-3139 (2010).
584 Rogakou, E. P., Pilch, D. R., Orr, A. H., Ivanova, V. S. & Bonner, W. M. DNA double-stranded breaks induce histone H2AX phosphorylation on serine 139. *J Biol Chem* **273**, 5858-5868 (1998).
585 Luijsterburg, M. S. *et al.* A new non-catalytic role for ubiquitin ligase RNF8 in unfolding higher-order chromatin structure. *EMBO J* **31**, 2511-2527 (2012).
586 Helbling Chadwick, L., Chadwick, B. P., Jaye, D. L. & Wade, P. A. The Mi-2/NuRD complex associates with pericentromeric heterochromatin during S phase in rapidly proliferating lymphoid cells. *Chromosoma* **118**, 445-457 (2009).

587 Luger, K., Rechsteiner, T. J. & Richmond, T. J. Preparation of nucleosome core particle from recombinant histones. *Methods Enzymol* **304**, 3-19 (1999).
588 Lederberg, J. Cell genetics and hereditary symbiosis. *Physiol Rev* **32**, 403-430 (1952).
589 Bolivar, F., Rodriguez, R. L., Betlach, M. C. & Boyer, H. W. Construction and characterization of new cloning vehicles. I. Ampicillin-resistant derivatives of the plasmid pMB9. *Gene* **2**, 75-93 (1977).
590 Bolivar, F. *et al.* Construction and characterization of new cloning vehicles. II. A multipurpose cloning system. *Gene* **2**, 95-113 (1977).
591 Vieira, J. & Messing, J. The pUC plasmids, an M13mp7-derived system for insertion mutagenesis and sequencing with synthetic universal primers. *Gene* **19**, 259-268 (1982).
592 Hartley, J. L. Use of the gateway system for protein expression in multiple hosts. *Curr Protoc Protein Sci* **Chapter 5**, Unit 5 17 (2003).
593 Walhout, A. J. *et al.* GATEWAY recombinational cloning: application to the cloning of large numbers of open reading frames or ORFeomes. *Methods Enzymol* **328**, 575-592 (2000).
594 Luckow, V. A. Baculovirus systems for the expression of human gene products. *Curr Opin Biotechnol* **4**, 564-572 (1993).
595 Li, M. Z. & Elledge, S. J. SLIC: a method for sequence- and ligation-independent cloning. *Methods Mol Biol* **852**, 51-59 (2012).
596 Li, M. Z. & Elledge, S. J. Harnessing homologous recombination in vitro to generate recombinant DNA via SLIC. *Nat Methods* **4**, 251-256 (2007).
597 Bertani, G. Studies on lysogenesis. I. The mode of phage liberation by lysogenic Escherichia coli. *J Bacteriol* **62**, 293-300 (1951).
598 Payne, J. W. & Gilvarg, C. Size restriction on peptide utilization in Escherichia coli. *J Biol Chem* **243**, 6291-6299 (1968).
599 Sezonov, G., Joseleau-Petit, D. & D'Ari, R. Escherichia coli physiology in Luria-Bertani broth. *J Bacteriol* **189**, 8746-8749 (2007).
600 Kinjo, K. & Nikaido, H. [Successive assimilation of amino acids. (2)]. *Nihon Saikingaku Zasshi* **16**, 926-931 (1961).
601 Studier, F. W. Protein production by auto-induction in high density shaking cultures. *Protein Expr Purif* **41**, 207-234 (2005).
602 Miller, L. K., Jewell, J. E. & Browne, D. Baculovirus induction of a DNA polymerase. *J Virol* **40**, 305-308 (1981).
603 Smith, G. E. *et al.* Modification and secretion of human interleukin 2 produced in insect cells by a baculovirus expression vector. *Proc Natl Acad Sci U S A* **82**, 8404-8408 (1985).
604 Hink, W. F., Thomsen, D. R., Davidson, D. J., Meyer, A. L. & Castellino, F. J. Expression of three recombinant proteins using baculovirus vectors in 23 insect cell lines. *Biotechnol Prog* **7**, 9-14 (1991).
605 Vaughn, J. L., Goodwin, R. H., Tompkins, G. J. & McCawley, P. The establishment of two cell lines from the insect Spodoptera frugiperda (Lepidoptera; Noctuidae). *In Vitro* **13**, 213-217 (1977).
606 Lathe, G. H. & Ruthven, C. R. The separation of substances and estimation of their relative molecular sizes by the use of colums of starch in water. *Biochem J* **62**, 665-674 (1956).
607 Lathe, G. H. & Ruthven, C. R. The separation of substances on the basis of their molecular weights, using columns of starch and water. *Biochem J* **60**, xxxiv (1955).
608 Porath, J. & Flodin, P. Gel filtration: a method for desalting and group separation. *Nature* **183**, 1657-1659 (1959).
609 Laemmli, U. K. Cleavage of structural proteins during the assembly of the head of bacteriophage T4. *Nature* **227**, 680-685 (1970).
610 Fried, M. & Crothers, D. M. Equilibria and kinetics of lac repressor-operator interactions by polyacrylamide gel electrophoresis. *Nucleic Acids Res* **9**, 6505-6525 (1981).

611 Garner, M. M. & Revzin, A. A gel electrophoresis method for quantifying the binding of proteins to specific DNA regions: application to components of the Escherichia coli lactose operon regulatory system. *Nucleic Acids Res* **9**, 3047-3060 (1981).
612 Ericsson, U. B., Hallberg, B. M., Detitta, G. T., Dekker, N. & Nordlund, P. Thermofluor-based high-throughput stability optimization of proteins for structural studies. *Anal Biochem* **357**, 289-298 (2006).
613 Price, W. N., 2nd *et al.* Understanding the physical properties that control protein crystallization by analysis of large-scale experimental data. *Nat Biotechnol* **27**, 51-57 (2009).
614 Watson, J. D. & Crick, F. H. The structure of DNA. *Cold Spring Harb Symp Quant Biol* **18**, 123-131 (1953).
615 Kendrew, J. C. *et al.* A three-dimensional model of the myoglobin molecule obtained by x-ray analysis. *Nature* **181**, 662-666 (1958).
616 Matthews, B. W. Solvent content of protein crystals. *J Mol Biol* **33**, 491-497 (1968).
617 Basak, S. & Ramaswamy, H. S. Simultaneous Evaluation of Shear Rate and Time Dependency of Stirred Yogurt Rheology as Influenced by Added Pectin and Strawberry Concentrate. *J Food Eng* **21**, 385-393 (1994).
618 Ramaswamy, H. S. & Basak, S. Pectin and Raspberry Concentrate Effects on the Rheology of Stirred Commercial Yogurt. *J Food Sci* **57**, 357-360 (1992).
619 Arndt, U. W., Crowther, R. A. & Mallett, J. F. A computer-linked cathode-ray tube microdensitometer for x-ray crystallography. *J Sci Instrum* **1**, 510-516 (1968).
620 Diederichs, K. & Karplus, P. A. Improved R-factors for diffraction data analysis in macromolecular crystallography. *Nat Struct Biol* **4**, 269-275 (1997).
621 Weiss, M. S. Global indicators of X-ray data quality. *J Appl Crystallogr* **34**, 130-135 (2001).
622 Brunger, A. T. Free R value: a novel statistical quantity for assessing the accuracy of crystal structures. *Nature* **355**, 472-475 (1992).
623 Karplus, P. A. & Diederichs, K. Linking Crystallographic Model and Data Quality. *Science* **336**, 1030-1033 (2012).
624 Krivanek, O. L., Dellby, N., Spence, A. J., Camps, R. A. & Brown, L. M. Aberration correction in the STEM. *Inst Phys Conf Ser*, 35-40 (1997).
625 Bruggeller, P. & Mayer, E. Complete Vitrification in Pure Liquid Water and Dilute Aqueous-Solutions. *Nature* **288**, 569-571 (1980).
626 Dubochet, J. & Mcdowall, A. W. Vitrification of Pure Water for Electron-Microscopy. *J Microsc-Oxford* **124**, Rp3-Rp4 (1981).
627 Toyoshima, C. & Unwin, N. Contrast transfer for frozen-hydrated specimens: determination from pairs of defocused images. *Ultramicroscopy* **25**, 279-291 (1988).
628 Leapman, R. D. & Sun, S. Q. Cryoelectron Energy-Loss Spectroscopy - Observations on Vitrified Hydrated Specimens and Radiation-Damage. *Ultramicroscopy* **59**, 71-79 (1995).
629 van Heel, M. *et al.* Single-particle electron cryo-microscopy: towards atomic resolution. *Quarterly reviews of biophysics* **33**, 307-369 (2000).
630 Frank, J. *Three-dimensional electron microscopy of macromolecular assemblies*. 2nd edn, (Oxford University Press, 2006).
631 Van Heel, M. Angular reconstitution: a posteriori assignment of projection directions for 3D reconstruction. *Ultramicroscopy* **21**, 111-123 (1987).
632 Dyer, P. N. *et al.* Reconstitution of nucleosome core particles from recombinant histones and DNA. *Chromatin and Chromatin Remodeling Enzymes, Pt A* **375**, 23-44 (2004).
633 Lowary, P. T. & Widom, J. New DNA sequence rules for high affinity binding to histone octamer and sequence-directed nucleosome positioning. *J Mol Biol* **276**, 19-42 (1998).

Structure-function study of the chromatin-remodelling complex NuRD

Résumé en anglais

The analysis of large macromolecular assemblies requires an integrated structural biology approach which combines notably biochemical preparation, biophysical characterization, single particle cryo-electron microscopy and X-ray crystallography. These techniques have been used in my work to study the structural organization of the Nucleosome Remodelling and histone Deacetylation (NuRD) complex, which comprises several subunits, among which histone deacetylases HDAC1/2, ATP-dependent remodelling enzymes CHD3/4, histone chaperones RbAp46/48, CpG-binding proteins MBD2/3 and specific regulatory proteins MTA1/2/3.
My work focuses especially on MBD3, RbAp46 and RbAp48. For this purpose, I set up the preparation of the individual subunits and characterized them by various biophysical methods, such as gel-filtration, analytical ultra-centrifugation (AUC), dynamic light scattering (DLS) or Thermofluor®. We next analyzed whether the purified NuRD subunits interact with nucleosomes. For this, we formed complexes by mixing purified protein with homemade human nucleosomes, and analyzed the samples by electromobility shift assays (EMSA).
For MBD3, despite a partial unfolding of the protein, optimisation of salt concentrations and ratios allowed us to observe positive results of the protein interacting with nucleosomes on EMSA gels, and to address the stoichiometry of the complex. Crystals diffracting up to 7 Å were obtained and are currently being optimised. Furthermore, a preliminary 3-D reconstruction at 25 Å resolution has been solved in cryo-EM, showing an extra-density on top of the nucleosome in which the crystal structure of the MBD domain of MBD3 could be fitted.
For RbAp46/48, crystal structures are available and suggest that those chaperones can't bind to nucleosomes, but only to free H4 histone. Nevertheless, in contrast to current belief, those proteins were shown to form stable complexes with the nucleosome. Together, the present results reveal that those proteins interact well with the nucleosome, paving the way for future structural analysis by cryo-EM or X-ray crystallography.

Keywords: chromatin, NuRD, nucleosome, remodelling, biochemistry, crystallography, cryo-EM.

Résumé en français

L'étude de grands complexes macromoléculaires requiert une approche de biologie structurale intégrative qui combine à la fois la préparation biochimique, la caractérisation biophysique, la cryo-microscopie électronique et la cristallographie aux rayons-X. Ces techniques ont été mises à profit dans mon travail de thèse, pour l'étude de l'organisation structurale du complexe de remodelage de la chromatine et de déacétylation « NuRD », qui comprend plusieurs sous-unités, dont les histones déacétylases HDAC1/2, les enzymes de remodelage ATP-dépendant CHD3/4, les chaperonnes d'histones RbAp46/48, les protéines de liaison à l'ADN méthylé MBD2/3 et les protéines de régulation MTA1/2/3.
Mon travail s'est focalisé essentiellement sur MBD3, RbAp46 et RbAp48. Dans ce but, j'ai mis en place les protocoles de production et de purification de ces différentes sous-unités, et les ai caractérisé biophysiquement par diverses méthodes comme la chromatographie d'exclusion stérique, l'ultracentrifugation analytique, la diffusion dynamique de lumière ou le Thermofluor®. Nous avons ensuite entrepris des études de liaisons sur des nucléosomes reconstitués au laboratoire, par expérience de retard sur gel.
Pour MBD3, malgré un dépliement partiel de la protéine, l'optimisation des concentrations saline et ratio a permis d'observer un résultat positif de cette protéine interagissant avec le nucléosome, et de définir la stœchiométrie du complexe. Des cristaux diffractant jusqu'à 7 Å de résolution ont été obtenus et sont en phase d'optimisation. Egalement, une reconstitution 3-D préliminaire à partir de données de cryo-microscopie électronique a pu être obtenue, à 25 Å de résolution, montrant une densité électronique supplémentaire sur le nucléosome, et dans laquelle la structure cristallographique du domaine MBD de MBD2 a pu être replacée.
Pour RbAp46/48, les structures cristallographiques étaient déjà publiées et suggéraient que ces chaperonnes ne pouvaient pas lier les nucléosomes mais uniquement l'histone H4 libre. Cependant, nous avons pu montrer que ces protéines formaient un complexe stable avec le nucléosome. Ainsi, ces résultats montrent que ces deux protéines interagissent avec le nucléosome et pavent la voie pour leur future étude structurale en cryo-microscopie électronique ou en cristallographie aux rayons-X.

Mots-clefs : chromatine, NuRD, nucléosome, remodelage, biochimie, cristallographie, cryo-EM