

*École Doctorale des Sciences de la Vie et de la Santé*

Centre de Biologie Intégrative  
Institut de Génétique et de Biologie Moléculaire et Cellulaire  
CNRS UMR 7104 - Inserm U 1258

## THÈSE

présentée par :

**Brice BEINSTEINER**

soutenue le : **16 octobre 2018**

pour obtenir le grade de : **Docteur de l'université de Strasbourg**

Discipline : Sciences de la Vie

Spécialité : Biophysique et Biologie Structurale

## Origine et évolution des récepteurs nucléaires et étude structurale du premier stéroïdien, ERR

**THÈSE dirigée par :**

**Dr. KLAHOLZ Bruno**  
**Dr. MORAS Dino**

DR, IGBMC, Université de Strasbourg  
DR, IGBMC, Université de Strasbourg

**RAPPORTEURS :**

**Dr. GAUTHIER-VANACKER Karine**  
**Pr. SAIBIL Helen**

DR, IGFL, Lyon  
DR, Birkbeck College, Londres

---

**AUTRES MEMBRES DU JURY :**

**Dr. BOURGUET William**  
**Dr. POCH Olivier**

DR, CBS, Montpellier  
DR, LBGI, Université de Strasbourg



# Remerciements

Je tiens à remercier les membres de mon jury de thèse, Dr. Karine GAUTHIER-VANACKER, Pr. Helen SAIBIL, Dr. William BOURGUET et Dr. Olivier POCH pour avoir accepté de lire le présent manuscrit et pour l'honneur qu'ils me font de juger cette thèse qui est l'aboutissement de 4 années de travail.

Je tiens également à exprimer ma profonde reconnaissance et à remercier chaleureusement Bruno KLAHOLZ et Dino MORAS, mes deux directeurs de thèse. Merci de m'avoir donné l'opportunité de réaliser cette présente thèse avec tout votre soutien et vos conseils aussi bien scientifiques que humains. Merci pour votre confiance, pour la grande liberté que vous m'avez accordé dans ma démarche scientifique. Merci également de m'avoir guidé tout au long de ce projet, et pour la formation scientifique de très grande qualité que j'ai pu avoir avec vous, chacun à votre façon. J'ai énormément appris au cours de ces dernières années et je suis heureux d'avoir fait mes premiers pas de chercheur à vos côtés.

Je souhaite également remercier Isabelle BILLAS qui m'a co-encadré pour de nombreux aspects de cette thèse, ainsi que Kareem MOHIDEEN ABDUL pour son implication sans faille dans les projets de cryo-microscopie électronique. Tous deux sont à l'origine de la préparation biochimique des échantillons et m'ont permis d'appréhender, de comprendre et de discuter de nombreux points pour l'optimisations des échantillons étudiés. Merci d'avoir été là pour toutes mes questions et toutes mes demandes de modifications de la préparation des échantillons ainsi que pour les nombreuses discussions enrichissantes qu'on a eu ensemble.

Je souhaite remercier Jean-François MENETRET de m'avoir appris les bases de la préparations d'échantillons et de l'utilisation du Polara.

Je remercie également, Jonathan MICHALON sans qui une bonne partie des développements informatiques réalisés durant cette thèse n'auraient pas été possible, on forme une bonne équipe ! Mais aussi pour tous les autres side projects qu'on a eu et qu'on aura encore, au labo ou en externe, aussi bien en photographie qu'en divers et nombreux autres projets.

Je voulais également remercier Rachel TABARONI, Simon PICHARD, Mathieu SCHAEFFER, Léo FRECHIN, Nils MARECHAL, Luka VANDERVELDEN et Tan Dat TRUONG pour tous les bons moments passés au cours de ces dernières années.

Je remercie également Sacha OURJOUNTSEV et Kundhavi NATCHIAR pour toutes les discussions intéressantes qu'on a eues dans le bureau et l'aide apportée quand j'en avais besoin.

Pour m'avoir permis d'apprendre la cryo-microscopie électronique et avoir répondu à mes nombreuses questions, je souhaite remercier Igor ORLOV, Sacha MYASNIKOV, Clara Otilie LOEFFELHOLZ VON COLBERG, Julio ORTIZ ESPINOZA, Gabor PAPAI, Corinne CRUCIFIX, Christine RUHLMANN, Nicolas LEMERCIER et Patrick SCHULTZ. Vous avez chacun d'entre vous participé à ma formation sur les points pratiques comme sur les points théoriques, aussi bien pour la préparation d'échantillons que pour l'utilisation des microscopes.

Je souhaite également remercier Didier HENTSCH et Jean-Luc VONESCH de la plateforme d'imagerie pour leur soutien et l'ensemble des discussions intéressantes que nous avons partagées.

Je remercie aussi l'ensemble des membres de l'équipe passé et présent et plus largement l'ensemble des personnes que j'ai côtoyé au cours de ces dernières années dans l'institut.

Pour finir je souhaite remercier mes proches qui m'ont soutenu tout au long de ses années et sans qui tout ceci n'aurait été possible. Mes amis, ma famille et en particulier ma compagne Thuy Thanh, ma maman et mes grands-parents qui ont tous toujours été là pour moi de manière inconditionnelle tout au long de ses quatre années. Pour tout ceci, je vous dis un énorme merci!

# Table des matières

Table des figures.....	IV
Liste des abréviations .....	VIII
Résumé de thèse .....	XI
Introduction.....	1
1. La superfamille des récepteurs nucléaires.....	4
1.1 Fonctions biologiques.....	4
1.2 Origine et évolution.....	6
1.2.1 La systématique.....	6
1.2.2 Origine et classification des récepteurs nucléaires .....	11
1.2.3 Diversification et évolution de la superfamille des récepteurs nucléaires .....	13
1.2.4 Acquisition de la capacité à fixer un ligand .....	15
1.3 Découverte et évolution des connaissances .....	17
1.4 Structure modulaire .....	19
1.4.1 Le domaine A/B .....	20
1.4.2 Le domaine C .....	21
1.4.3 Le domaine D.....	22
1.4.4 Le domaine E .....	23
1.4.5 Le domaine F .....	26
1.5 Les éléments de réponses .....	26
1.5.1 Les récepteurs ligand-dépendants .....	27
1.5.2 Les récepteurs orphelins .....	27
1.6 Les ligands .....	29
1.7 Les complexes récepteurs nucléaires-ADN .....	32
2. Le récepteur nucléaire orphelin ERR apparenté à ER .....	36
2.1 ERR et ses 3 isotypes $\alpha$ , $\beta$ , $\gamma$ .....	36
2.2 La régulation de l'activité transcriptionnelle d'ERR.....	37
2.2.1 Le coactivateur PGC-1 $\alpha$ .....	37
2.2.2 Le corépresseur NCoR .....	39
2.2.3 Régulation d'ERR avec PGC-1 $\alpha$ et NCoR.....	40
2.2.4 Les modifications post-transcriptionnelles .....	41
2.3 ERR $\alpha$ est un régulateur du métabolisme énergétique.....	42
2.3.1 Métabolisme du glucose .....	42
2.3.2 $\beta$ -oxydation des acides gras .....	42
2.3.3 Fonctions mitochondriales.....	43
2.4 La balance ERR $\alpha$ /ERR $\gamma$ dans les cancers.....	43
2.5 Modulation d'ERR $\alpha$ par des ligands synthétiques.....	45
3. La microscopie électronique à transmission .....	49
3.1 Les débuts et l'évolution au cours du 20 <sup>ème</sup> siècle .....	49
3.2 Le microscope électronique à transmission.....	50
3.2.1 Les sources d'électrons.....	51
3.2.2 La formation de l'image.....	51
3.2.3 Les limites physiques rencontrées par la MET .....	54
3.3 L'étude d'un échantillon biologique.....	57
3.3.1 La coloration négative .....	58
3.3.2 La cryo-microscopie électronique .....	60

3.4	Les conditions d'acquisition .....	63
3.5	Principes généraux de la reconstruction 3D en microscopie électronique.....	64
3.6	La révolution de la résolution.....	68
3.6.1	Des caméras à détections directes.....	68
3.6.2	Le traitement d'images.....	71
3.6.3	Des microscopes plus stables et plus performants .....	72
3.6.4	Des échantillons plus reproductibles .....	74
3.6.5	De nouvelles grilles plus performantes .....	75
Chapitre 1 : Origine et évolution des récepteurs nucléaires.....		78
1.	Introduction.....	78
2.	Résultats - Article scientifique.....	80
3.	Conclusion et perspectives.....	103
Chapitre 2 : Détermination structurale d'ERR $\alpha$ par cryo-microscopie électronique .....		104
1.	Introduction.....	104
2.	Matériel et méthodes.....	106
2.1	La production et purification des complexes .....	107
2.2	La congélation de l'échantillon.....	108
2.2.1	Les grilles .....	108
2.2.2	Système à décharge luminescente.....	109
2.2.3	Cryogénéisation des échantillons.....	110
2.3	Acquisition des images .....	111
2.4	Le traitement d'images et reconstruction 3D .....	112
2.4.1	Le prétraitement .....	112
2.4.2	Le tri des micrographes .....	113
2.4.3	La sélection des particules.....	114
2.4.4	La classification 2D .....	117
2.4.5	Détermination d'une structure initiale.....	118
2.4.6	Classification 3D .....	119
2.4.7	Affinement de la structure 3D.....	120
3.	Résultats et discussion .....	121
3.1	Complexe ERR $\alpha$ -ADN BE33- <i>tff1</i> ERRE.....	122
3.1.1	Préparation d'échantillons .....	122
3.1.2	Acquisitions .....	125
3.1.3	Classification 2D .....	129
3.1.4	Reconstruction 3D.....	129
3.1.5	Limites-problèmes.....	130
3.2	Complexe ERR-Di Nucléosome et ERR-Nucléosome .....	131
3.2.1	Préparation d'échantillons .....	131
3.2.2	Acquisitions .....	135
3.2.3	Classification 2D .....	138
3.2.4	Reconstruction 3D.....	139
3.2.5	Classification 3D .....	140
3.2.6	Limites-problèmes.....	141
3.3	Complexe ERR $\alpha$ -ADN BE33-embedded ERRE/ERE .....	141
3.3.1	Préparation d'échantillons .....	141
3.3.2	Acquisitions .....	141
3.3.3	Classification 2D .....	143
3.3.4	Reconstruction 3D.....	143
3.3.5	Limites-problèmes.....	144
3.4	Complexe ERR $\alpha$ - BE29-embedded ERRE/ERE -PGC-1 $\alpha$ .....	144
3.4.1	Préparation d'échantillons .....	144
3.4.2	Acquisitions .....	145

3.4.3	Classification 2D .....	146
3.4.4	Reconstruction 3D .....	149
3.4.5	Classification 3D .....	155
3.4.6	Limites-problèmes .....	156
4.	Conclusions et perspectives .....	159
Chapitre 3 : Développements informatiques .....		161
1.	IBiSS, un développement bioinformatique d'un outil interactif pour l'analyse structure-séquence de grands complexes macromoléculaires .....	161
1.1	Introduction et contexte .....	161
1.2	Résultats - Article scientifique .....	164
1.3	Discussion et perspective .....	171
2.	Grid Files Manager un logiciel utilitaire de suivis d'échantillons en Cryo-ME .....	173
2.1	Introduction et contexte .....	173
2.2	Résultats .....	173
2.3	Discussion et perspectives .....	179
3.	BackPhylo, un logiciel d'évolution et de phylogénie .....	181
3.1	Introduction et contexte .....	181
3.2	Résultats .....	182
3.3	Discussion et perspectives .....	195
Discussion générale .....		196
Publications .....		205
Bibliographie .....		207

# Table des figures

Tableau 1 : Niveau d'expression et diagnostique associé pour les cancers où les isotypes d'ERR sont impliqués. ....	44
Tableau 2 : Ligands antagonistes connus pour ERR $\alpha$ .....	48
Tableau 3 : Récapitulatif des jeux de données principaux acquis au cours de la thèse.....	159

---

Figure 1 : Mécanisme d'action général d'un récepteur nucléaire de classe I ligand dépendant (homodimère). ....	5
Figure 2 : Mécanisme d'action général d'un récepteur nucléaire de classe II ligand dépendant (hétérodimère). ....	5
Figure 3 : Arbre phylogénétique du vivant.....	7
Figure 4 : Anatomie de polype et méduse. ....	8
Figure 5 : La symétrie corporelle des Eumetazoas.....	8
Figure 6 : Les cavités corporelles des animaux triploblastiques .....	9
Figure 7 : Comparaison des modes de développement protostomien et deutérostomien. ....	10
Figure 8 : Phylogénie des métazoaires basée sur des données moléculaires.....	11
Figure 9 : Les six sous-familles de la superfamille des récepteurs nucléaires. ....	12
Figure 10 : Duplications et pertes des nucléaires récepteurs dans l'évolution des Métazoaires .....	14
Figure 11 : Répartition des récepteurs nucléaires pour 12 modèles vertébrés.....	15
Figure 12 : Schéma illustrant l'acquisition de la capacité de fixation du ligand .....	16
Figure 13 : Premier postulat du mécanisme du récepteur d'œstrogène .....	18
Figure 14 : Organisation structurale générale des récepteurs nucléaires .....	19
Figure 15 : Représentation schématique du DBD de RXR.....	22
Figure 16 : Sandwich d'hélices $\alpha$ antiparallèles du domaine LBD .....	23
Figure 17 : La fixation du ligand et le mécanisme de piège à souris. ....	25
Figure 18 : Schéma des principaux types d'éléments de réponses des récepteurs nucléaires. ....	28
Figure 19 : Structures de différents types d'éléments de réponses. ....	29
Figure 20 : Ligands et récepteurs nucléaires associés.....	30
Figure 21 : Biosynthèse des stéroïdes sexuels .....	31
Figure 22 : Complexes de récepteurs nucléaires avec leur ADN.....	33
Figure 23 : Modèle de la structure ADN/ERR $\alpha$ /SRC-3/CARM1/p300.....	35
Figure 24 : Niveau d'expression des ERR en fonction des tissus.....	37
Figure 25 : Organisation en une dimension des domaines fonctionnels de PGC-1 $\alpha$ . ....	38
Figure 26 : A : Structure d'ERR $\alpha$ avec un peptide du coactivateur PGC-1 $\alpha$ .....	39
Figure 27 : Organisation en une dimension des domaines fonctionnels de NCoR. ....	40

Figure 28 : Régulation du métabolisme oxydatif par PGC-1 $\alpha$ et NCoR.....	41
Figure 29 : Le composé antagoniste 1a.....	46
Figure 30 : Le composé antagoniste 29.....	47
Figure 31 : Anatomie schématique d'un microscope électronique. ....	50
Figure 32 : Les différentes sources d'électrons utilisées en microscopie électronique.....	51
Figure 33 : A gauche, une photo d'une lentille magnétique. ....	52
Figure 34 : Schéma des différents types d'interactions entre les électrons et l'échantillon. ....	53
Figure 35 : Schéma du chemin optique de la formation de l'image .....	54
Figure 36 : Courbe de la fonction FTC théorique pour plusieurs valeur de défocalisation.....	56
Figure 37 : Effet de la fonction enveloppe sur la FTC théorique.....	57
Figure 38 : Exemple d'image de coloration négative de virus prise en 1959.....	58
Figure 39 : Photographie du CM120 (Philips) de l'institut IGBMC/CBI .....	59
Figure 40 : Exemple d'image de Cryo-microscopie électronique de virus prise début des années 80. 60	
Figure 41 : Photographies de grilles .....	61
Figure 42 : Photographie du Titan Krios (FEI) de l'institut IGBMC/CBI.....	62
Figure 43: Photographie du Polara (FEI) de l'institut IGBMC/CBI .....	63
Figure 44 : Schéma d'une ligne commune (C).....	65
Figure 45 : Schéma de l'inclinaison conique aléatoire et de la reconstruction inclinée orthogonale ..	66
Figure 46 : Schéma du principe de tomographie .....	67
Figure 47 : Comparaison des deux systèmes de détections pour la cryo-microscopie électronique... 69	
Figure 48 : Comparaison DQE entre plusieurs caméras et films photographiques .....	70
Figure 49 : Effet de l'alignement des images d'un film .....	71
Figure 50 : Les différents plongeurs actuels.....	74
Figure 51 : Plongeur Spotiton V1.0.....	75
Figure 52 : Mouvements de l'échantillon dus à l'irradiation du faisceau d'électrons .....	76
Figure 53 : Grille en or.....	77
Figure 54 : Eléments de réponses reconnus par ER et ERR.....	105
Figure 55 : Gel natif 5% EMSA d'un complexe ERR $\alpha$ -ADN .....	108
Figure 56 : Schéma de la préparation de grilles avec du carbone flotté.....	109
Figure 57 : Photographie d'une décharge lumineuse avec une machine Elmo cordouan.....	110
Figure 58 : Exemple d'une interface de trie des micrographes (cisTEM).....	114
Figure 59 : Exemple d'une interface de sélection de particules manuelle (Relion). ....	115
Figure 60 : Exemple d'une interface de sélection de particules semi manuelle (EMAN2). ....	116
Figure 61 : Exemple d'une interface de sélection de particules automatique (cisTEM).....	117
Figure 62 : Exemple de classification 2D pour un complexe ERR $\alpha$ -ADN-PGC-1 $\alpha$ (IMAGIC).....	118
Figure 63 : Exemple de structure initiale pour un complexe ERR-ADN (cryoSPARC).....	119

Figure 64 : Exemple de classification 3D pour un complexe de nucléosome-ERR (Relion) .....	120
Figure 65 : Exemple de carte affinée pour le complexe $ERR\alpha$ -ADN-PGC-1 $\alpha$ (cisTEM).....	121
Figure 66 : Séquence de l'ADN BE33-tff1 avec l'élément de réponse ERRE encadré. ....	122
Figure 67 : Micrographe de tampon Tris pH7 seul.....	123
Figure 68 : Micrographe illustrant les problèmes de distributions des particules.....	124
Figure 69 : Micrographe de l'acquisition du complexe $ERR\alpha$ -ADN BE33-tff1 ERRE .....	126
Figure 70 : Micrographe réalisé sur le microscope Titan Krios du complexe $ERR\alpha$ -ADN BE33-tff1 ERRE .....	127
Figure 71 : Micrographe réalisé sur le microscope Titan Krios du complexe $ERR\alpha$ -ADN BE33-tff1 ERRE .....	128
Figure 72 : Classes 2D du complexe $ERR\alpha$ -ADN BE33-tff1 ERRE (EMAN2).....	129
Figure 73 : Première structure du complexe $ERR\alpha$ -ADN BE33-tff1 ERRE .....	130
Figure 74 : Gel natif 5% EMSA de plusieurs complexe NCP- $ERR\alpha$ .....	132
Figure 75 : Séquence des éléments de réponse utilisés dans les complexes nucléosomes .....	133
Figure 76 : Schéma de la position de l'élément de réponse (ER).....	133
Figure 77 : Micrographe du complexe Di-NCP- $ERR\alpha$ .....	134
Figure 78 : Micrographe du complexe Di-NCP- $ERR\alpha$ .....	135
Figure 79 : Micrographe du complexe NCP- $ERR\alpha$ .....	136
Figure 80 : Micrographe du complexe NCP- $ERR\alpha$ .....	137
Figure 81 : Comparatif de l'évolution de l'équipement d'acquisition durant la thèse .....	138
Figure 82 : Exemple de classes 2D du complexe NCP- $ERR\alpha$ .....	138
Figure 83 : Structure initiale du complexe ERR-nucléosome généré avec cryoSPARC.....	139
Figure 84 : Structure après le premier cycle d'affinement fait avec Relion 2.1 .....	140
Figure 85 : Classification 3D pour un complexe de nucléosome-ERR (Relion).....	140
Figure 86 : Séquence de l'ADN BE33-embedded IR3 avec l'élément de réponse ERRE/ERE .....	141
Figure 87 : Micrographe du complexe $ERR\alpha$ -ADN .....	142
Figure 88 : Classe 2D du complexe $ERR\alpha$ -ADN (Relion 2.0) .....	143
Figure 89 : La structure initiale générée et affinée dans cryoSPARC .....	143
Figure 90 : La structure initiale générée et affinée dans cryoSPARC (tournée de 90° dans l'axe Y) ...	144
Figure 91 : Séquence de l'ADN WC29-embedded-AA avec l'élément de réponse ERRE/ERE .....	145
Figure 92 : Micrographe du complexe $ERR\alpha$ -ADN-PGC-1 $\alpha$ .....	146
Figure 93 : Classe 2D du complexe $ERR\alpha$ -ADN-PGC-1 $\alpha$ (Relion 2.1). .....	147
Figure 94 : Classe 2D du complexe $ERR\alpha$ -ADN-PGC-1 $\alpha$ (cisTEM).....	148
Figure 95 : Classe 2D du complexe $ERR\alpha$ -ADN-PGC-1 $\alpha$ (IMAGIC sans filtre). .....	148
Figure 96 : Classe 2D du complexe $ERR\alpha$ -ADN-PGC-1 $\alpha$ (IMAGIC avec filtre passe-bande). .....	149
Figure 97 : Comparaison entre une structure cristallographique et la carte de cryo-ME .....	149

Figure 98 : Distribution des angles d'Euler des particules du jeu de données (cisTEM).....	150
Figure 99 : A gauche, la structure initiale, à droite le résultat après plusieurs cycles d'affinement avec cisTEM .....	151
Figure 100 : A gauche, la structure initiale, à droite le résultat après plusieurs cycles d'affinement avec CisTEM.....	151
Figure 101 : Structure après plusieurs cycles d'affinement avec Relion2.1.....	152
Figure 102 : Classification 2D de Relion2.1. ....	153
Figure 103 : Classification 2D de Relion2.1 .....	154
Figure 104 : Positionnement d'une structure cristallographique des LBD de ERR $\alpha$ avec un petit fragment de PGC-1 $\alpha$ dans la carte de cryo-ME.....	155
Figure 105 : Classification 3D ciblé sur la partie ADN/DBD de la structure (focus refinement) .....	156
Figure 106 : Classification 3D ciblé sur la partie ADN/DBD de la structure. Rotation de 90° sur l'axe Y. ....	156
Figure 107 : Micrographe du complexe ERR $\alpha$ -ADN-PGC-1 $\alpha$ .....	157
Figure 108 : Exemple de classes 2D de particules mal centré.....	158
Figure 109 : Schéma des différentes étapes de traitement d'images.....	158
Figure 110 : Interface principale dans Grid Files Manager .....	175
Figure 111 : Interface de saisie de la composition de l'échantillon dans Grid Files Manager .....	175
Figure 112 : Interface de saisie des détails de cryogénéisation des échantillons dans Grid Files Manager. ....	176
Figure 113 : Interface de saisie des détails d'acquisition et du microscope dans Grid Files Manager. ....	177
Figure 114 : Interface d'ajout de l'image de l'atlas de la grille dans Grid Files Manager.....	178
Figure 115 : Interface d'ajout d'une image exemple de l'échantillon à grandissement de travail dans Grid Files Manager. ....	179
Figure 116 : Interface générale de BackPhylo.....	182
Figure 117 : Interface générale de BackPhylo, phase de préparation. ....	183
Figure 118 : Interface générale de BackPhylo, partie filtrage.....	185
Figure 119 : Interface générale de BackPhylo, ajout manuel de séquence. ....	186
Figure 120 : Interface générale de BackPhylo, Calcule du fichier newick.....	187
Figure 121 : Interface générale de BackPhylo, interface de sélection du display du newick .....	188
Figure 122 : Interface générale de BackPhylo, interface de phylogénie, display du newick. ....	189
Figure 123 : Interface générale de BackPhylo, interface de phylogénie, zoom display du newick. ...	190
Figure 124 : Interface générale de BackPhylo, interface de fusion de deux newick.....	190
Figure 125 : Interface générale de BackPhylo, interface CSV export.....	191
Figure 126 : Interface générale de BackPhylo, résultat CSV export (première moitié). ....	193
Figure 127 : Interface générale de BackPhylo, résultat CSV export (seconde moitié).....	194

# Liste des abréviations

<b>2D :</b>	<b>2 Dimensions</b>
<b>3D :</b>	<b>3 Dimensions</b>
<b>ADN :</b>	<b>Acide DésoxyriboNucléique</b>
<b>AF-1 :</b>	Fonction d'activation 1 / <b>Activation-Function 1</b>
<b>AF-2 :</b>	Fonction d'activation 2 / <b>Activation-Function 2</b>
<b>AP-1 :</b>	Protéine activatrice 1 / <b>Activator Protein 1</b>
<b>AR :</b>	Récepteur des androgènes / <b>Androgen Receptor</b>
<b>ARN :</b>	<b>Acide RiboNucléique</b>
<b>BLAST :</b>	<b>Basic Local Alignment Search Tool</b>
<b>CARM1 :</b>	Arginine méthyltransférase associée au coactivateur 1 / <b>Coactivator-Associated Arginine Methyltransferase 1</b>
<b>CBP :</b>	Protéine liant le CREB / <b>CREB-Binding Protein</b>
<b>Cc :</b>	Coefficient d'aberration Chromatique
<b>CCD :</b>	Dispositif à transfert de charges / <b>Charge-Coupled Device</b>
<b>CDK9 :</b>	Kinase dépendante de la cycline 9 / <b>Cyclin-Dependant Kinase 9</b>
<b>CLI :</b>	Interface en ligne de commande / <b>Command Line Interface</b>
<b>CMC :</b>	Concentration <b>Micellaire Critique</b>
<b>CMOS :</b>	<b>Complementary Metal-Oxide-Semiconductor</b>
<b>CoA :</b>	<b>CoActivateur</b>
<b>CoR :</b>	<b>CoRrépresseur</b>
<b>SMRT :</b>	<b>Silencing Mediator for Retinoic acid and Thyroid hormone receptors</b>
<b>CREB :</b>	Protéine de liaison à l'élément de réponse CAMP / <b>CAMP-Response-Element-Binding</b>
<b>Cryo-ME :</b>	<b>Cryo-Microscopie Electronique</b>
<b>Cs :</b>	Coefficient d'aberration Sphérique
<b>CS (model) :</b>	Propagation Conformationnelle / <b>Conformational Spread</b>
<b>CSV :</b>	<b>Comma-Separated Values</b>
<b>CTE :</b>	Extension Carboxyl-Terminale / <b>C-Terminal Extension</b>
<b>DBD :</b>	Domaine de liaison à l'ADN / <b>DNA Binding Domain</b>
<b>DDBJ :</b>	<b>DNA Data Bank of Japan</b>
<b>DHEA :</b>	<b>DéHydroEpiAndrostérone</b>
<b>DHT :</b>	<b>DiHydroTestostérone</b>
<b>DQE :</b>	Efficacité quantique de détection / <b>Detective Quantum Efficiency</b>
<b>DR :</b>	Répétition directe / <b>Direct Repeat</b>
<b>EBI :</b>	<b>European Bioinformatics Institute</b>
<b>EcR :</b>	Récepteur de l'ecdysone / <b>Ecdysone Receptor</b>
<b>ER :</b>	Récepteurs des œstrogènes / <b>Estrogen Receptor</b>
<b>ERE :</b>	Élément de réponse aux œstrogènes / <b>Estrogen Response Element</b>
<b>ERH :</b>	Élément de Réponse aux Hormones
<b>ERR :</b>	Récepteur apparenté-aux-œstrogènes / <b>Estrogen-Related Receptor</b>
<b>ERRE :</b>	Élément de réponse de ERR / <b>ERR Response Element</b>
<b>FEG :</b>	Canon à émission de champs / <b>Field Emission Gun</b>

<b>FRET :</b>	Transfert d'énergie entre molécules fluorescentes / <b>F</b> luorescence <b>R</b> esonance <b>E</b> nergy <b>T</b> ransfer
<b>FTC :</b>	<b>F</b> onction de <b>T</b> ransfert de <b>C</b> ontraste
<b>GIF :</b>	Filtre d'énergie / <b>G</b> atan <b>I</b> maging <b>F</b> ilter
<b>GR :</b>	Récepteur des glucocorticoïdes / <b>G</b> lucocorticoid <b>R</b> eceptor
<b>HAT :</b>	Histone acétyltransférase / <b>H</b> istone <b>A</b> cetyl <b>T</b> ransferase
<b>HDAC-3 :</b>	Histone désacétylase 3 / <b>H</b> istone <b>D</b> e <b>A</b> cetylase <b>3</b>
<b>HNF-4 :</b>	Facteur nucléaire hépatocytaire 4 / <b>H</b> epatocyte <b>N</b> uclear <b>F</b> actor- <b>4</b>
<b>HSP :</b>	Protéine de choc thermique / <b>H</b> eat <b>S</b> hock <b>P</b> rotein
<b>HTTP :</b>	protocole de transfert hypertexte / <b>H</b> yper <b>T</b> ext <b>T</b> ransfer <b>P</b> rotocol
<b>IR :</b>	Répétition inversée / <b>I</b> nverted <b>R</b> epeat
<b>JPEG :</b>	<b>J</b> oint <b>P</b> hotographic <b>E</b> xperts <b>G</b> roup
<b>KCl :</b>	Chlorure de potassium
<b>KNF :</b>	<b>K</b> oshland, <b>N</b> emethy, et <b>F</b> ilmer
<b>LBD :</b>	Domaine de liaison du ligand; <b>L</b> igand <b>B</b> inding <b>D</b> omain
<b>LXR :</b>	Récepteur nucléaire des oxystérols / <b>L</b> iver <b>X</b> <b>R</b> eceptors
<b>MAPK :</b>	Protéines kinases activées par un mitogène / <b>M</b> itogen- <b>A</b> ctivated <b>P</b> rotein <b>K</b> inase
<b>MCAD :</b>	Acyl-coenzyme A déshydrogénase des acides gras à chaîne moyenne / <b>M</b> edium- <b>C</b> hain acyl-coenzyme <b>A</b> <b>D</b> ehydrogénase
<b>MET :</b>	<b>M</b> icroscopie <b>É</b> lectronique à <b>T</b> ransmission
<b>ML :</b>	Maximum de vraisemblance / <b>M</b> aximum <b>L</b> ikelihood
<b>MR :</b>	Récepteur des minéralocorticoïdes / <b>M</b> ineralocorticoid <b>R</b> eceptor
<b>MSA :</b>	Analyse statistique multivariée / <b>M</b> ultivariate <b>S</b> tatistical <b>A</b> nalysis
<b>MWC :</b>	<b>M</b> onod, <b>W</b> yman, et <b>C</b> hangeux
<b>NCBI :</b>	National Center for Biotechnology Information
<b>NCoR :</b>	Co-répresseur de récepteurs nucléaires / <b>N</b> uclear-receptor <b>C</b> o <b>R</b> epressor
<b>NCP :</b>	Particule nucléosomique / <b>N</b> ucleosome <b>C</b> ore <b>P</b> article
<b>NLS :</b>	Signal de localisation nucléaire / <b>N</b> uclear <b>L</b> ocalization <b>S</b> ignal
<b>NoSQL :</b>	<b>N</b> ot <b>O</b> nly <b>S</b> tructured <b>Q</b> uery <b>L</b> anguage
<b>p300 :</b>	Histone AcetylTransferase <b>p300</b>
<b>pCAF :</b>	<b>p300</b> Coactivation Associated Factor
<b>PDB :</b>	<b>P</b> rotein <b>D</b> ata <b>B</b> ank
<b>PDH :</b>	<b>P</b> yruvate <b>D</b> és <b>H</b> ydrogénase
<b>PDSM :</b>	Motif de sumoylation dépendant de la phosphorylation / <b>P</b> hosphorylation- <b>D</b> ependant <b>S</b> umoylation <b>M</b> otif
<b>PGC-1α :</b>	<b>P</b> eroxisome proliferator-activated receptor- <b>G</b> amma <b>C</b> oactivator- <b>1α</b>
<b>PHP :</b>	<b>H</b> ypertext <b>P</b> re <b>P</b> rocessor
<b>PNG :</b>	<b>P</b> ortable <b>N</b> etwork <b>G</b> raphics
<b>PPAR :</b>	<b>P</b> eroxisome <b>P</b> roliferator- <b>A</b> ctivated <b>R</b> eceptor
<b>PR :</b>	Récepteur de la progestérone / <b>P</b> rogesterone <b>R</b> eceptor
<b>PXR :</b>	<b>P</b> regnane <b>X</b> <b>R</b> eceptor
<b>RAR :</b>	Récepteur de l'acide rétinoïque / <b>R</b> etinoic <b>A</b> cid <b>R</b> eceptor
<b>RD :</b>	Domaine de répression / <b>R</b> epression <b>D</b> omain
<b>RMN :</b>	Résonance <b>M</b> agnétique <b>N</b> ucléaire

<b>RM</b> s :	<b>R</b> écepteurs <b>M</b> embranaires
<b>RN</b> s :	<b>R</b> écepteurs <b>N</b> ucléaires
<b>RRM</b> :	Motif de reconnaissance de l'ARN / <b>R</b> NA <b>R</b> ecognition <b>M</b> otif
<b>RS</b> :	<b>A</b> rginine/ <b>S</b> erine rich domain
<b>RXR</b> :	Récepteur X des rétinoïdes / <b>R</b> etinoid <b>X</b> <b>R</b> eceptor
<b>SANS</b> :	Diffusion des neutrons aux petits angles / <b>S</b> mall- <b>A</b> ngle <b>N</b> eutron <b>S</b> cattering
<b>SANT</b> :	<b>SWI3</b> , <b>ADA2</b> , <b>N-CoR</b> et <b>TFIIIB</b>
<b>SAXS</b> :	Diffusion des rayons X aux petits angles / <b>S</b> mall- <b>A</b> ngle <b>X</b> -ray <b>S</b> cattering
<b>SRC-1</b> :	Coactivateur des récepteurs Stéroïdiens-1 / <b>S</b> teroid- <b>R</b> eceptor <b>C</b> oactivator- <b>1</b>
<b>SRC-3b</b> :	Coactivateur des récepteurs Stéroïdiens -3b / <b>S</b> teroid <b>R</b> eceptor <b>C</b> oactivator- <b>3b</b>
<b>TAD</b> :	domaine d'activation de la transcription / <b>T</b> ranscription <b>A</b> ctivation <b>D</b> omain
<b>TAR</b> :	<b>T</b> ape <b>A</b> Rchiver
<b>TFF1</b> :	Facteur de trèfle 1 / <b>T</b> re <b>F</b> oil <b>F</b> actor <b>1</b>
<b>TFIIH</b> :	Facteur de transcription II H / <b>T</b> ranscription <b>F</b> actor <b>II H</b>
<b>TR</b> :	Récepteur des hormones thyroïdiennes / <b>T</b> hyroid hormone <b>R</b> eceptor
<b>UCP-1</b> :	Thermogénine / <b>U</b> n <b>C</b> oupling <b>P</b> rotein- <b>1</b>
<b>UniProt</b> :	<b>U</b> niversal <b>P</b> rotein <b>R</b> esource
<b>USP</b> :	<b>U</b> ltra <b>S</b> piracle
<b>VDR</b> :	Récepteur de la vitamine D / <b>V</b> itamin <b>D</b> <b>R</b> eceptor
<b>VPP</b> :	<b>V</b> olta <b>P</b> hase <b>P</b> late
<b>YAML</b> :	<b>Y</b> et <b>A</b> nother <b>M</b> arkup <b>L</b> anguage

# Résumé de thèse

---

Les récepteurs nucléaires sont des protéines impliquées dans la régulation de la transcription des gènes. Ils ont une structure modulaire comprenant un domaine de liaison à l'ADN (DBD) et un domaine de liaison du ligand (LBD). Cette thèse a pour objectif la détermination de la structure de complexes fonctionnels du récepteur nucléaire ERR (récepteur apparenté-aux-œstrogènes ; Estrogen-related receptor) entier et de l'analyse des relations structure-fonction dans le but de mieux comprendre les mécanismes d'actions moléculaires. Pour ce faire nous avons utilisé une approche intégrative basée sur la bioinformatique et la biologie structurale, et plus particulièrement la cryo-microscopie électronique (cryo-ME) qui permet d'étudier des complexes de relativement petite taille tels que les récepteurs nucléaires. Ces recherches permettront de mieux comprendre les bases moléculaires de la régulation de la transcription par ERR et contribuerons au développement de nouvelles stratégies thérapeutiques et pharmacologiques.

La thèse comporte deux volets :

- Une analyse bioinformatique des récepteurs nucléaires à partir des bases de données de séquences et de structures afin de positionner ERR $\alpha$  dans la famille des récepteurs nucléaires d'un point de vue évolutif.
- La détermination de la structure 3D du récepteur nucléaire ERR $\alpha$  en complexe avec son ADN et un coactivateur PGC-1 $\alpha$ .

## 1) Etude bioinformatique

L'analyse de séquences et de structures porte sur l'ensemble des séquences et structures de récepteurs nucléaires connus à ce jour. Cette étude a permis de caractériser et de corrélérer les différences fonctionnelles et structurales entre différents récepteurs nucléaires. Cette approche a déjà été utilisée dans le cadre d'une revue (Beinsteiner and Moras, 2015). Au cours de ma thèse j'ai mis l'accent sur la comparaison de ERR avec les autres membres des familles de récepteurs stéroïdiens (ER) et oxo-stéroïdiens (AR, GR, MR et PR). Deux aspects en particulier ont été abordés, la dimérisation des LBD et la reconnaissance de l'ADN.

Cette analyse a également permis de mettre en évidence des détails importants sur l'évolution des récepteurs nucléaires, notamment sur la présence d'une particularité structurale présente chez les récepteurs nucléaires ancestraux et certains de leurs descendants qui a probablement permis le passage entre les homodimères ancestraux et présents, et les hétérodimères actuels, c'est à dire les récepteurs nucléaires de classe II qui dimérisent avec le récepteur aux rétinoïdes X (RXR). Ces travaux seront bientôt soumis pour publication. Pour mener à bien ces recherches, en plus de l'utilisation des outils de bioinformatiques courants, nous avons développé un logiciel de détermination évolutive de séquences au sein d'une famille de protéines (BackPhylo, non publié), et une base de données intégré à plusieurs outils d'alignement de séquences et de structures du nom de IBiSS (Beinsteiner et al., 2015). Pour ce faire, avec un jeu de données de 30 000 séquences de récepteurs nucléaires, nous avons établi un historique de l'évolution de la famille des récepteurs nucléaires depuis leur apparition à nos jours. Par ce biais nous pouvons restituer notre cible d'intérêt dans un contexte plus large et comprendre également quels évènements d'évolution/mutation ont conduit à son fonctionnement actuel, notamment par rapport à un élément de structure caractéristique dans le LBD (le  $\pi$ -turn).

## 2 ) Détermination de la structure de divers complexes

ERR $\alpha$  est apparenté aux récepteurs des œstrogènes (ER $\alpha$  et ER $\beta$ ) et lie des séquences similaires à celles liées par les ERs sous forme d'homodimères. Au vue de la meilleure solubilité et stabilité de ERR lors de sa purification biochimique, son utilisation comme système modèle des récepteurs nucléaires homodimériques nous permet d'étudier l'architecture moléculaire et l'organisation topologique des récepteurs nucléaires stéroïdiens homodimériques, le but étant d'analyser le complexe entier en complexe avec l'ADN, et en présence ou en absence de coactivateurs.

La singularité du récepteur ERR se traduit également par son mode d'interaction avec l'ADN; en effet, la quasi-totalité des récepteurs nucléaires dimérisent et de fait se lient à des éléments de réponse dimériques (2x6 nucléotides espacés par 1 ou plusieurs acides nucléiques). ERR est également homodimérique, en raison des fortes interactions entre LBD, cependant le dimère se lie préférentiellement à un élément de réponse monomérique comprenant un demi-site étendu. La raison moléculaire de ce mode de fixation n'était pas connue et constituait l'un des points d'intérêts de l'étude structurale. En effet cette spécificité d'interaction devrait permettre de comprendre le décodage de l'information contenue dans les séquences d'ADN des éléments de réponse des promoteurs, à partir de l'analyse de l'interface avec l'ADN dans la présente structure et sa comparaison avec celles de complexes liés à des séquences consensus.

L'expression et la purification du récepteur nucléaire ERR $\alpha$  et du coactivateur PGC-1 $\alpha$  ainsi que la formation des différents complexes ont été optimisées au sein de l'équipe pour les études structurales de cryo-ME. Je me suis donc focalisé sur l'analyse structurale du complexe par cryo-ME. Après une optimisation des conditions adéquates pour la production de grilles de cryo-ME qui a nécessité de très nombreux tests, j'ai réussi à produire de façon reproductible des grilles de qualité (bonne concentration, distribution homogène des complexes, stabilisation du complexe lors de l'étape de cryogénéisation, conditions de cryogénéisation). Cela nous a permis d'acquérir des images de bonne qualité sur les microscopes électroniques à haute résolution Polara et Titan Krios de l'institut (CBI-IGBMC).

Une première structure (ERR $\alpha$  + ADN + PGC- $\alpha$ 1) est en cours d'affinement. Cette première structure devrait permettre de positionner les LBD et DBD dont les structures cristallographiques sont connues individuellement; elle sera un point de départ pour des futurs traitements d'images pour atteindre par la suite la haute résolution (~3-4 Å) dans laquelle on pourra construire un modèle atomique plus détaillé.

Les deux approches complémentaires de ma thèse permettent d'une part une meilleure compréhension générale sur l'origine et l'évolution des récepteurs nucléaires; d'autre part de nouvelles connaissances structurales sur ERR, récepteur nucléaire à l'origine du groupe des récepteurs stéroïdiens. Son étude est intéressante pour la connaissance fondamentale qu'il peut apporter sur la famille des stéroïdiens mais aussi pour des applications biomédicales puisqu'il est une importante cible pharmaceutique, notamment dans le cas de plusieurs cancers.



# Introduction

---

## Introduction

La cellule est l'unité biologique fondamentale de tout organisme vivant. Les traces de vie les plus anciennes connues sont des cellules procaryotes. Par la suite au cours de l'évolution apparaissent les organismes eucaryotes, d'abord unicellulaires puis pluricellulaires. La multicellularité est apparue plusieurs fois au cours de l'évolution, et ceci, aussi chez des procaryotes. C'est le cas par exemple chez certaines cyanobactéries, actinomycètes ou encore quelques archées qui présentent une organisation multicellulaire. Dans ces cas, il n'est pas question d'un organisme pluricellulaire, mais bien d'une colonie d'organismes qui collaborent et commencent à se spécialiser tout en restant potentiellement indépendants. Les organismes pluricellulaires sont spécifiques des eucaryotes et sont représentés dans les trois règnes suivants : *Plantae*, *Fungi* et *Animalia*. Les plus anciens pluricellulaires connus sont datés de 2.1 milliards d'années et ont été découverts au Gabon (Albani et al., 2010).

Les eucaryotes pluricellulaires peuvent avoir plus de 100 types de cellules différentes, communément regroupés en organes qui jouent alors des rôles et fonctions précis pour l'intégrité de l'organisme entier : on parle de spécialisation cellulaire. Coordonner ses fonctions et les utiliser le moment propice est important afin de permettre à l'organisme de s'adapter au mieux à son milieu. Il est alors nécessaire d'établir des voies de communications et de signalisations cellulaires pour transférer une information entre les cellules elles-mêmes et entre les cellules et l'environnement. Il y a plusieurs voies de communications qui se sont mises en place au cours de l'évolution, mais les principes de bases sont toujours les mêmes. Il faut un récepteur cellulaire qui détecte le signal, qui peut être à l'intérieur comme à l'extérieur de la cellule. Il transmet l'information à des effecteurs qui permettent l'intégration et la propagation du signal ce qui conduit en fin de chaîne à l'adaptation du comportement cellulaire. Il est très courant que cette adaptation soit une modulation de l'expression génétique de la cellule. Une des solutions importantes apparues pour ces voies de communication réside dans le concept d'allostérie. Dans ce cas la fixation d'une molécule effectrice induit un changement de conformation de la cible, une protéine enzymatique. Cet effet est un effet indirect. Ce faisant, il y a un changement de conformation de l'enzyme et un ou plusieurs site(s) actif(s) deviennent actif ou inactif.

Il y a deux modèles allostériques principaux, le modèle MWC (Monod, Wyman, and Changeux) (Monod et al., 1965) et le modèle KNF (Koshland, Nemethy, and Filmer) (Koshland et al., 1966). Dans le modèle MWC, on considère une protéine régulée telle qu'une enzyme ou un récepteur existe dans différents états interchangeables en l'absence de tout régulateur. Le rapport des différents états conformationnels est déterminé par l'équilibre thermique. Ce modèle est défini par les règles suivantes:

- Une protéine allostérique est un oligomère de monomères qui sont liés de façon symétrique.
- Chaque monomère peut exister dans au minimum deux états conformationnels différents, désignés par T (pour Tendue, par convention la forme de faible affinité au ligand) et R (Pour relaxée, par convention la forme de forte affinité au ligand). Ces états sont en équilibre, que le ligand soit lié ou non au monomère.
- Le ligand peut se lier à un monomère dans l'une ou l'autre conformation. Seul le changement conformationnel modifie l'affinité d'un monomère pour le ligand. Les régulateurs ne font que déplacer l'équilibre vers un état ou un autre.

Dans ce modèle, tous les monomères adoptent la même conformation, R ou T, c'est à dire que la fixation d'un ligand ne modifie pas la conformation si les voisins n'ont pas aussi de ligand.

Dans le modèle KNF, on parle de modèle séquentiel. Ici la fixation d'un ligand change la conformation du monomère indépendamment des autres monomères, et favorise la fixation de ligand et le changement de conformation de ces derniers.

Plus récemment un modèle de propagation conformationnel (CS pour Conformational Spread) a été proposé (Bray and Duke, 2004). Il s'agit non plus de système allostérique uniquement formé de monomères identiques, mais de complexes qui peuvent être grands et avec une population hétérogène de protéines. Il y a en fonction du cas la possibilité que la fixation du ligand induise un changement conformationnel mais aussi que le changement conformationnel d'une protéine voisine identique ou différente puisse induire par cascade d'autres changements et ainsi permette le transfert du signal.

Il y a deux principaux mécanismes de reconnaissance et transmission des signaux externe à la cellule. Pour chacun d'eux, il est fréquent que leurs activations aboutissent au recrutement de facteurs de transcriptions avec plus ou moins d'intermédiaires. Il s'agit des récepteurs membranaires (RMs) et des récepteurs nucléaires (RNs).

Pour les récepteurs membranaires, la molécule de signalisation ne pénètre pas dans la cellule, elle est captée par les RMs dans le milieu extérieur. L'activation peut être faite par une petite molécule comme la dopamine ou l'adrénaline, par un petit peptide comme le glucagon, ou encore par une protéine comme les hormones de croissances et les interleukines. Après activation, le signal est transmis à l'intérieur de la cellule et s'en suit une réaction en chaîne jusqu'à atteindre des facteurs de transcription nucléaires comme par exemple le complexe AP-1 qui est une réponse aux facteurs de croissance (Vesely et al., 2009).

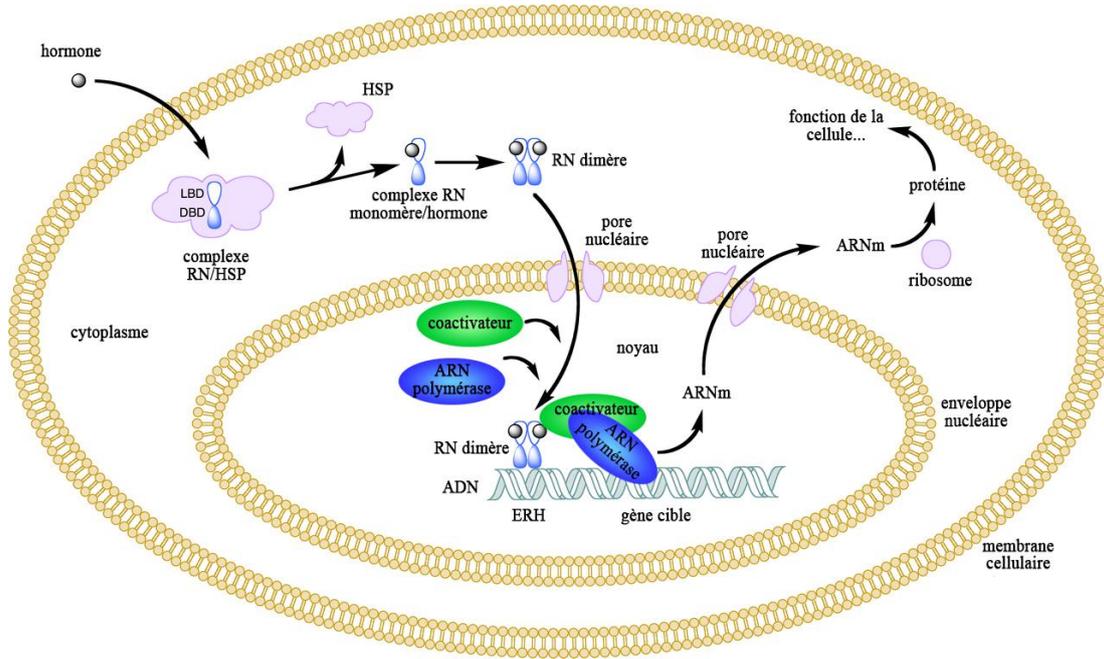
Les récepteurs nucléaires captent des ligands qui contrairement aux ligands des RMs peuvent entrer dans la cellule. Ils sont présents uniquement chez les Métazoaires (règne *Animalia*) et permettent de réguler directement la transcription des gènes. De nombreux récepteurs nucléaires fonctionnent directement comme des facteurs de transcription inductible par un ligand majoritairement hydrophobe pouvant avoir une extrémité hydrophile. Les RNs sont capables à la fois de capter un signal et d'induire la transcription directement contrairement aux RMs.

Ces différentes voies de communications, RMs et RNs fonctionnent en parallèle et interagissent entre elles pour moduler finement l'activité transcriptionnelle d'une cellule.

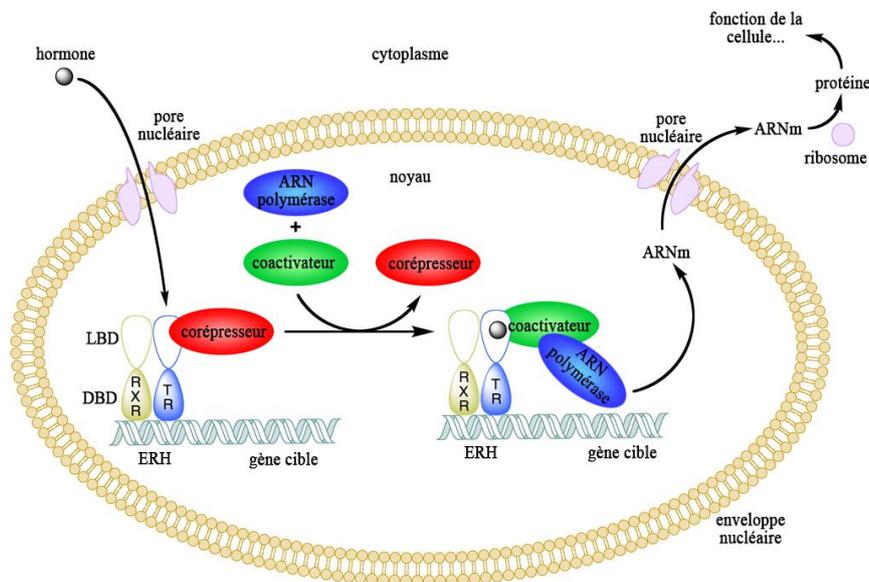
### 1. La superfamille des récepteurs nucléaires

#### 1.1 Fonctions biologiques

Durant toute la vie d'un organisme complexe, du développement embryonnaire à la croissance puis au maintien de l'homéostasie, il est nécessaire de réguler de manière précise l'expression des gènes, aussi bien spatialement que temporellement. Ces régulateurs sont appelés facteurs de transcription. La superfamille des récepteurs nucléaires est un des sous-groupes de facteurs de transcription. Ils sont potentiellement activables par un signal principalement lipophile. Ce signal peut être une hormone lipophile, telles que les hormones stéroïdiennes et thyroïdiennes qui jouent un rôle central dans de nombreux processus physiologiques. Elles sont capables de traverser librement les membranes cellulaires et de se lier à des récepteurs nucléaires dont la fonction est de transmettre le signal hormonal directement au noyau. Cependant, un grand nombre de récepteurs fonctionnent aussi sans ligand ou sans ligand connus à ce jour, ce sont des récepteurs nucléaires orphelins. D'un point de vue fonctionnel, on peut définir trois groupes de récepteurs nucléaires : les récepteurs endocriniens ayant un ligand de forte affinité, les récepteurs nucléaires orphelins "adoptés" possédant un ligand de faible affinité et les récepteurs nucléaires orphelins, pour lesquels aucun ligand naturel n'a encore été identifié à ce jour.



**Figure 1 :** Mécanisme d'action général d'un récepteur nucléaire de classe I ligand dépendant (homodimère).



**Figure 2 :** Mécanisme d'action général d'un récepteur nucléaire de classe II ligand dépendant (hétérodimère).

Certaines molécules doivent être d'abord métabolisées par la cellule avant de pouvoir être reconnues par le récepteur correspondant (proligands), c'est le cas par exemple de la vitamine D, ou encore pour la testostérone qui est d'abord transformé en androstanolone (DHT) avant d'être reconnues par AR. Certains ligands sont même directement synthétisés de novo par la cellule où se trouve le récepteur. Dans ce cas le ligand n'a pas un rôle endocrinien dans le sens classique du terme,

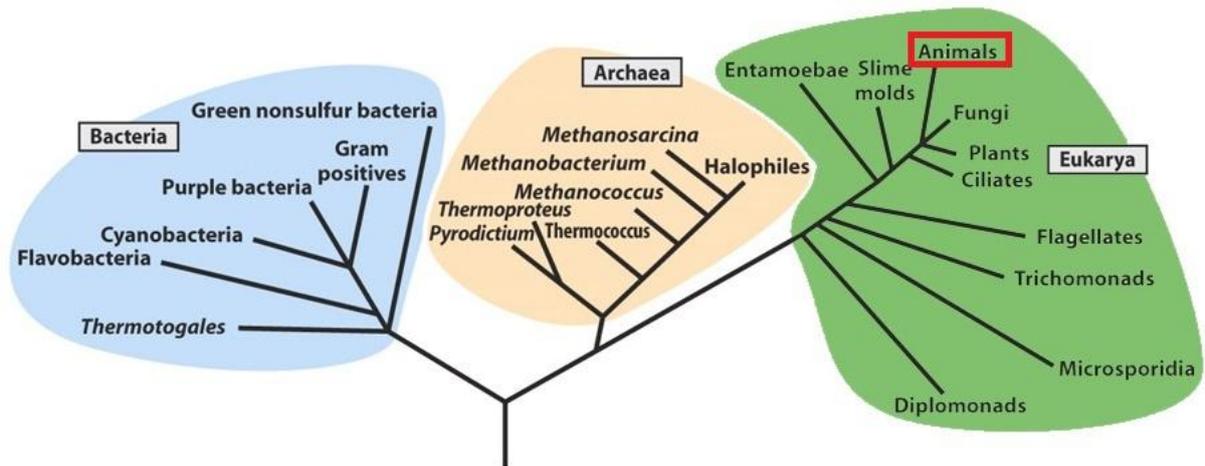
on dit qu'il agit de manière intracrine, par exemple les hormones stéroïdiennes ont une action intracrine en parallèle de leurs actions endocriniennes, car les cellules productrices les utilisent également. Ces mécanismes différents des voies endocriniennes classiques expliquent pourquoi avant les études par génomique de nombreux récepteurs n'ont pas pu être identifiés.

Pour les autres, une fois le complexe récepteur-ligand formé, ils agissent comme régulateurs de la transcription. Cette activité est également modulée par des co-facteurs activateurs ou inhibiteurs dont les interactions avec le récepteur est dépendant de la présence ou l'absence du ligand.

## 1.2 Origine et évolution

### 1.2.1 La systématique

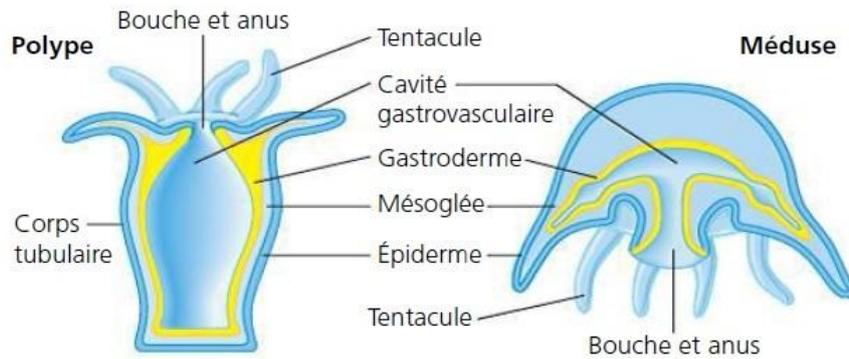
La systématique est un domaine de la biologie qui cherche à classer les organismes vivants, elle n'est pas synonyme de taxonomie qui elle cherche à décrire les taxons. Mais les deux disciplines vont de pair, une cherchant à décrire et l'autre à classer les taxons décrits les uns par rapport aux autres. Cette classification se fait suivant 2 critères majeurs. Premièrement, des critères phénotypiques qui concernent les données anatomiques et de développements. Deuxièmement et plus récemment, sur des critères génotypiques qui sont classiquement basés sur les ARN ribosomiques. Le but étant de déterminer les organismes et les taxons qui proviennent d'une même lignée évolutive. Cette classification est en continuel changement au rythme des nouvelles découvertes et des taxons sont fréquemment modifiés afin de mieux correspondre à la réalité de l'évolution. C'est principalement le cas pour les organismes unicellulaires ou multicellulaires qui ont une origine très ancienne et pour lesquels il manque de nombreux chaînons.



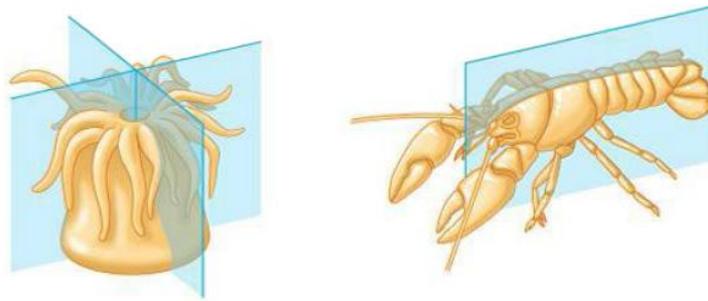
**Figure 3** : Arbre phylogénétique du vivant (Adaptée de Raven et al., 2005).

La racine de la branche du règne animal ou encore des métazoaires (Taxon *metazoa*) débute il y a environ 600 à 900 millions d'années avec l'apparition des Parazoaires (Du grec "à côté des animaux"). Ce nom est dû à l'absence de vrais tissus contrairement aux autres animaux qui eux sont donc situés dans le taxon des Eumetazoa (du grec "vrais animaux"). Les parazoaires regroupent environ 10 000 espèces d'éponges (aussi appelées spongiaires), qui sont des organismes aquatiques filtreurs ne présentant ni plan de symétrie ni vrai organe, mais uniquement des types de cellules différentes jouant le rôle des organes sans organisation particulière, elles sont toutes "en vrac" dans ce qu'on appelle la mésoglée.

Les Eumetazoas sont divisés en fonction de leur plan de symétrie. Tout d'abord les animaux à symétrie radiale qui sont également diploblastiques comme les spongiaires, c'est à dire que leur paroi corporelle est formée de deux couches, interne (endoderme) et externe (ectoderme) entre lesquelles est présente la mésoglée (mésenchyme).



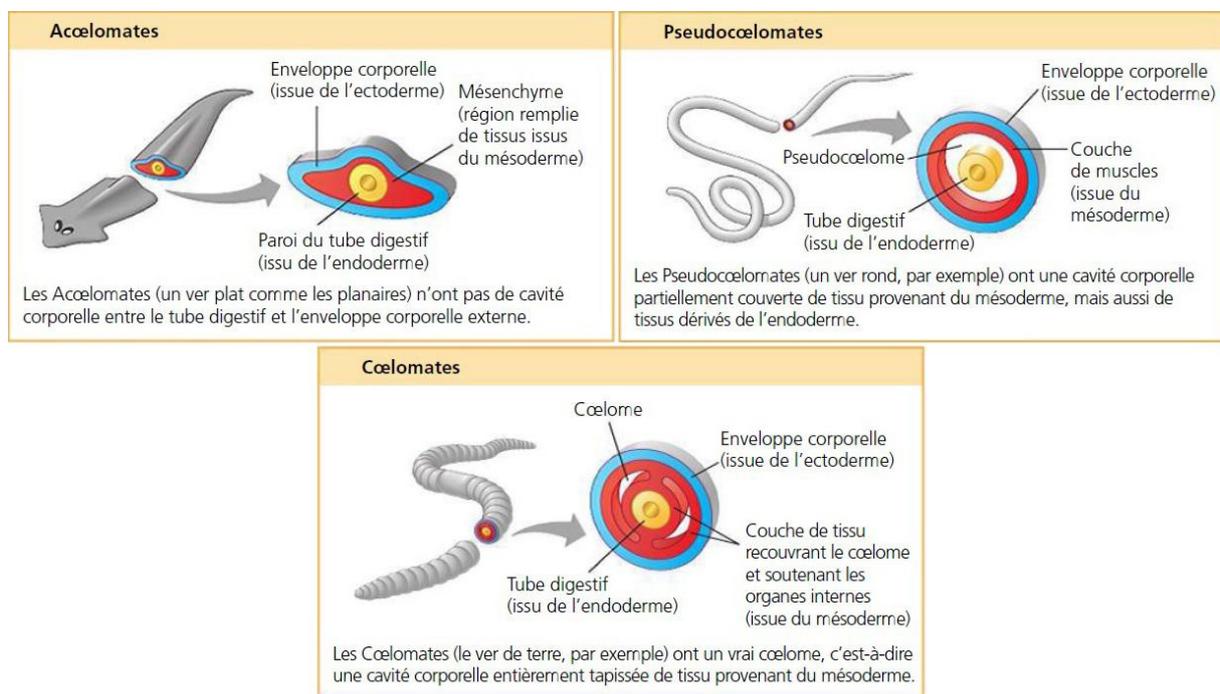
**Figure 4** : Anatomie de polype et méduse : les deux formes des Cnidaires. (Reproduit de Reece, 2011). L'enveloppe corporelle des Cnidaires se compose de deux couches de cellules. L'épiderme (en bleu foncé ; provenant de l'ectoderme) forme la couche externe, et le gastroderme (en jaune ; provenant de l'endoderme), la couche interne. Une couche gélatineuse et parfois épaisse, la mésogée (en bleu), sépare l'épiderme et le gastroderme.



**Figure 5** : La symétrie corporelle des Eumetazoas (Reproduit de Reece, 2011). A gauche, un exemple de symétrie radiaire avec une anémone de mer (Cnidaire). N'importe quelle coupe qui passe par le centre est un plan de symétrie. A droite un exemple de symétrie bilatérale avec un homard (Arthropode), un seul plan de symétrie est présent, il y a un côté gauche et un côté droit.

Il s'agit du taxon des Cnidaires (du grec "comme une ortie" à cause de leur propriétés souvent urticantes) (méduses, anémones, coraux) qui sont les premiers animaux à proprement parlé avec une organisation des cellules différenciées en tissus. L'autre grand groupe des Eumetazoas comprend tous les autres animaux qui suivent dans l'évolution, il s'agit du taxon des Bilateria, qui ont une symétrie bilatérale et sont des triploblastiques, c'est à dire sont composés de 3 feuilles distincts, l'ectoderme, le mésoderme et l'endoderme. Le mésoderme est un feuillet qui est à l'origine des appareils circulatoire, excréteur, génital et des muscles. Les bilatériens ont deux axes de polarité, un axe antéro-postérieur et un axe dorso-ventral.

Les bilatériens sont à leur tour divisés mais cette division est sujette à controverse entre des marqueurs purement génotypiques ou des marqueurs phénotypiques. Pour simplifier, la classification basée sur le phénotype permet de diviser les bilatériens en 3 groupes définis par la présence d'une cavité interne que l'on appelle le coelome. L'absence de coelome est caractéristique du groupe des Acoelomates. Ce groupe est essentiellement caractérisé par les Plathelminthes qui sont des vers plats. La présence d'une cavité qui n'est pas un vrai coelome forme le groupe des Pseudocoelomates caractérisé par l'embranchement des Nématelminthes. Enfin la présence d'un vrai coelome qui donne le groupe des Eucelomates qui comprend tous les autres animaux non encore cités.

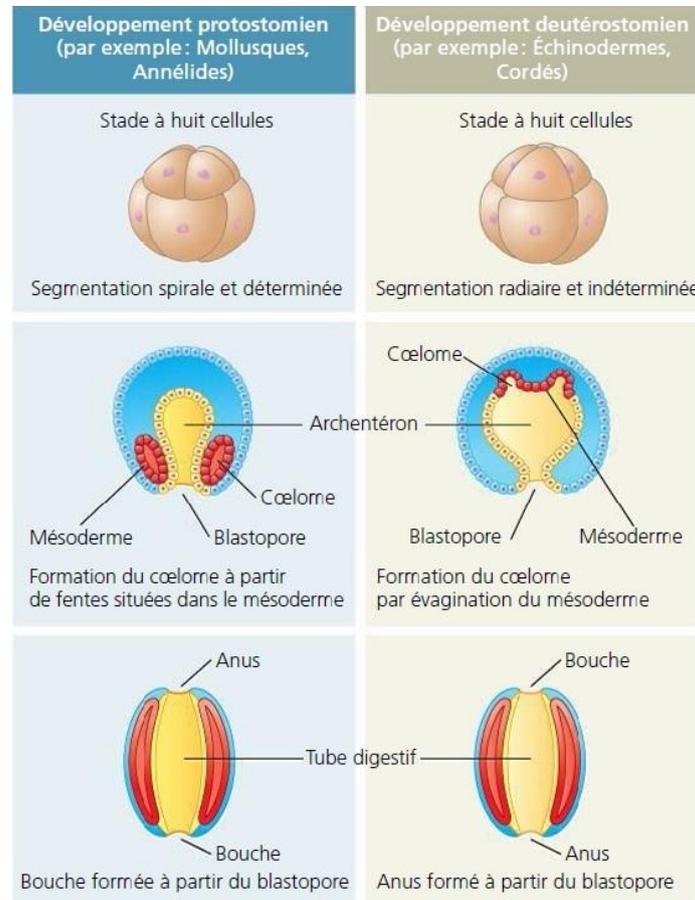


**Figure 6 :** Les cavités corporelles des animaux triploblastiques (Adaptée de Reece, 2011).

Les différents organes des animaux triploblastiques se développent à partir des trois feuillet embryonnaires. Par convention, les feuillet embryonnaires portent des couleurs précises : l'ectoderme est en bleu, le mésoderme en rouge et l'endoderme en jaune.

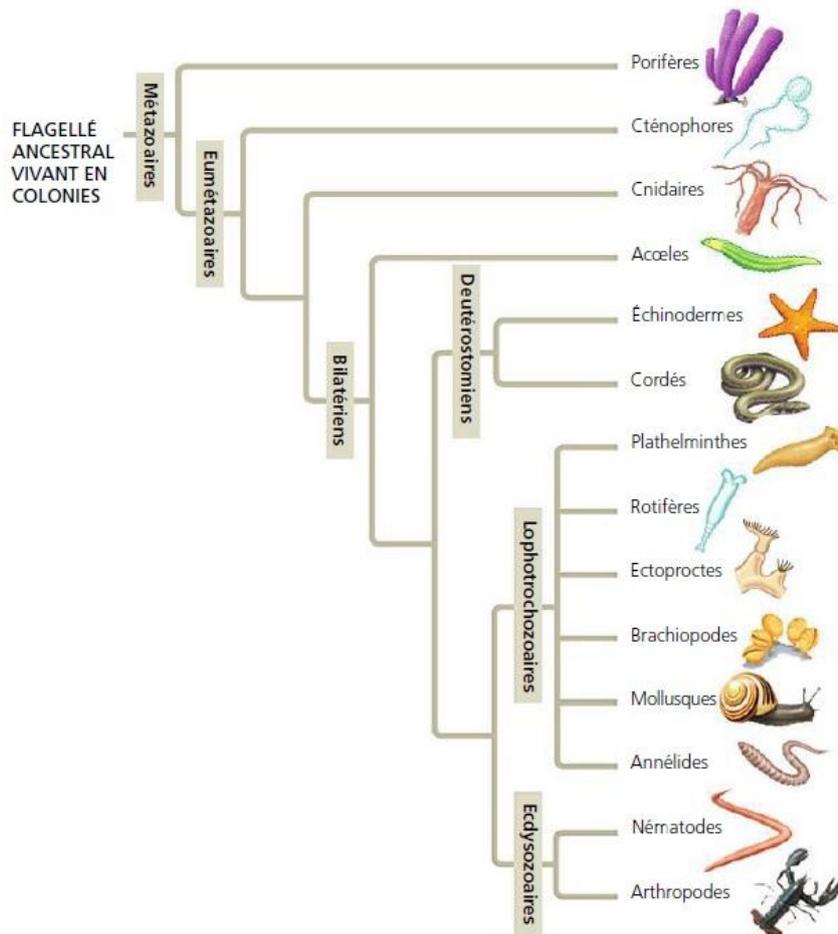
Ce dernier groupe est subdivisé en deux groupes, les protostomiens et les deutérostomiens. Actuellement, cette classification est remise en cause et déjà annulée dans certains cas, laissant place à une classification basée sur le génotype où les bilatériens comprennent uniquement deux groupes, les protostomiens et les deutérostomiens. En effet, l'absence et la pseudo absence de cavité serait une perte évolutive survenue plus tard. Dans ce cas les Acoelomates et les

Pseudocoelomates sont rattachés au groupe des protostomiens. Les protostomiens (du grec première bouche) et les deutérostomiens (du grec seconde bouche) sont différenciés en fonction du devenir du blastopore embryonnaire (premier orifice de la blastula au cours de l'embryogénèse).



**Figure 7** : Comparaison des modes de développement protostomien et deutérostomien (Reproduit de Reece, 2011). La première étape est la segmentation. La majorité des Protostomiens subissent une segmentation spirale et déterminée, tandis que la plupart des Deutérostomiens subissent une segmentation radiaire et indéterminée. La formation du cœlome se produit pendant le stade de la gastrula. Dans le développement protostomien, le cœlome se forme à partir de fentes situées dans le mésoderme. Dans le développement deutérostomien, il se forme par évagination du mésoderme depuis la paroi de l'archentéron. Le blastopore devient la bouche chez les Protostomiens, tandis que c'est l'ouverture du côté opposé qui devient la bouche chez les Deutérostomiens.

Dans le cas des protostomiens, le blastopore deviendra la bouche, dans le cas des deutérostomiens, le blastopore deviendra l'anus. Les protostomiens sont divisés en deux groupes majeurs, le groupe des Lophotrochozoaires, qui est le taxon donnant les groupes des mollusques, annélides et plathelminthes, et le groupe des Ecdysozoaires qui comprend les Arthropodes et les Nématodes. Les deutérostomiens quant à eux sont également divisés en deux groupes, le groupe des Échinodermes (étoiles de mer, oursins) et le groupe des Chordés (vertébrés, urochordés).

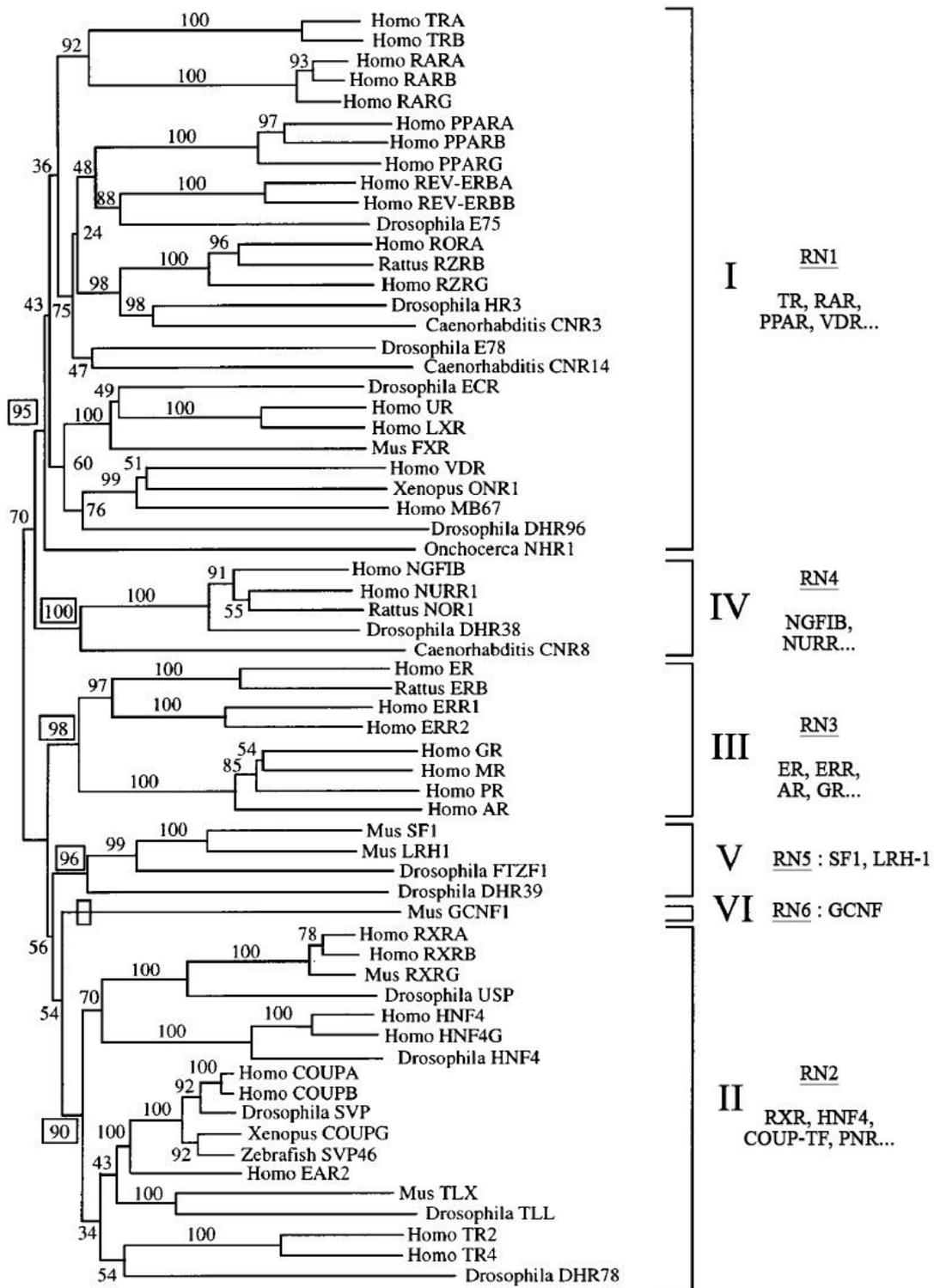


**Figure 8 :** Phylogénie des métazoaires basée sur des données moléculaires (Reproduit de Reece, 2011).

### 1.2.2 Origine et classification des récepteurs nucléaires

Au cours de l'évolution, la première trace connue des récepteurs nucléaires est dans le taxon des Parazoaires (Porifera, Spongiaires) qui est placé à la racine des métazoaires. Ils sont absents du taxon des Choanoflagellés qui sont les plus proches parents des métazoaires. Ils sont également retrouvés dans les taxons suivants comme les Cnidaires et les Triploblastiques etc.

Une analyse phylogénétique de l'ensemble des récepteurs nucléaires connus a permis de les classer en divisant la superfamille des récepteurs nucléaires en 6 sous-familles (Laudet, 1997).



**Figure 9 :** Les six sous-familles de la superfamille des récepteurs nucléaires basées sur l'étude phylogénétique de 63 gènes codant pour les récepteurs nucléaires des Vertébrés, Arthropodes et Nématodes (adaptée de Laudet, 1997).

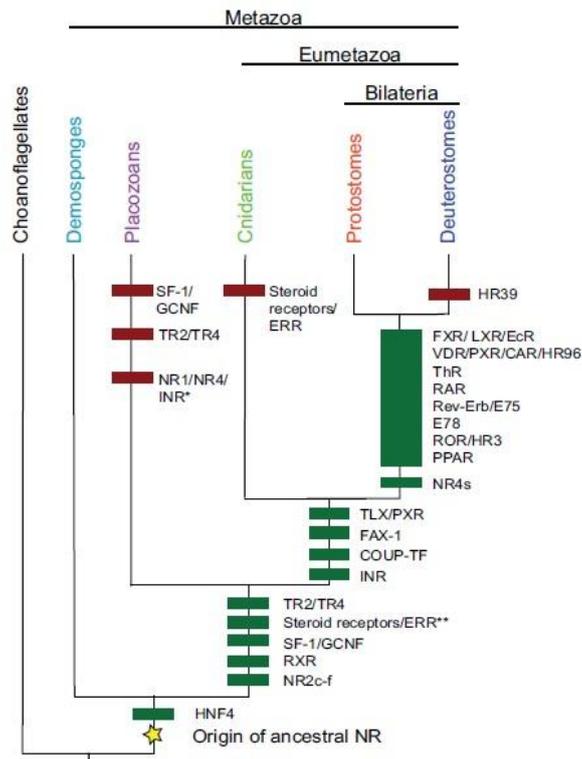
De cette classification découle une nomenclature précise afin de réduire la confusion due au grand nombre de récepteurs nucléaires et à leur nom, parfois différent pour le même récepteur. Le nom

assigné avec cette nouvelle nomenclature indique la sous-famille, le groupe puis la forme paralogue (Nuclear Receptors Nomenclature Committee 1999).

### 1.2.3 Diversification et évolution de la superfamille des récepteurs nucléaires

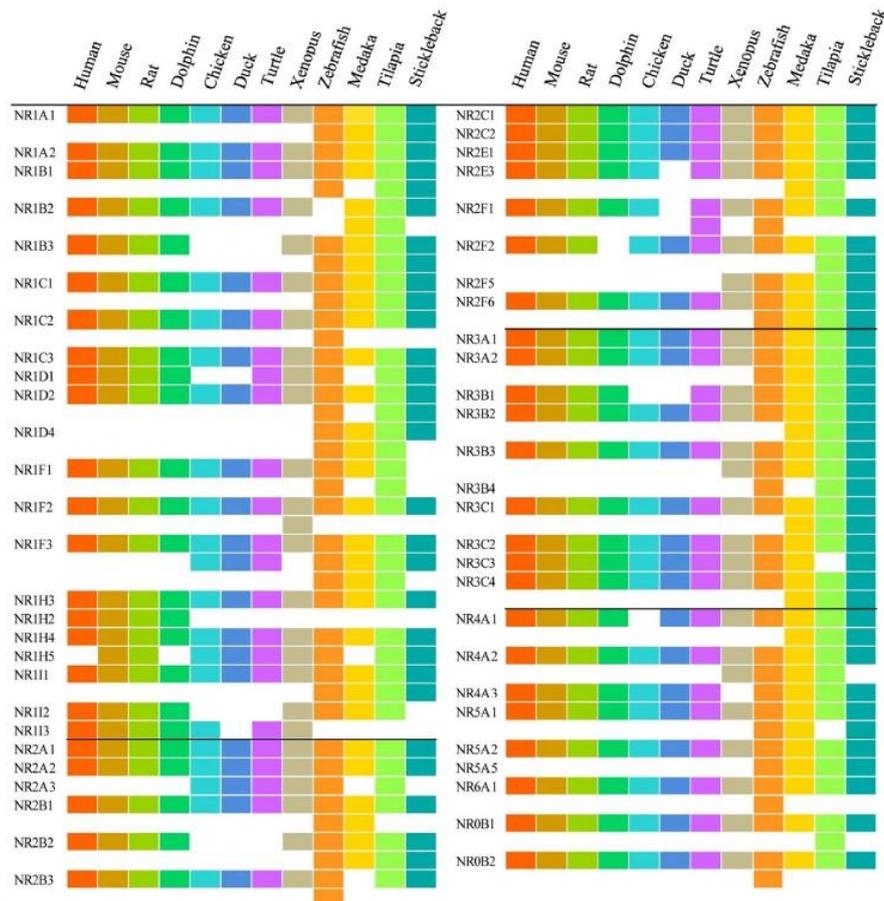
Un modèle concernant la diversification des récepteurs nucléaires a été proposé et décrit la succession de deux périodes importantes de duplications de gènes (Escriva et al., 2000.; Laudet, 1997). La première période serait survenue très tôt lors du développement des premiers métazoaires, un peu avant la divergence entre les cnidaires et les bilatériens. Cette première période a probablement permis de mettre en place les principales familles de Récepteurs nucléaires connues aujourd'hui. Vient ensuite la seconde période après la séparation entre les arthropodes et les vertébrés. Cette période propre aux vertébrés aurait conduit à l'émergence de nombreux paralogues à l'intérieur de chaque sous-famille. En parallèle à cette période de duplication vient s'ajouter des gains et des pertes de gènes qui jouent un rôle important dans l'évolution des différents Récepteurs nucléaires (Bertrand et al., 2004).

Pour illustrer ces évolutions de gènes, nous pouvons prendre la sous-famille des NR3 (récepteurs nucléaires stéroïdiens). En effet, aucun homologue de cette sous-famille n'a été caractérisé dans le groupe des arthropodes qui ont pourtant de nombreux génomes séquencés, rendant ce résultat fiable. Ces observations suggèrent que la sous famille NR3 s'est formée après la séparation entre les arthropodes et les vertébrés et donc qu'elle est spécifique aux Deutérostomiens. Cependant en 2003, un récepteur nucléaire ER d'un mollusque (*Aplysia californica*) a été découvert (Thornton et al., 2003). Cette observation laisse donc penser que la sous-famille NR3 était déjà présente à l'émergence des premiers bilatériens et qu'ils ont été perdus dans certain taxons. Une étude plus récente de 2010 met HNF-4 comme premier récepteur nucléaire ancestral présent dès le taxon des Poriferas (Bridgham et al., 2010).



**Figure 10** : Duplications et pertes des nucléaires récepteurs dans l'évolution des Métazoaires. Les barres vertes représentent les duplications et les barres rouges les délétions. Les duplications sont étiquetées avec les lignées nommées de Récepteurs nucléaires qu'elles ont générées. (Reproduit de Bridgham et al., 2010)

Les récepteurs nucléaires sont présents chez tous les métazoaires et sont propres à ce taxon, par contre leur distribution n'est pas la même dans tous les taxons. Cela permet de constater les évènements de délétions et de duplications qui ont eu lieu au cours de l'évolution. Il y a eu 2 périodes qui ont eu de nombreuses duplications, la première au niveau basal des métazoaires et la seconde chez les vertébrés. On compte par exemple 2 récepteurs nucléaires chez les Poriferas et 3 chez les Cnidaires. Il y a une particularité parmi les nématodes, en effet chez *Caenorhabditis elegans* il a été constaté la présence de plus de 270 récepteurs nucléaires. Parmi ces nombreux récepteurs se trouvent 13 orthologues de récepteurs nucléaires présents chez les arthropodes et les vertébrés. Les récepteurs nucléaires supplémentaires proviennent d'un grand nombre de duplications du récepteur orphelin HNF-4. Chez les différentes familles de vertébrés, nous retrouvons pour les mammifères 48 récepteurs nucléaires chez l'homme, 49 chez la souris et le rat et 47 chez le dauphin. Les oiseaux en ont un peu moins avec 44 chez la poule et 42 chez le canard. Les tortues qui représente les reptiles en ont 48 comme pour l'homme, et il y en a 52 chez la grenouille, représentant ainsi le taxon des amphibiens. Chez les poissons par contre on retrouve environ 70 récepteurs nucléaires. Il y a en a 73 pour le poisson-zèbre (zebrafish), 67 pour le médaka, 74 pour le tilapia et 66 chez l'épinoche qui est un poisson osseux (Zhao et al., 2015).

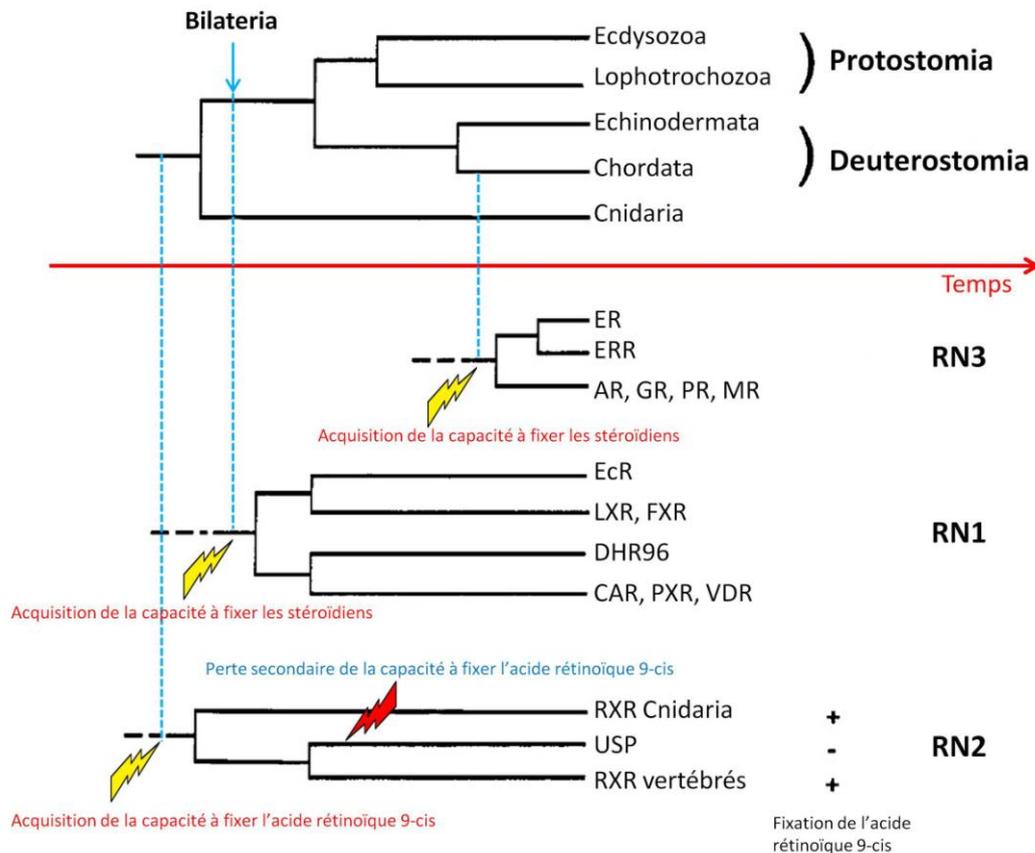


**Figure 11** : Répartition des récepteurs nucléaires pour 12 modèles vertébrés. (Reproduit de Zhao et al., 2015)

#### 1.2.4 Acquisition de la capacité à fixer un ligand

A partir de l'arbre phylogénétique des récepteurs nucléaires, on s'aperçoit que les récepteurs nucléaires fixant un ligand ne sont pas regroupés entre eux mais sont répartis dans l'ensemble de l'arbre phylogénétique. De plus, les récepteurs fixant une même nature de ligand ne sont pas regroupés non plus. En effet, si on prend des récepteurs nucléaires de classes II comme TR et VDR (NR1) qui fixent respectivement hormones thyroïdiennes et vitamine D, ils sont proches phylogénétiquement mais n'ont pas de ligands se ressemblant. Par contre, si l'on regarde RAR (NR1) et RXR (NR2) qui sont très différents, ils peuvent fixer tous les deux le même ligand, l'acide rétinoïque all-trans et 9-cis. Il y a donc très probablement des gains de fonctions indépendants quant à la capacité de fixer un ligand. Il est aussi important de noter que contrairement aux récepteurs orphelins que l'on retrouve dans un large éventail de métazoaires, la plupart des récepteurs nucléaires de vertébrés ayant un ligand n'ont pas d'homologues chez les arthropodes. Les Récepteurs nucléaires

ancestraux seraient dans ce contexte alors orphelins, et les Récepteurs nucléaires à ligand seraient plus récents dans l'évolution (Detera-Wadleigh and Fanning, 1994; Escriva et al., 2000; Laudet, 1997).



**Figure 12** : Schéma illustrant l'acquisition de la capacité de fixation du ligand (adapté de Escriva et al. 1997). Arbre phylogénétique schématique des Métazoaires (en haut). Les différentes périodes durant lesquelles la fixation du ligand a été acquise au cours de l'évolution sont indiquées pour les récepteurs NR3, NR1H et NR1I, ainsi que la perte secondaire de la capacité à fixer l'acide rétinoïque 9-cis propre aux arthropodes.

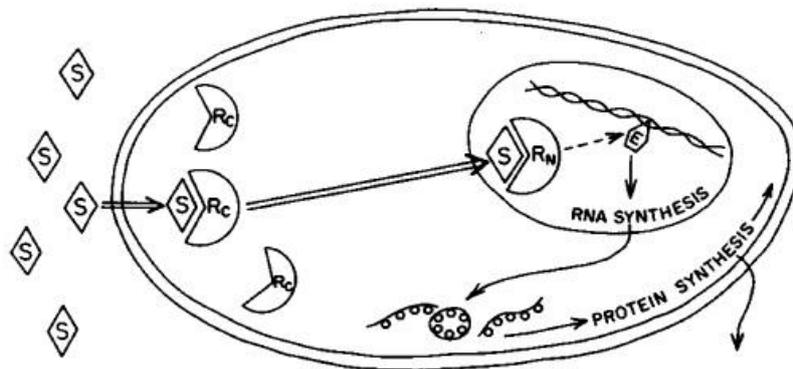
On peut prendre l'exemple de RXR qui peut fixer l'acide rétinoïque 9-cis déjà chez les Cnidaires. Or, son homologue chez les Arthropodes, appelé USP, ne peut pas fixer cette molécule (Oro et al., 1990). Au début on pensait que la capacité à fixer un ligand a été acquise très tôt pour ce récepteur nucléaire, puis perdu dans la branche des Arthropodes. Cependant, depuis, il y a un débat quant à la véritable nature du ligand naturel, actuellement aucune donnée ne soutient l'hypothèse que RXR ait un ligand naturel car le ligand physiologique n'a pas été identifié jusqu'à présent. RXR est donc considéré comme étant un récepteur nucléaire orphelin (Iwema et al., 2007).

Ce modèle d'évolution des récepteurs nucléaires pose la question, notamment, du mode de fonctionnement des récepteurs nucléaires en absence de ligands. En effet, nous savons que pour les

Récepteurs nucléaires dont nous connaissons le ligand, une fixation de ce dernier induit un changement de conformation permettant le recrutement de cofacteurs pour réguler la transcription des gènes (Horwitz et al., 1996). Les structures des récepteurs nucléaires orphelins avec des ligands synthétiques montrent qu'ils fonctionnent de la même manière. Cependant nous savons également que certains récepteurs comme le récepteur aux œstrogènes peuvent être activés par d'autres processus comme par exemple la phosphorylation. Nous pouvons faire l'hypothèse que les premiers Récepteurs nucléaires étaient régulés de cette manière, comme des facteurs de transcription classiques dont l'activité est induite par des phosphorylations par exemple. La capacité à fixer un ligand serait donc venue plus tard de façon à affiner et diversifier les modes de régulations de cette superfamille de protéines.

### 1.3 Découverte et évolution des connaissances

Avant la découverte des récepteurs nucléaires, on a longtemps supposé leur existence. En 1905, Ernest Starling introduit pour la première fois le terme "hormone" à partir d'observation de la sécrétion pancréatique qui était provoquée par un stimulus sanguin alors inconnu. En 1926, Edward Calvin Kendall et Tadeus Reichstein isolent et déterminent les structures de la cortisone et de la thyroxine, qui sont respectivement une hormone glucocorticoïdes synthétisée par les glandes surrénales et une hormone thyroïdienne agissant comme une pro hormone et synthétisée dans la thyroïde. En 1929, de manière totalement indépendante du premier groupe, Adolf Butenandt et Edward Adelbert Doisy isolent et déterminent la structure des œstrogènes, hormones stéroïdiennes principalement produites dans les gonades femelles. Le premier récepteur nucléaire découvert est le récepteur d'œstrogènes par Elwood V Jensen en 1958 (Jensen, 1962). Il a montré que seuls les tissus répondant aux œstrogènes, tels que ceux de l'appareil reproducteur féminin, étaient capables de concentrer l'œstradiol injecté dans le sang. Ce schéma d'absorption suggère que ces cellules doivent contenir des protéines de liaison spécifique, qu'il appelle « récepteurs d'œstrogènes ». Il a ensuite identifié le récepteur d'œstrogène dans les cellules de l'utérus. En 1971 Bert W O'Malley décrit pour la première fois une théorie d'action des récepteurs nucléaires stéroïdiens (O'Malley, 1971).



**Figure 13** : Premier postulat du mécanisme du récepteur d'œstrogène (O'Malley, 1971).

En 1975, Keith R. Yamamoto confirme cette théorie en démontrant que la liaison du ligand 17 $\beta$ -œstradiol au récepteur aux œstrogènes induit la translocation du complexe dans le noyau cellulaire et sa liaison à des régions précises de l'ADN (Yamamoto and Alberts, 1975).

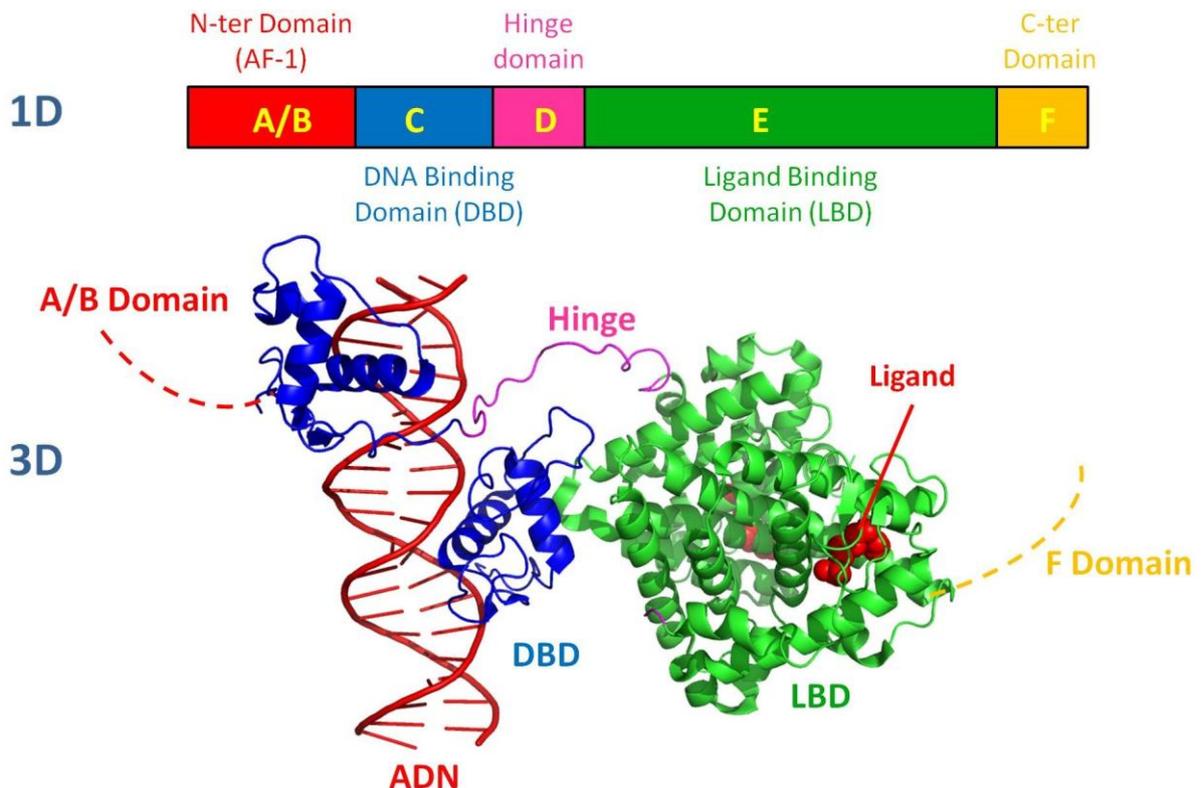
En 1985, le premier clonage de récepteur nucléaire est réalisé, il s'agit du récepteur nucléaire des glucocorticoïdes (GR) (Hollenberg et al., 1985) dans l'équipe de Ronald M Evans. En 1986, 3 nouveaux récepteurs nucléaires sont clonés avec succès. Le récepteur nucléaire aux œstrogènes (ER) est cloné (Jeltsch et al., 1986) dans l'équipe de Pierre Chambon. Dans le même temps, l'équipe de Björn Vennström démontre en 1986 que le gène c-erbA code pour un gène de récepteur nucléaire thyroïdien (TR) (Sap et al., 1986), ce même gène qu'ils avaient cloné deux ans plus tôt en 1984 (Spurr et al., 1984). Le récepteur de la progestérone (PR) quant à lui est cloné par l'équipe de Henri Rochefort (Chalbos et al., 1986). L'identification et le clonage de plusieurs récepteurs nucléaires a permis d'observer une forte homologie de séquences entre eux, notamment le domaine de liaison à l'ADN (DBD : DNA binding domain) qui est fortement conservé dans la famille. Une recherche basée sur ces caractéristiques génétiques a permis l'identification de nouveaux gènes. Le premier récepteur nucléaire à être découvert par cette méthode est le récepteur de l'acide rétinoïque (RAR $\alpha$ ) en 1987 dans l'équipe de Pierre Chambon (Petkovich et al., 1987). Cette technique de criblage est le début d'un nouveau concept, "l'endocrinologie inverse" décrit par l'équipe de Timothy M. Willson (Kliwer et al., 1999).

Dans ce contexte la découverte du gène précède la découverte du ligand et de la fonction biologique. L'avantage majeur de cette méthode est qu'elle permet également la découverte de récepteurs nucléaires sans ligand, ces Récepteurs nucléaires sont dit orphelins. Elle a également permis la découverte de voies de signalisation nucléaire insoupçonnées pour les rétinoïdes, les acides gras, les eicosanoïdes et les stéroïdes, avec d'importantes ramifications physiologiques et pharmacologiques.

Avec le nombre croissant de récepteurs nucléaires identifiés chez une multitude d'organismes une classification de la superfamille des récepteurs nucléaires en 6 familles a été proposée en 1997 (Laudet, 1997). Plus récemment une analyse de conservation d'acides aminés dans les séquences du domaine de liaison du ligand des Récepteurs nucléaires a permis de déterminer des résidus signatures et ainsi de classer les Récepteurs nucléaires en 2 classes, regroupant pour l'une d'elle (classe I) les homodimères et pour l'autre (classe II) les hétérodimères avec le partenaire ubiquitaire d'hétérodimérisation RXR. Nous détaillons cette classification ultérieures dans ce manuscrit (Brelivet et al., 2004).

#### 1.4 Structure modulaire

Les récepteurs nucléaires sont organisés en une architecture modulaire commune de 5 à 6 domaines (Krust et al., 1986)



**Figure 14** : Organisation structurale générale des récepteurs nucléaires. Ici pour l'exemple il s'agit de HNF-4 PDB 4IQR (Chandra et al., 2013).

Chacun des domaines est fonctionnellement autonome et l'ensemble de leurs activités permettent d'assurer l'activité de régulation transcriptionnelle.

---

#### 1.4.1 Le domaine A/B

Le domaine amino-terminal A/B est très variable en séquence et en longueur, ce qui en fait la région la moins conservée des récepteurs nucléaires. Par exemple, il représente une vingtaine de résidus seulement pour le récepteur nucléaire de la vitamine D (VDR) mais plus de 600 résidus pour le récepteur nucléaire au minéralocorticoïdes (MR). Le domaine A/B porte ce nom car historiquement, la première définition des domaines (Krust et al., 1986) est basé sur le récepteur aux œstrogènes (ER) humain où les domaines A et B peuvent être distingué, mais dans l'ensemble des autres cas ce n'est pas possible, d'où l'appellation domaine A/B.

Il contient une fonction d'activation de la transcription indépendante de la fixation du ligand AF-1 (Activation-Function 1). Cette fonction est régulée par des modifications post-traductionnelles, en effet c'est une cible de phosphorylation (Wärnmark et al., 2003). La phosphorylation de ER $\alpha$  par le facteur général de transcription TFIIH est ligand-dépendante (Chen et al., 2000), mais celle de ER $\beta$  par la kinase MAPK (Mitogen-Activated Protein Kinase) est ligand-indépendante. Cette dernière permet le recrutement du coactivateur (CoA) SRC-1 (Tremblay et al., 1999). La kinase MAPK étant impliquée dans la voie de communication des récepteurs membranaires, cet exemple illustre la communication parfois croisée entre les voies des RNs et des RMs. Ces phosphorylations permettent le recrutement de différents CoA et régulent l'activité AF-1 de manière importante. Cette régulation peut mener à l'activation du récepteur nucléaire en absence de ligand.

Il n'y a à ce jour aucune donnée cristallographique concernant le domaine A/B. En effet ce domaine est probablement désordonné en absence de partenaire (Dahlman-Wright and McEwan, 1996). Cette caractéristique est valable pour beaucoup de domaines d'activation des facteurs de transcriptions chez les eucaryotes. Des données de résonance magnétique nucléaire (RMN) montrent que la région AF-1 est en effet désordonnée en solution.

#### 1.4.2 Le domaine C

Le domaine C, plus communément appelé domaine de liaison à l'ADN (DBD : DNA Binding Domain), est un domaine comprenant environ 70 à 80 résidus. Contrairement au domaine A/B, le domaine C est le domaine le mieux conservé parmi les récepteurs nucléaires. Ce DBD permet la reconnaissance spécifique et la fixation à l'ADN cible.

Il y a à ce jour de très nombreuses structures cristallographiques et par RMN disponibles pour les DBD d'un grand éventail de Récepteurs nucléaires. Les deux premières structures de DBD liés à leur ADN sont celle de GR et de ER, tous deux des homodimères (Luisi et al., 1991; Schwabe et al., 1993). Il est constitué de 2 doigts de zinc de type C2-C2 contenant chacun 4 cystéines chélatant un ion de zinc ( $Zn^{2+}$ ) (Freedman et al., 1988), ainsi que 2 hélices  $\alpha$  perpendiculaires entre elles qui interagissent spécifiquement avec les éléments de réponse hormonale (HRE Hormone-Response Element) qui sont situés au niveau des promoteurs des gènes. Les analyses de mutation suggèrent que l'hélice  $\alpha$  est indispensable pour la liaison à l'ADN. La séquence des sites de liaison de GR affecte de manière différentielle la conformation du récepteur et l'activité transcriptionnelle (Meijsing et al., 2009). Bien que seuls des changements structurels mineurs aient pu être observés lors de la comparaison des structures cristallines des DBD liés à différents éléments de réponse, une corrélation avec l'activité des GR supporte la proposition selon laquelle l'ADN serait un effecteur allostérique pour moduler l'activité du récepteur. Pour chaque DBD un seul motif de liaison au zinc est impliqué dans la liaison à l'ADN, l'autre permettant la dimérisation au contact de l'ADN. Plusieurs éléments ont été caractérisés sur ce domaine C, il s'agit des boîtes P, D, A et T. La boîte P (Proximale) est impliquée dans la reconnaissance du demi-site de l'élément de réponse sur l'ADN (Giguère, 1999). Elle se situe au niveau de la première hélice du DBD et du premier doigt de zinc. La boîte D (Distale) est impliquée dans l'interface de dimérisation des DBD, notamment pour les récepteurs nucléaires stéroïdiens pour qui la boîte D favorise les interactions pour des éléments de réponses de type palindromique (IR inverse repeat) (Aumais et al., 1996). Cette boîte se situe au niveau du second doigt de zinc. Les boîtes T et A interviennent dans la dimérisation et la reconnaissance de l'ADN au niveau du petit sillon. Cette action a été caractérisée pour le cas de TR et RXR. Ces deux boîtes sont situées dans ce qu'on appelle l'extension carboxyl-terminale (CTE). (Giguère, 1999)

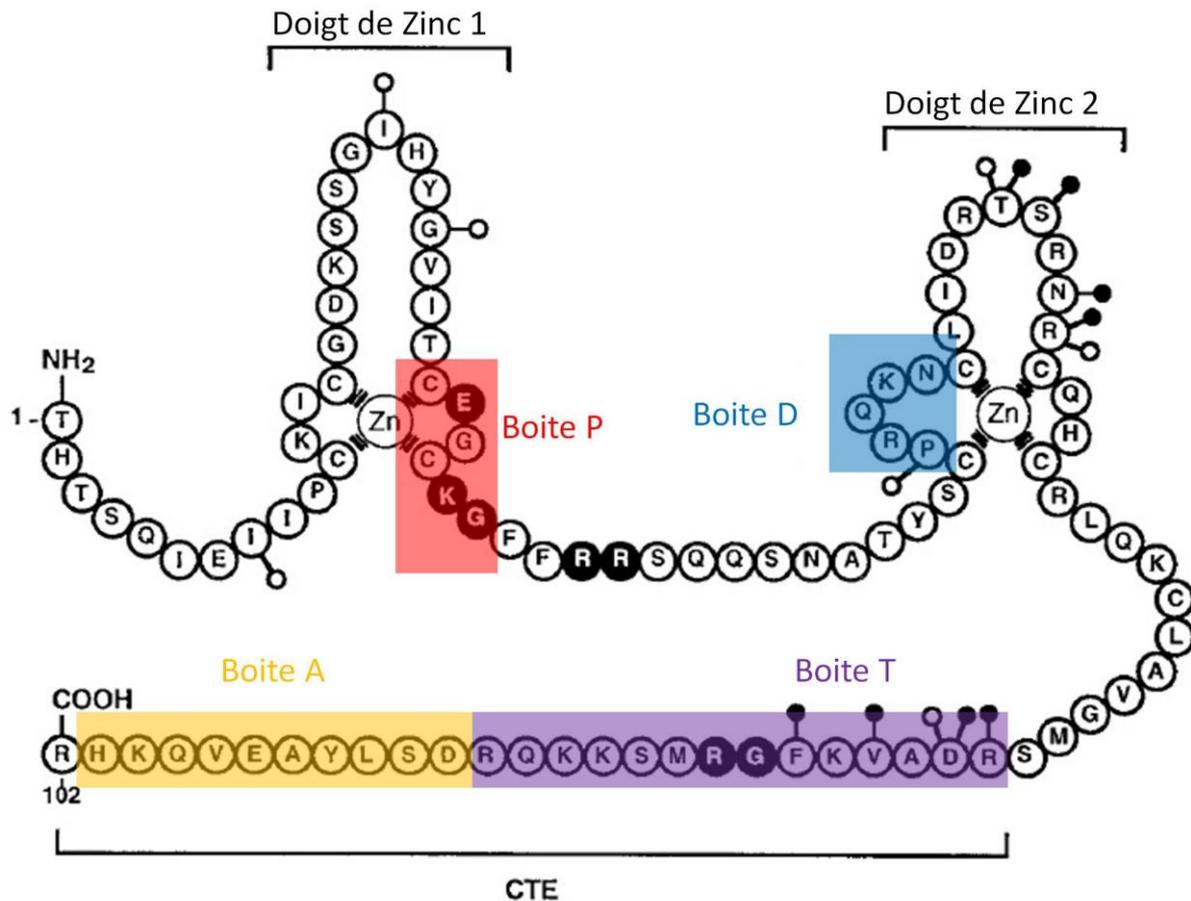


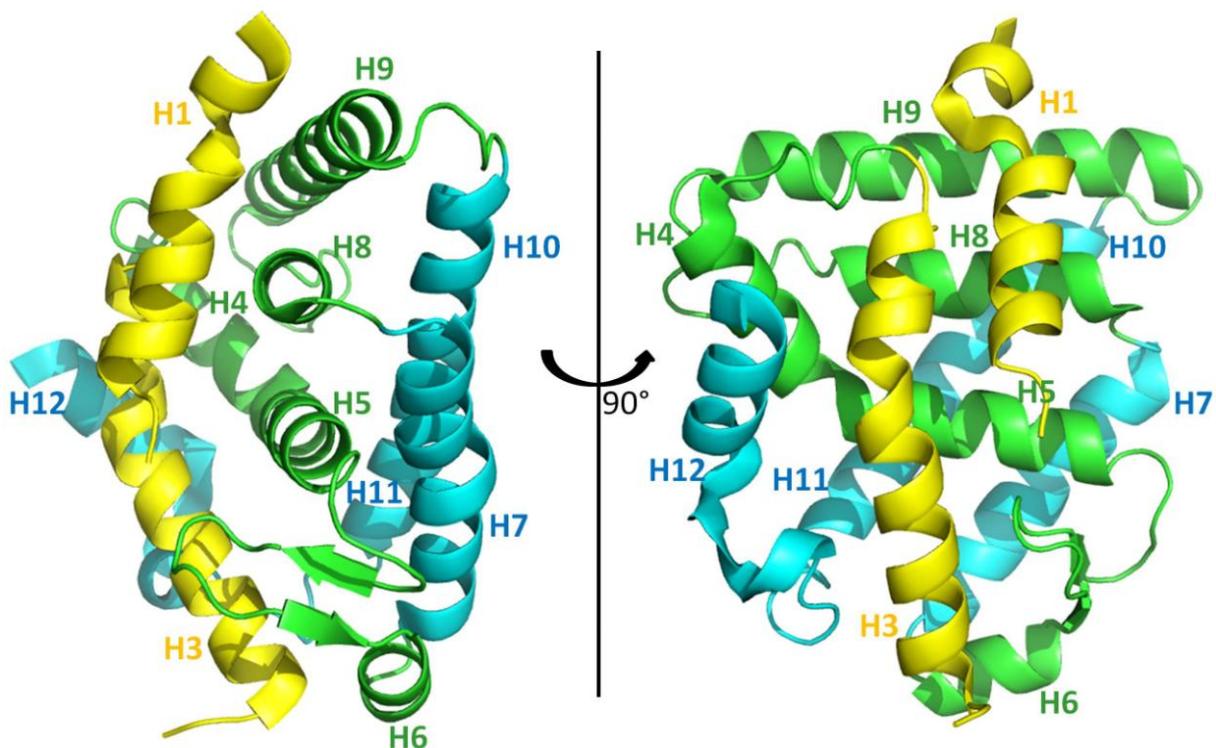
Figure 15 : Représentation schématique du DBD de RXR (adaptée de Giguère, 1999)

### 1.4.3 Le domaine D

Ce domaine est variable en longueur et en séquence. C'est un domaine charnière entre le DBD et le LBD (ligand binding domain). Initialement, ce domaine était considéré comme une zone flexible permettant de moduler la position du DBD par rapport au LBD et ainsi permettre les rotations nécessaires pour la fixation à l'ADN. Dans ce contexte, cette flexibilité permet la fixation des récepteurs nucléaires à la fois sur des éléments de réponses directs (DR direct repeat) et inversés (IR inverted repeat) (Glass, 1994). Une fois le récepteur fixé, ce domaine adopte une structure secondaire. Plus récemment, il a été montré que ce domaine a également une activité variée en intégrant de nombreux signaux de régulation. Par exemple pour le récepteur nucléaire aux androgènes (AR), le domaine D peut être acétylé, et influence ainsi l'activité transcriptionnelle du récepteur (Fu et al., 2000).

## 1.4.4 Le domaine E

Le domaine E est le second domaine le plus conservé chez les récepteurs nucléaires. Il est plus communément appelé domaine de liaison du ligand (LBD : Ligand Binding Domain). Ce domaine est impliqué dans de nombreuses fonctions. Il contient la fonction d'activation de la transcription ligand-dépendante AF-2 (Activation-Fonction 2) sur l'hélice H12 (Gronemeyer and Laudet, 1995; Renaud et al., 1995). Il assure aussi la fonction de liaison du ligand dans une poche (LBP, Ligand Binding Pocket). Ce domaine est également le siège de la dimérisation d'homo- ou d'hétérodimères et interagit avec de nombreuses protéines co-activatrices ou co-inhibitrices ainsi que des protéines de choc thermique (HSP, Heat Shock Protein). L'organisation structurale du repliement (Fold) du LBD qui est très bien documentée, est un repliement unique parmi les structures protéiques connues actuellement. D'un récepteur à l'autre, il peut y avoir de légères variations, mais de manière générale il est constitué de 12 hélices  $\alpha$  (H1 à H12) antiparallèles disposées en trois couches (Moras and Gronemeyer, 1998; Wurtz et al., 1996) formant un sandwich d'hélices  $\alpha$  antiparallèles et d'un petit feuillet  $\beta$  de deux brins antiparallèles, situé entre les hélices H5 et H6. Les hélices H4, H5, H6, H8 et H9 sont situées entre les hélices H1 à H3 d'un côté et les hélices H7, H10, H11 et H12 de l'autre.

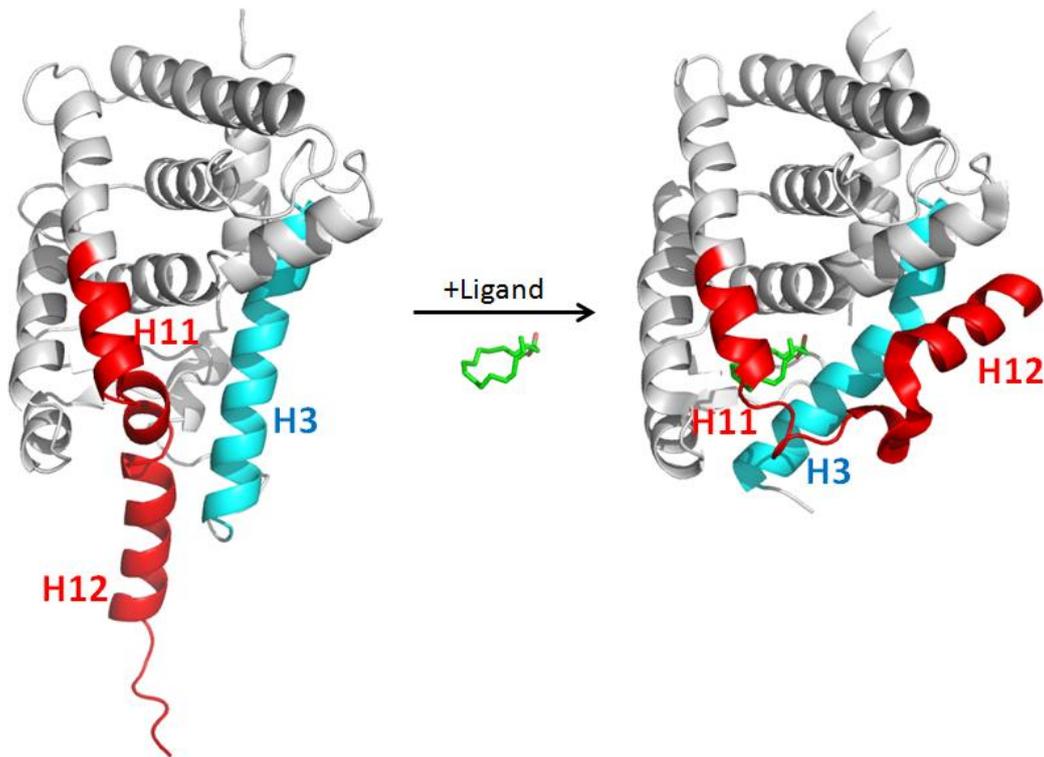


**Figure 16** : Sandwich d'hélices  $\alpha$  antiparallèles du domaine LBD. Structure du domaine LBD de RXR (Code PDB : 1DKF chaîne A). Les hélices H1 et H3 sont représentées en jaune; les hélices H4, H5, H6, H8 et H9 sont représentées en vert; les hélices H7, H10, H11 et H12 sont représentées en bleu.

De plus, une analyse des séquences protéiques des LBDs des récepteurs nucléaires a mis en évidence un motif signature d'une vingtaine d'acides aminés situés entre la partie C-ter de l'hélice H3 et l'hélice H5, [(F,W)AKX<sub>4</sub>FX<sub>2</sub>LX<sub>3</sub>DQX<sub>2</sub>LL] (Wurtz et al., 1998). Ce motif contient de nombreux résidus conservés nécessaires à la stabilisation du cœur du repliement du LBD.

L'interface de dimérisation du LBD est principalement le long de l'axe des hélices H10 et H11 et implique également des contacts entre l'hélice H7 et la boucle entre les hélices H8 et H9, ainsi que des contacts entre l'hélice H9 et la partie N-ter de H10 (Bourguet et al., 1995, 2000; Chandra et al., 2013; Egea et al., 2002; Gampe et al., 2000; Svensson et al., 2003). Il n'y a pas de variation majeure de topologie entre les homodimères et les hétérodimères, cependant chez les hétérodimères on constate une asymétrie des liaisons de dimérisation. Cette association asymétrique pour un hétérodimère impliquant RXR est plus stable que celle de l'homodimère RXR (Bourguet et al., 2000).

La poche de fixation du ligand est une zone majoritairement hydrophobe au cœur du LBD. L'entrée et la sortie du ligand est possible grâce à un mécanisme appelé "piège à souris" (Moras and Gronemeyer, 1998; Renaud et al., 1995). Il n'y a pas beaucoup de structures disponibles de LBD qui n'ont pas de ligand. En comparant les structures cristallographiques du LBD sous forme apo (absence du ligand) (Bourguet et al., 1995) et sous forme holo (présence de ligand) (Bourguet et al., 2000; Egea et al., 2000; Renaud et al., 1995), on observe un réarrangement de l'hélice H12 qui en fonction de sa position, ferme ou ouvre la poche (Blondel et al., 1999).



**Figure 17** : La fixation du ligand et le mécanisme de piège à souris. A droite, RXR forme apo (absence du ligand) (Code PDB : 1LBD). A gauche, RXR forme holo (présence du ligand) (code PDB : 1DKF). Lorsque le ligand se fixe, l'hélice H12 bascule pour fermer la poche du ligand et l'hélice H3 bascule.

Ce faisant le récepteur nucléaire devient transcriptionnellement actif après la fixation du ligand. L'orientation de l'hélice H12 est une conséquence allostérique induite par la nature du ligand qui est fixé dans la poche de liaison du ligand. Cette modification conformationnelle génère de nouvelles surfaces d'interaction au niveau de H5 et H12 qui participent au recrutement de cofacteurs.

Un grand nombre de récepteurs nucléaires sont dit orphelins, c'est à dire qu'ils n'ont pas de ligand naturels connus. Les récepteurs nucléaires non orphelins ont une forte sélectivité et affinité pour leur ligand respectif. Cette sélectivité est possible notamment grâce aux chaînes latérales des résidus qui forment la poche du ligand, la variabilité de la nature hydrophobe ou hydrophile de la surface interne de la poche et la taille de la poche. En effet par exemple pour le récepteur activé par les proliférateurs de peroxyosomes (peroxisome proliferator-activated receptor, PPAR) qui est orphelin, il a une poche de très grande taille, de l'ordre de  $1400 \text{ \AA}^3$  alors que DHR38 a une très petite poche de l'ordre de  $30 \text{ \AA}^3$  (Li et al., 2003). A titre de comparaison le ligand  $17\beta$ -œstradiol a un volume moléculaire de  $214.817 \text{ \AA}^3$ . La grande taille de la poche permet de fixer de nombreux ligands sans réelle sélectivité. C'est pourquoi les récepteurs non orphelins ont une petite poche du ligand. La "rigidité" du ligand joue aussi un rôle, en effet les ligands stéroïdiens sont moins flexibles que les

acides rétinoïques par exemple. Une taille réduite de la poche permet une plus grande sélectivité en général. On peut cependant nuancer avec la poche de VDR pour qui un ligand plus grand que la poche peut également rentrer en déplaçant des chaînes latérales et créant ainsi une poche secondaire (Mizwicki et al., 2004).

### 1.4.5 Le domaine F

Le domaine F est le domaine C-terminal des récepteurs nucléaires, mais il n'est pas présent chez tous les récepteurs nucléaires. On le retrouve chez les oxo-stéroïdiens, chez ER et chez HNF-4. C'est un domaine dont la fonction est encore mal connue. D'un point de vue structural, on ne l'a que pour les structures des oxostéroïdiens, par exemple pour la structure du récepteur aux glucocorticoïdes (Bledsoe et al., 2002). La délétion du domaine F chez ER $\alpha$  augmente son affinité pour l'œstradiol, par contre la délétion du domaine F pour PR entraîne une diminution de la fixation de la progestérone. Chez HNF-4 $\alpha$  une étude démontre que la perte du domaine F diminue la capacité de fixer CoR SMRT (Silencing Mediator for Retinoic acid and Thyroid hormone receptors) (Chen and Evans, 1995). Le domaine F a donc un rôle de régulateur pour HNF-4 $\alpha$  qui intervient dans la discrimination entre les CoA et CoR. De plus pour ER $\alpha$  le domaine F participe à l'homodimérisation des LBD (Skafar and Zhao, 2008). Les interactions entre le domaine F de ER et les protéines 14-3-3 a un effet négatif sur la dimérisation de ER, les protéines 14-3-3 étant une famille de protéines spécialisées pour lier et réguler divers types de protéines de signalisations (Leeuwen et al., 2013).

## 1.5 Les éléments de réponses

Une grande partie des Récepteurs nucléaires reconnaissent des éléments de réponse, ou demi-sites, sur l'ADN qui est constitué d'une séquence hexanucléotidique canonique 5'-PuGGTCA-3' où Pu est une purine (adénine ou guanine) (Gronemeyer and Laudet, 1995). Il existe quelques variantes à cette séquence comme par exemple les oxostéroïdiens qui fixent la séquence AGAACA. On observe une corrélation entre le mode de dimérisation des Récepteurs nucléaires et leur mode de liaison à l'ADN grâce à une multitude de géométries différentes existantes.

---

### 1.5.1 Les récepteurs ligand-dépendants

Pour les récepteurs à activité ligand-dépendante, la majorité des sites de liaison à l'ADN sont composés de 2 demi-sites de 6 paires de bases orientés soit en tête à queue (DR : Direct Repeat) soit en tête bêche (IR : Inverted Repeat). Entre ces deux demi-sites, il peut y avoir un espacement plus ou moins grand avec des bases aléatoires, cet espacement étant compris entre 0 et 5 nucléotides. Ainsi on peut décrire un élément de réponses de l'ADN avec les critères suivants : la séquence des demi-sites, leur orientation l'un par rapport à l'autre, la distance les séparant et la séquence du séparateur.

Les répétitions directes sont les éléments de réponses pour les hétérodimères impliquant RXR avec ses partenaires de classe II mais aussi des récepteurs homodimériques non stéroïdiens tel que HNF-4. Ces séquences présentent une polarité de liaison. Par exemple pour les hétérodimères, RXR peut occuper l'un ou l'autre des demi-sites en fonction du cas. Pour les demi-sites DR2 et DR5, dans le cas de l'hétérodimère RXR-RAR, RXR occupe le demi site en 5' tandis que cette polarité est inversé pour une séquence DR1 (Rastinejad et al., 2000). RXR occupe également le demi-site en 5' dans d'autres cas. Par exemple lorsqu'il est associé à VDR sur un ADN DR3 (Orlov et al., 2012), mais aussi sur un ADN DR4 lorsqu'il hétérodimérise avec TR et LXR (Lou et al., 2014; Rastinejad et al., 2000).

Les récepteurs nucléaires stéroïdiens et homodimères quant à eux se fixent sur des éléments de réponses inversés qui sont des palindromes. Pour les oxostéroïdiens et le récepteur aux œstrogènes il s'agit de séquences IR3, c'est à dire des demi-sites en tête bêche séparés par 3 nucléotides (Helsen et al., 2012; Hudson et al., 2014; Roemer et al., 2006; Schwabe et al., 1993).

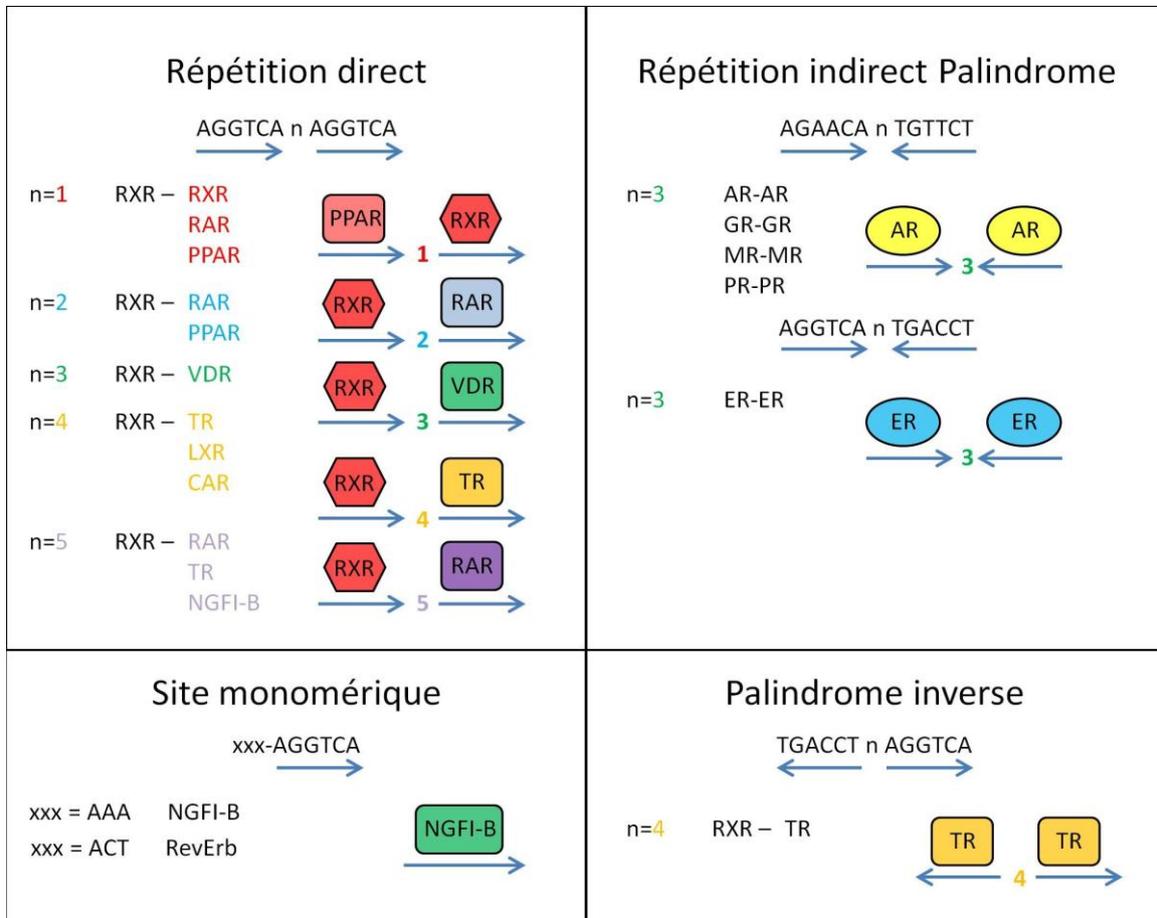
La complexité et la diversité des séquences DR associées à de nombreuses combinaisons d'homo- et d'hétérodimères sont à l'origine de la régulation fine de la transcription des gènes régulés par ces récepteurs.

---

### 1.5.2 Les récepteurs orphelins

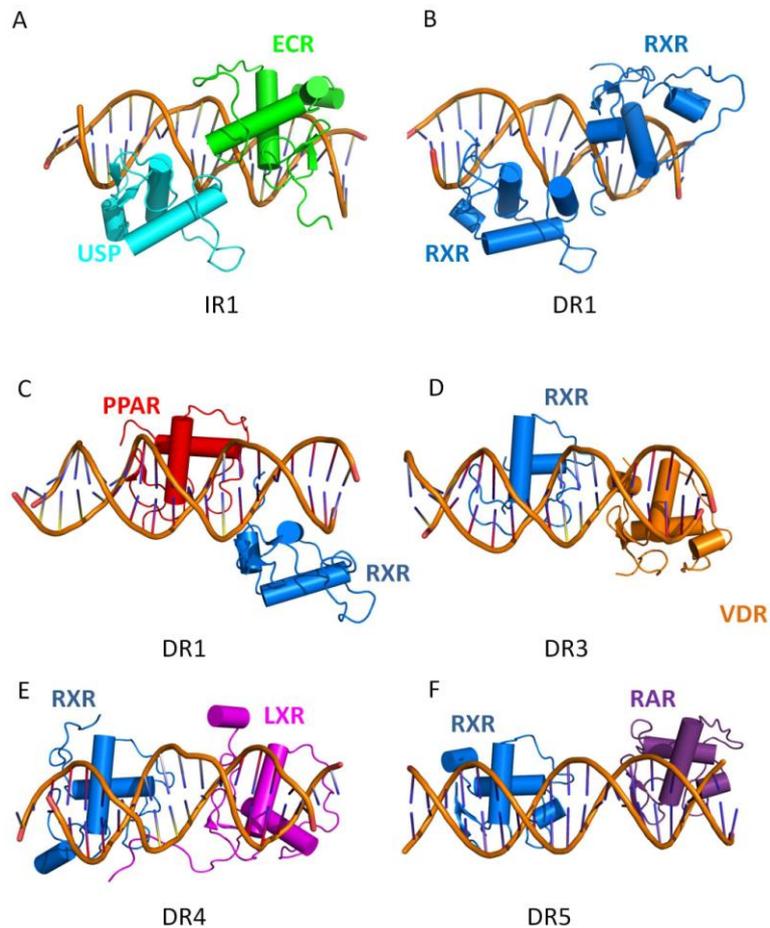
Une grande partie des récepteurs nucléaires orphelins se fixent à l'ADN grâce à un demi-site unique qui possède une extension en 5'. C'est le cas par exemple de NGFI-B qui se lie à un demi-site précédé de 3 nucléotides A-T en 5' (Meinke and Sigler, 1999). Dans le cas d'ERR (Estrogen Related Receptor), l'élément de réponse est également un demi-site seul avec une extension de 3 nucléotides en 5' de séquence TNA, où N peut être C, G, T, A. Cet élément de réponse est appelé demi site étendu ERRE

(Vanacker et al., 1999). Parmi les gènes régulés par ERR, on retrouve le gène *tff1* qui est exprimé dans la muqueuse gastro-intestinale. Ce gène est régulé par deux Récepteurs nucléaires à la fois, ER et ERR, grâce à un site d'élément de réponse combinant un ERRE et un ERE ne se superposant pas (Deblois et al., 2009).



**Figure 18** : Schéma des principaux types d'éléments de réponses des récepteurs nucléaires.

De manière général, le type d'éléments de réponses ont un impact important sur le positionnement et le type d'interaction des DBD. Un même récepteur nucléaire peut avoir des topologies différentes et donc probablement être régulé différemment si il se fixe sur des éléments de réponses différents.



**Figure 19** : Structures de différents types d'éléments de réponses. A : Structure IR1 d'un hétérodimère USP-EcR (PDB : 2HAN (Jakób et al., 2007)). B : Structure DR1 d'un homodimère RXR (PDB : 4CN2 (Osz et al., 2015)). C : Structure DR1 d'un hétérodimère RXR-PPAR (PDB : 3DZU (Chandra et al., 2008)). D : Structure DR3 d'un hétérodimère RXR-VDR (structure cryo-ME (Orlov et al., 2012)). E : Structure DR4 d'un hétérodimère RXR-LXR (PDB : 4NQA (Lou et al., 2014)). F : Structure DR5 d'un hétérodimère RXR-RAR (structure SAXS/SANS/FRET (Rochel et al., 2011)). (Adaptée de Beinsteiner and Moras, 2015).

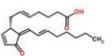
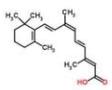
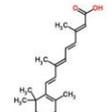
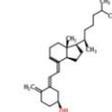
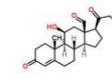
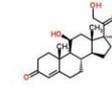
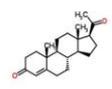
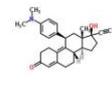
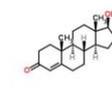
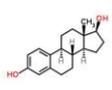
## 1.6 Les ligands

On trouve 4 grandes familles de ligands distincts pour les récepteurs nucléaires ligand-dépendants : les hormones stéroïdiennes (ER), les rétinoïdes qui sont des dérivés de la vitamine A (RAR), les hormones thyroïdiennes (TR) et les stérols comme la vitamine D (VDR).

Les hormones stéroïdiennes sont divisées en deux classes: d'une part les corticostéroïdes, qui sont synthétisés dans le cortex surrénal comprenant les glucocorticoïdes et les minéralocorticoïdes. et

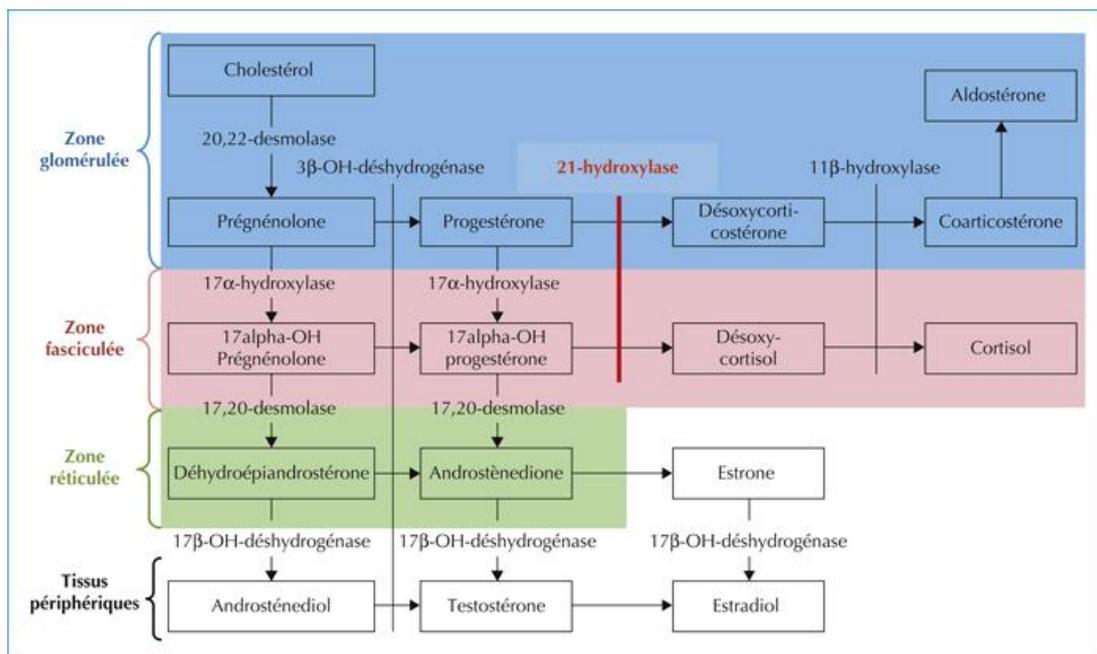
## Introduction - La superfamille des récepteurs nucléaires

d'autre part les stéroïdes sexuels, qui sont synthétisés dans les gonades et le placenta comprenant les androgènes, les progestérones et les œstrogènes.

	Ligands	Récepteurs	Familles
Thyroïde		TR	NR1
		3,5,3'Triiodothyronine (T3)	
Isoprénoides et Eicosanoïdes		PPAR	NR1
		Prostaglandine J2 (PTG)	
		RAR RXR	NR1 NR2
		Acide Rétinoïque 9-cis (9CR)	
		RAR	NR1
		Acide Rétinoïque tout-trans (REA)	
Stéroïdes		VDR	NR1
		Vitamine D3 (VD3)	
		MR	NR3
		Aldosterone (AS4)	
		GR	NR3
		Cortisol (HCY)	
		PR	NR3
		Progesterone (STR)	
		PR	NR3
		Mifepristone (486)	
	AR	NR3	
	Testostérone (TES)		
	ER	NR3	
	Oestradiol-17β (EST)		

**Figure 20** : Ligands et récepteurs nucléaires associés.

Les hormones œstrogènes ciblent plusieurs tissus dont l'endomètre, les seins, les os, le cerveau et le cœur et jouent un rôle important pour la fertilité, la grossesse, la détermination sexuelle et le système cardiovasculaire. Il s'agit d'un dérivé du cholestérol. Le cholestérol est dans un premier temps transformé en prégnénolone au niveau de la zone glomérulée des glandes surrénales. Dans un second temps, la prégnénolone est modifiée en 17-OH-prégnénolone puis en déhydroépiandrostérone (DHEA) grâce au cytochrome P450 c17 (CYP17) au niveau de la zone fasciculée des glandes surrénales. Par la suite la DHEA est transformé par l'intermédiaire de la 3 $\beta$ -hydroxystéroïde déshydrogénase (3 $\beta$ HSD) en androstènedione au niveau de la zone réticulée des glandes surrénales. Les dernières étapes de la voie de biosynthèse des œstrogènes se fait dans les tissus périphériques. Deux protéines entre en jeu et n'ont pas d'ordre d'action définis, il s'agit de l'aromatase (CYP19) et de la 17 $\beta$ -Hydroxystéroïde déshydrogénase (17 $\beta$ HSD). S'il y a d'abord action de la CYP19 puis de la 17 $\beta$ HSD, l'androstènedione est transformé en œstrone puis en œstradiol. Si à l'inverse, la 17 $\beta$ HSD agit en première, l'androstènedione est transformé en testostérone puis en œstradiol.



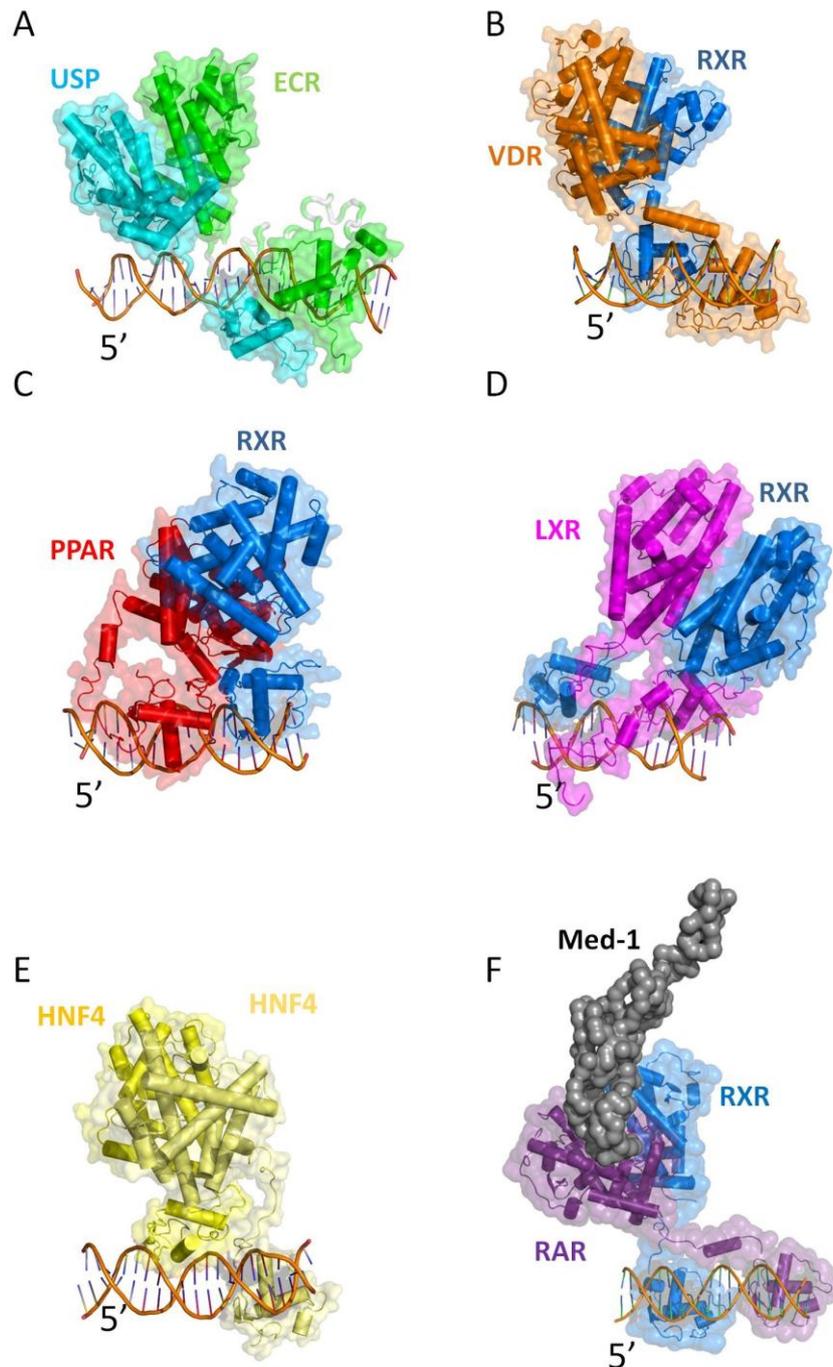
**Figure 21 :** Biosynthèse des stéroïdes sexuels. (Reproduit de Robin et al., 2012).

## 1.7 Les complexes récepteurs nucléaires-ADN

Les structures actuellement disponibles montrent une architecture moléculaire des complexes et illustrent l'adaptabilité des récepteurs tout en révélant certaines caractéristiques communes telles que leur conformation ouverte en forme de L avec une position asymétrique des domaines de liaison du ligand. Les données structurales informent sur le rôle important du promoteur et de la protéine qui s'articule dans l'organisation spatiale des domaines de liaison de l'ADN et du ligand. Elles donnent des informations importantes sur la manière dont l'ADN dicte la topologie du complexe et son asymétrie, remodelant ainsi les surfaces d'interaction en fonction des différents éléments de réponse.

L'étude du rôle de la phosphorylation dans le processus d'activation de RAR a fourni une première illustration de ce concept (Gaillard et al., 2006). Leurs partenaires (RXR ou USP) appartiennent à la classe I et forment des homodimères stables en l'absence de récepteurs nucléaires de classe II (Billas et al., 2001; Bourguet et al., 1995). Les récepteurs nucléaires interagissent avec les corépresseurs, les coactivateurs et d'autres cofacteurs protéiques qui participent à la transduction du signal de la machinerie transcriptionnelle (Bulyanko and O'Malley, 2011). Plus de 300 cofacteurs primaires ou secondaires ont été identifiés ([www.nursa.org](http://www.nursa.org)), mais leurs rôles n'ont pas été complètement élucidés. Un modèle général propose que les récepteurs nucléaires non stéroïdiens forment des hétérodimères avec les RXR. En l'absence de ligand, les hétérodimères sont associés à des complexes corépresseurs avec une activité histone-désacétylase qui modifient la chromatine pour établir et maintenir un état transcriptionnel réprimé (Glass and Rosenfeld, 2000; Nagy et al., 1999).

En ce qui concerne l'étude structurale de récepteurs nucléaires comprenant l'essentiel du complexe, c'est-à-dire DBD et LBD, trois structures cristallines d'hétérodimères sont actuellement connues (PPAR $\gamma$ /RXR $\alpha$ /ADN ; LXR $\beta$ /RXR $\alpha$ /ADN et RXR $\alpha$ /RAR $\beta$ /ADN) (Chandra et al., 2008, 2017; Lou et al., 2014). Deux structures en solution de cryo-microscopie électronique (cryo-ME) sont également disponibles pour VDR/RXR $\alpha$ /ADN et USP/EcR/ADN (Maletta et al., 2014; Orlov et al., 2012). Une dernière structure en solution (RAR $\alpha$ /RXR $\alpha$ /ADN/Med1) à basse résolution déterminé par SAXS/SANS/FRET (Rochel et al., 2011). Une structure cristalline d'homodimère (HNF-4 $\alpha$ /HNF-4 $\alpha$ /ADN) lié à son ADN cible est également disponible (Chandra et al., 2013). Une dernière structure à basse résolution (ER $\alpha$ /SRC-3b/CARM1/p300/ADN), obtenue par cryo-microscopie électronique d'un complexe bien plus gros est également disponible (Yi et al., 2015, 2017). Les résultats illustrent la diversité des récepteurs nucléaires tout en révélant des caractéristiques communes à l'architecture moléculaire des complexes. Certaines corrélations fonctionnelles apparaissent.



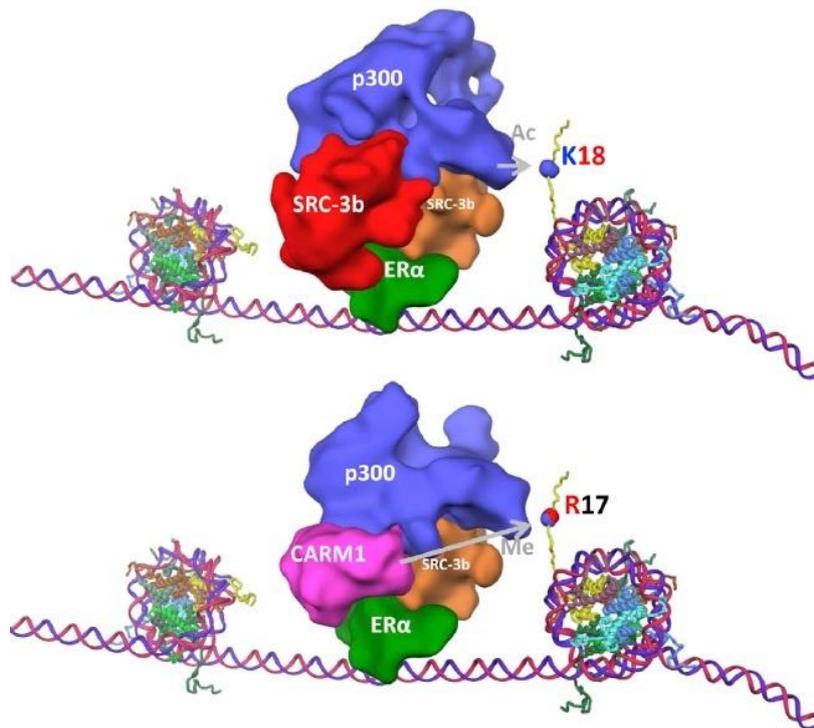
**Figure 22** : Complexes de récepteurs nucléaires avec leur ADN. A : Structure d'un hétérodimère USP/Ecr/ADN (PDB : 4UMM structure cryo-ME (Maletta et al., 2014)). B : Structure d'un hétérodimère VDR/RXR $\alpha$ /ADN (structure cryo-ME (Orlov et al., 2012)). C : Structure d'un hétérodimère PPAR $\gamma$ /RXR $\alpha$ /ADN (PDB : 3DZU (Chandra et al., 2008)). D : Structure d'un hétérodimère LXR $\beta$ /RXR $\alpha$ /ADN (PDB : 4NQA (Lou et al., 2014)). E : Structure d'un homodimère HNF-4 $\alpha$ /HNF-4 $\alpha$ /ADN (PDB : 4IQR (Chandra et al., 2013)). F : Structure d'un hétérodimère RAR $\alpha$ /RXR $\alpha$ /ADN/Med1 (structure SAXS/SANS/FRET (Rochel et al., 2011)). (Adaptée de Beinsteiner and Moras, 2015).

Les structures de solution ont été déterminées à l'aide d'approches intégratives combinant des méthodes de diffusion aux petits angles, de diffraction de rayons X (SAXS) et de neutrons (SANS), de techniques optiques comme le FRET avec des molécules marquées et de microscopie électronique (cryo-ME). A savoir que la méthode cryo-ME utilisée pour la détermination de la structure analyse une solution gelée qui préserve l'échantillon dans un état fonctionnel. Le SAXS et le SANS sont des méthodes structurales puissantes pour étudier les protéines multi-domaines flexibles en solution en évitant les artefacts de compression des cristaux (Petoukhov et al., 2013). Ils fournissent des enveloppes moléculaires pouvant être interprétées au niveau moléculaire lorsque des structures à haute résolution des domaines individuels sont disponibles, ce qui est le cas pour de nombreux récepteurs nucléaires, par contre il n'y a pas de structures à haute résolution pour les différents cofacteurs. Ces méthodes permettent donc de répondre sans ambiguïté à la question de la topologie correcte globale d'un complexe en solution. En outre, ils fournissent des informations supplémentaires importantes telles que la présence ou non d'un conformère en solution unique ou largement dominant et sa corrélation avec les modèles moléculaires proposés. La cristallographie fournit des informations à haute résolution sur la partie ordonnée des molécules mais repose sur l'existence de bons cristaux diffractant. La structure cristalline capture la conformation la plus favorable à la croissance cristalline et la structure cristalline fournit ainsi un instantané du conformère sélectionné. Ce dernier ne peut représenter qu'une petite fraction des conformères de la solution car le processus de cristallisation peut piéger et stabiliser une conformation non dominante en solution. En fait, des cas ont été rapportés dans la littérature où un contaminant, présent à moins de 5% de la préparation protéique, était cristallisé (Vesely et al., 2009).

Ces structures cristallines et la plupart des autres, telles que celles des récepteurs complets, ont été obtenues avec des séquences d'ADN consensus, extrêmement rares dans des séquences fonctionnelles réelles de promoteurs cibles. La première preuve structurelle de l'importance de la séquence réelle a été fournie par la structure cristalline des DBD USP/EcR liés aux demi-sites avec une répétition inversée espacée de 1 paire de bases (IR1), un élément de réponse ADN palindromique naturel. La comparaison de la structure avec celle obtenue en utilisant un élément de réponse consensus a montré comment le récepteur nucléaire EcR s'adapte aux modifications structurelles induites par l'ADN. Une partie de l'extension C-terminale (CTE) du DBD de EcR se replie en une hélice  $\alpha$  dont l'emplacement dans le petit sillon ne correspond à aucun des emplacements précédemment observés (Jakób et al., 2007).

Dans un contexte cellulaire, les récepteurs nucléaires recrutent plusieurs coactivateurs séquentiellement pour activer la transcription. Ce recrutement « ordonné » permet d'effectuer les

différentes étapes nécessaires à la transcription. Le récepteur aux œstrogènes (ER) recrute le coactivateur primaire SRC-3 (Steroid Receptor Coactivator-3; coactivateurs des récepteurs stéroïdiens-3) et les coactivateurs secondaires, p300 / CBP et CARM1 (Histone Acetyltransferase p300 / CREB-Binding Protein et Coactivator-Associated Arginine Methyltransferase 1). Le recrutement de CARM1 est en retard par rapport à la liaison de SRC-3 et de p300 avec ER. Il existe un lien étroit entre les coactivateurs recrutés précocement et tardivement. Le recrutement séquentiel de CARM1 ajoute non seulement une activité arginine méthyltransférase au complexe, mais modifie également l'organisation structurale du complexe ADN/ER $\alpha$ /SRC-3/p300 préexistant. Il induit un changement de conformation de la protéine p300 et augmente significativement l'activité histone acetyltransferase (HAT) de p300 sur les résidus de l'histone H3-K18, ce qui favorise l'activité de la méthylation de CARM1 sur les résidus H3-R17 pour améliorer l'activité transcriptionnelle (Yi et al., 2017).



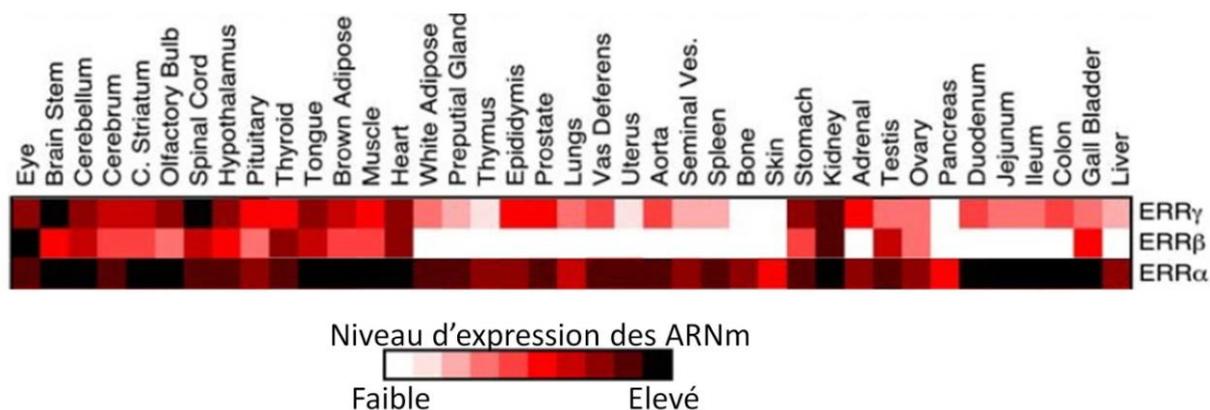
**Figure 23** : Modèle de la structure ADN/ER $\alpha$ /SRC-3/CARM1/p300 sur un segment de chromatine (Reproduit de Yi et al., 2017).

## 2. Le récepteur nucléaire orphelin ERR apparenté à ER

### 2.1 ERR et ses 3 isotypes $\alpha$ , $\beta$ , $\gamma$

ERR (Estrogen-Related Receptor) est un récepteur nucléaire orphelin, pour lequel aucun ligand naturel n'a été identifié à ce jour. Cependant, il joue un rôle de régulation centrale dans plusieurs voies métaboliques. Chez l'homme, il y a 3 isotopes d'ERR,  $\alpha$ ,  $\beta$  et  $\gamma$ . Chacun d'entre eux jouant des rôles différents. ERR $\alpha$  est le premier membre de la famille à avoir été identifié en 1988 sur la base de sa forte similarité de séquence protéique de son DBD avec celui de ER $\alpha$ , il a été isolé à partir du rein murin. Dans un second temps, également en 1988, c'est au tour d'ERR $\beta$  d'être identifié à partir de l'ADNc d'ERR $\alpha$ , dans le cœur murin (Giguère et al., 1988). ERR $\gamma$  quant à lui n'a été identifié qu'en 2000 grâce à sa forte homologie avec ERR $\beta$  (Heard et al., 2000). D'un point de vue phylogénétique, ERR $\beta$  et ERR $\gamma$  sont plus proches entre eux que de ERR $\alpha$  (May, 2014).

Ces 3 isotypes sont exprimés différemment entre les différents tissus de l'organisme. ERR $\alpha$  est le plus largement exprimé des trois isotypes et est retrouvé dans divers tissus. Il intervient dans le métabolisme lipidique et l'homéostasie énergétique. On le retrouve en concentration élevée dans les organes du système nerveux central tels que les yeux, la corde spinale et l'hypothalamus, mais aussi dans les organes du système gastro-intestinal comme l'estomac, le colon et le duodénum, ainsi que dans les organes du système cardiovasculaire comme l'aorte, le cœur et les poumons. Les tissus à haute activité métabolique comme les reins, les tissus adipeux ou les muscles présentent aussi une forte expression d'ERR $\alpha$ . Il est également exprimé plus discrètement dans les tissus à activité endocrinienne, le système immunitaire, les organes reproducteurs, les os et la peau. ERR $\beta$  quant à lui est le moins abondant des trois. On le retrouve principalement dans le système nerveux central, mais on peut aussi le retrouver plus modérément dans les tissus endocrinien, gastro-intestinal et cardiovasculaire mais son rôle majeur est au niveau des cellules souches et de l'embryogénèse. Le dernier isotype ERR $\gamma$  est quant à lui modérément exprimé dans les systèmes nerveux central, gastro-intestinal et cardiovasculaire et est faiblement exprimé dans les tissus endocriniens, immunitaires et reproducteurs (Bookout et al., 2006; Giguère et al., 1988). Les expressions respectives des trois isotypes suivent un rythme circadien. Par exemple dans les tissus adipeux bruns et le foie on constate un pic d'ERR $\gamma$  pendant la période diurne tandis que pendant cette même période ERR $\alpha$  est plus faiblement exprimé dans ces mêmes tissus. Il y a probablement un lien moléculaire entre l'horloge circadienne et la régulation du métabolisme énergétique (Yang et al., 2006).



**Figure 24** : Niveau d'expression des ERR en fonction des tissus (adaptée de (Bookout et al., 2006)).

## 2.2 La régulation de l'activité transcriptionnelle d'ERR

### 2.2.1 Le coactivateur PGC-1 $\alpha$

La protéine PGC-1 $\alpha$  (Peroxisome proliferator-activated receptor-gamma coactivator 1 $\alpha$ ) a été caractérisée en 1998 comme coactivateur des récepteurs nucléaires PPAR $\gamma$  et TR pour activer l'expression de la protéine mitochondriale UCP-1 (Uncoupling Protein-1). Elle est fortement exprimée dans les tissus qui ont une forte demande énergétique comme les muscles squelettique, le cœur et le cerveau. PGC-1 $\alpha$  est un régulateur de la thermogénèse qui induit la production de chaleur physiologique en fonction des variations environnementales. Lors de l'exposition de l'organisme au froid, cette protéine stimule l'expression de UCP-1 qui dissipe le gradient électrochimique mitochondrial et induit ainsi la production de chaleur (Puigserver et al., 1998). Le jeûne provoque une augmentation de PGC-1 $\alpha$  qui stimule la gluconéogenèse hépatique en induisant la transcription d'une série de protéines indispensables à cette dernière (Yoon et al., 2001). ERR $\alpha$  et PGC-1 $\alpha$  interviennent dans l'autorégulation de l'expression d'ERR $\alpha$  en se liant au promoteur du gène *esrra*. L'expression d'ERR $\alpha$  augmente alors. L'augmentation de PGC-1 $\alpha$  étant induite par l'exposition au froid, le jeûne et l'activité physique.

PGC-1 $\alpha$  favorise la transcription des gènes régulés par ERR en se liant à ce dernier avec nos connaissances actuelles pas d'interaction avec l'ADN. Pour se fixer à ERR, PGC-1 $\alpha$  a trois motifs LXXLL (où L est une Leucine et X un acide aminé quelconque) d'interaction avec les récepteurs nucléaires. Ces motifs se nomment L1, L2 et L3. Le premier d'entre eux est retrouvé dans le domaine d'activation de la transcription (TAD) de PGC-1 $\alpha$  mais n'est pas utilisé pour la liaison pour la liaison avec les

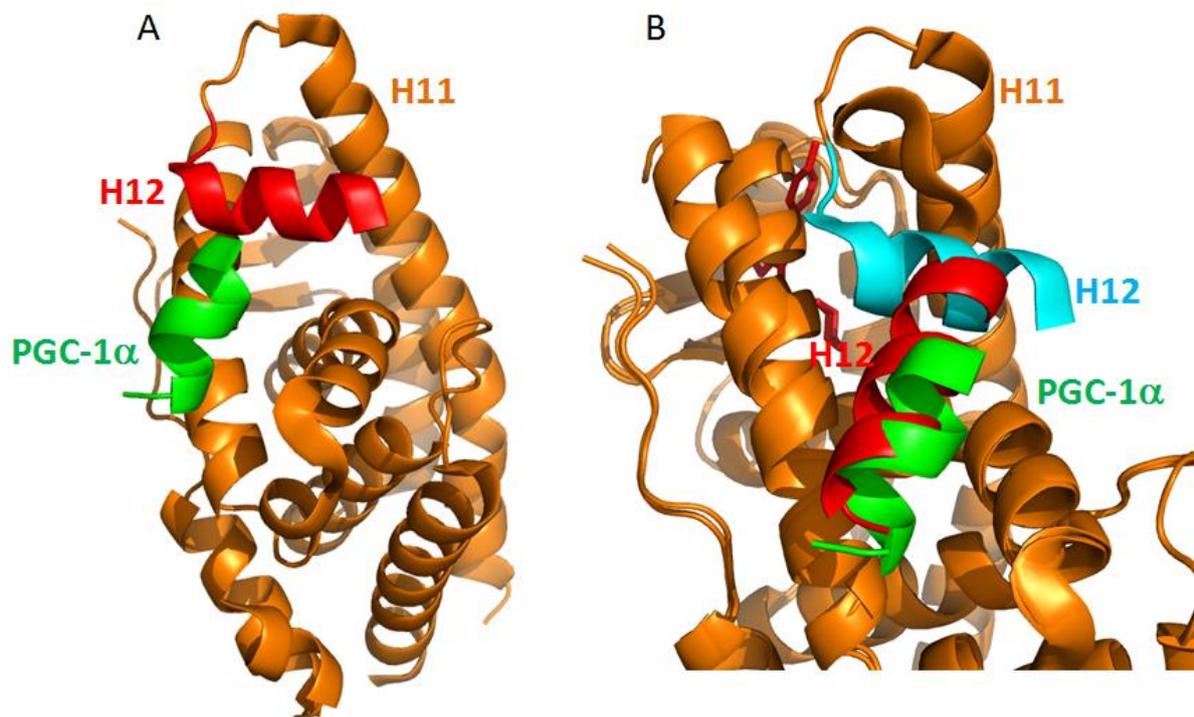
récepteurs nucléaires. Seuls L2 et L3 sont impliqués dans la liaison avec les Récepteurs nucléaires, L3 étant spécifique à ERR.



**Figure 25** : Organisation en une dimension des domaines fonctionnels de PGC-1 $\alpha$ .

L1, L2 et L3 sont les motifs LXXLL (en bleu) qui sont les sites d'interactions avec les LBD des Récepteurs nucléaires. Les autres domaines sont les suivants : TAD (Transcription Activation Domain); NLS (Nuclear Localization Signal); RS (Arginine/Serine rich domain); RRM (RNA recognition motif).

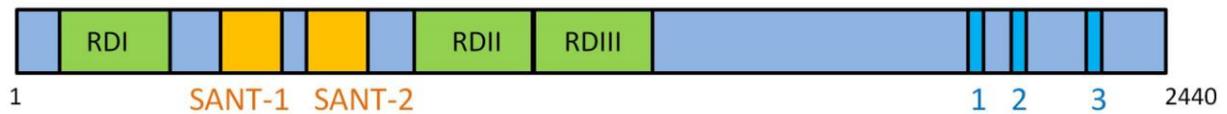
PGC-1 $\alpha$  interagit avec plusieurs récepteurs nucléaires au niveau des LBDs, pour certains comme pour ER, PPAR et PXR $\alpha$ , l'activation est ligand dépendante, tandis qu'elle est ligand indépendante pour HNF-4 $\alpha$  et ERR. Pour activer la transcription il est nécessaire de décondenser la chromatine pour permettre aux complexes transcriptionnels de se fixer à l'ADN. Pour ce faire, PGC-1 $\alpha$  recrute avec sa partie amino-terminale des protéines telles que SRC-1 (Steroid-Receptor Coactivator-1) et CREB (cAMP-Response-Element-Binding protein) qui ont une activité enzymatique d'acétyltransférase (HAT; Histone Acetyltransferase). Cette acétylation des histones est nécessaire pour le remodelage de la chromatine (Puigserver et al., 1999). En partie carboxy-terminale il y a deux domaines d'interaction avec l'ARN, les domaines RS et le domaine RRM. Ces domaines permettent l'interaction entre l'ARN polymérase II et plusieurs facteurs d'élongation comme CDK9 (Cyclin-Dependant Kinase 9) et la Cycline T (Monsalve et al., 2000). En amont de la région RS se trouvent deux séquences NLS (Nuclear Localization Signal) qui permettent à PGC-1 $\alpha$  de se localiser dans le noyau pour assurer son activité. Cette activité est régulée grâce à la p38 MAP Kinase qui phosphoryle PGC-1 $\alpha$  et renforce ainsi son activité et sa stabilité (Puigserver et al., 2001).



**Figure 26 : A :** Structure d'ERR $\alpha$  avec un peptide du coactivateur PGC-1 $\alpha$  (PDB : 1XB7 (Kallen et al., 2004)). L'hélice H12 est en position agoniste. **B :** Superposition de la structure 1XB7 avec une structure d'ERR $\alpha$  lié par un ligand synthétique antagoniste (1a) (PDB : 2PJL (Kallen et al., 2007)). On voit que l'hélice H12 en rouge de la structure avec l'antagoniste prend la place de fixation du peptide de PGC-1 $\alpha$  en vert.

### 2.2.2 Le corépresseur NCoR

La protéine NCoR (Nuclear-receptor CoRepressor) réprime l'activité transcriptionnelle des récepteurs nucléaires de façon ligand indépendante. Pour ce faire il interagit avec la partie charnière (Hinge) du récepteur nucléaire et non avec le LBD où se lient les coactivateurs (Hörlein et al., 1995). Dans sa partie carboxy-terminale, NCoR a trois motifs hydrophobes I/LXXII qui interagissent avec les Récepteurs nucléaires (Pérez-Schindler et al., 2012). NCoR n'a pas de fonction de désacétylation, cependant il permet de recruter l'histone désacétylase HDAC-3 afin de réprimer l'initiation de la transcription. Le recrutement de HDAC-3 se fait grâce aux régions SANT-1 et SANT-2 de NCoR. La désacétylation de la queue amino-terminale des histones induit la compaction de la chromatine et donc diminue fortement l'expression génique (Li et al., 2000; Zhang et al., 2002).



**Figure 27** : Organisation en une dimension des domaines fonctionnels de NCoR.

1, 2 et 3 sont les motifs I/LXXLL (en bleu) qui sont les sites d'interactions avec les Récepteurs nucléaires. Les autres domaines sont les suivants : RDI, RDII et RDIII (Repression Domain); SANT (SWI3, ADA2, N-CoR et TFIIIB).

### 2.2.3 Régulation d'ERR avec PGC-1 $\alpha$ et NCoR

Dans les cellules musculaires murines, après une délétion de la protéine NCoR on constate une augmentation du métabolisme oxydatif qui se traduit par une augmentation de l'expression des protéines de la chaîne respiratoire mitochondriale, telles que par exemple la succinate déshydrogénase ou la NADH déshydrogénase. A l'opposé, la délétion de PGC-1 $\alpha$  induit la diminution de l'expression de ces mêmes protéines. PGC-1 $\alpha$  et NCoR partagent donc la régulation des mêmes gènes du métabolisme oxydatif en ayant des effets opposés sur leur expression. NCoR inhibe l'activité transcriptionnelle d'ERR tandis que PGC-1 $\alpha$  active cette activité transcriptionnelle (Pérez-Schindler et al., 2012). Lors d'inactivité physique, NCoR inhibe donc l'activité d'ERR $\alpha$  en recrutant la protéine HDAC qui elle-même permet la désacétylation de la chromatine et bloque l'expression des gènes du métabolisme oxydatif. Lors d'une activité physique, il se passe l'inverse, PGC-1 $\alpha$  permet le recrutement des protéines HAT qui acétylent la chromatine et induit par conséquent la fixation du complexe de pré-initiation de la transcription pour les gènes du métabolisme oxydatif.

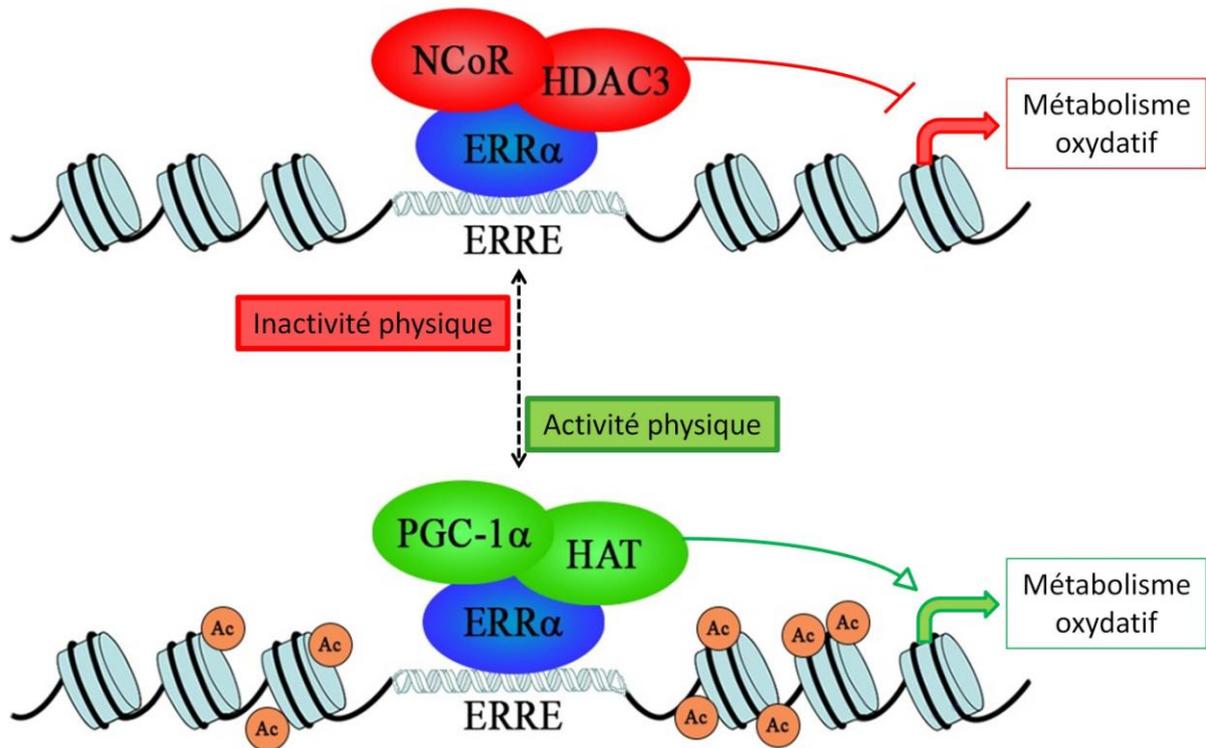


Figure 28 : Régulation du métabolisme oxydatif par PGC-1α et NCoR.

#### 2.2.4 Les modifications post-transcriptionnelles

Il n'y a pas de ligand naturel connu pour ERR, cependant les cofacteurs ne sont pas les seuls moyens de régulation de l'activité du récepteur. Son activité est également dynamiquement régulée par des modifications post-transcriptionnelles.

Parmi ces modifications, il y a la sumoylation dépendante de la phosphorylation qui induit un ralentissement de l'activité transcriptionnelle. Ces modifications se font au niveau de motifs PDSM (Phosphorylation-Dependant Sumoylation Motif) qui ont la séquence consensus  $\psi$ KXE/DXXSP (ou  $\psi$  est un acide aminé hydrophobe et X un acide aminé quelconque). Les trois isotypes d'ERR possèdent le motif PDSM au niveau de leurs domaines amino-terminaux A/B, ces motifs sont préalablement phosphorylés au niveau de certaines sérines, ce qui permet la réaction de sumoylation de lysines. La mutation des lysines en arginines ou encore la mutation des sérines en alanines qui empêche également indirectement la sumoylation provoque une augmentation de l'activité d'ERR. La régulation par sumoylation a donc un effet répresseur sur l'activité d'ERR (Tremblay et al., 2008).

L'acétylation d'ERR induit également un ralentissement de son activité transcriptionnelle. L'acétyltransférase pCAF (p300 Coactivation Associated Factor) catalyse l'acétylation de 4 lysines situées dans le domaine DBD du récepteur. Ces acétylations par pCAF se traduisent par une diminution de l'activité transcriptionnelle du complexe ERR/ PGC-1 $\alpha$ . En parallèle, des mécanismes de désacétylation sont assurés par HDAC8 et Sirt1. Ils désacétylent les mêmes lysines du DBD et provoquent ainsi une augmentation de la liaison entre ERR et son ADN cible et augmentent donc son activité transcriptionnelle. L'activité transcriptionnelle d'ERR est donc finement régulée grâce à des mécanismes d'acétylation et de désacétylation des DBD (Wilson et al., 2010).

## 2.3 ERR $\alpha$ est un régulateur du métabolisme énergétique

### 2.3.1 Métabolisme du glucose

La glycolyse permet de réguler l'assimilation du glucose et la production d'énergie. Le glucose est alors dégradé en pyruvate à la suite d'une cascade de réactions chimiques. Ce dernier est ensuite converti en acétyl-coenzyme A grâce au complexe pyruvate déshydrogénase (PDH). Le PDH est régulé post-traductionnellement par des phosphorylations réversibles par les pyruvates déshydrogénases kinases 1, 2, 3 et 4. Le gène *pdk4* de la pyruvate déshydrogénase kinase 4 est régulé positivement par le complexe ERR $\alpha$ /PGC-1 $\alpha$  (Araki and Motojima, 2006). De plus, la synthèse du glucose lors de la gluconéogenèse fait intervenir la phosphoénolpyruvate carboxykinase, pour laquelle l'expression est induite par le complexe GR/PGC-1 $\alpha$  (Yoon et al., 2001) avec lequel ERR $\alpha$  interfère. En effet, au niveau du promoteur du gène codant pour la phosphoénolpyruvate carboxykinase, il y a un élément de réponse ERRE reconnu par ERR $\alpha$  qui se comporte comme un agoniste en empêchant le recrutement de PGC-1 $\alpha$  par GR (Herzog et al., 2006).

### 2.3.2 $\beta$ -oxydation des acides gras

L'enzyme medium-chain acyl-coenzyme A Dehydrogénase (MCAD) catalyse la première étape de la  $\beta$ -oxydation des acides gras dans la mitochondrie. Elle est fortement exprimée dans tous les tissus qui ont un besoin important en énergie comme le cœur, les reins et le foie par exemple. Au niveau du promoteur du gène de cette protéine, on retrouve un élément de réponse ERRE qui est la cible

d'ERR $\alpha$  (Sladek et al., 1997). Cette étude est appuyée par la diminution de l'expression de MCAD lorsque l'on introduit des ARN interférents ciblant ERR $\alpha$  (Schreiber et al., 2003).

### 2.3.3 Fonctions mitochondriales

ERR $\alpha$  contrôle la transcription de nombreux gènes. En effet, le couple ERR $\alpha$  et PGC-1 $\alpha$  sont exprimés dans les tissus à haute demande énergétique. Ensemble ils assurent l'expression de plus de 150 gènes impliqués dans les fonctions mitochondriales. Il y a par exemple plusieurs gènes impliqués dans la chaîne de phosphorylation oxydative et le cycle des acides carboxyliques (Giguère, 2008). Les mitofusines 1 et 2 sont des protéines membranaires mitochondriales qui sont essentielles à la fusion mitochondriale et au maintien cohésif de l'organelle. ERR $\alpha$  est impliqué dans la régulation de ces protéines et est ainsi impliqué dans la biogénèse et le maintien mitochondrial (Cartoni et al., 2005).

## 2.4 La balance ERR $\alpha$ /ERR $\gamma$ dans les cancers

Dans la famille ERR, les isotypes ERR $\alpha$  et ERR $\gamma$  sont donc les deux protagonistes principaux responsables des tumeurs cancéreuses. Ils sont impliqués dans plusieurs cancers tel que le cancer du sein, des ovaires, du colon (Cavallini et al., 2005), de l'endomètre (Fujimoto and Sato, 2009) et de la prostate (Fujimura et al., 2007) où les cas les plus agressifs sont lorsqu'il y a une forte expression d'ERR $\alpha$  et une faible expression d'ERR $\gamma$ . ERR $\beta$  est plus présent au cours du développement embryonnaire et n'est à ce jour pas considéré comme un marqueur cancéreux. ERR $\alpha$  et ERR $\gamma$  sont fortement impliqués dans le contrôle cellulaire de l'énergie métabolique et leur participation est probablement importante dans le métabolisme dérégulé des cellules cancéreuses.

La surexpression d'ERR $\alpha$  et ERR $\gamma$  dans les cellules cancéreuses active fortement l'expression génique des enzymes glycolytiques via les ERRE. Les cellules cancéreuses préfèrent utiliser la glycolyse aérobie, phénomène connu sous le nom de l'effet Warburg et favorisent ainsi leur prolifération cellulaire aboutissant souvent à des phénotypes très agressifs (Cai et al., 2013). L'expression d'ERR $\alpha$  favorise la progression du cancer du sein en régulant positivement les gènes impliqués dans les fonctions oxydatives de la mitochondrie. ERR $\alpha$  représente un biomarqueur de choix dans le cas du cancer du sein puisque sa très forte expression est synonyme d'un diagnostic défavorable (Ariazi et al., 2002). ERR $\alpha$  est présent dans 55% des cancers du sein invasif et y est principalement détecté

dans les noyaux cellulaires de carcinomes. Le niveau d'ARNm d'ERR $\alpha$  est souvent supérieur à celui de ER $\alpha$  dans un grand nombre de cancers du sein, dans le cas où ER $\alpha$  n'est plus exprimé (tumeurs ER-négatives) ce niveau est encore plus élevé (Suzuki et al., 2004). Dans une cellule saine, ER $\alpha$  et ERR $\alpha$  se partagent la régulation de la transcription de gènes communs comme les gènes de la *lactoferrine*, *l'ostéopontine*, *l'aromatase* et *tff1* (Lu et al., 2001; Vanacker et al., 1998; Yang et al., 1998, 1996). Dans le cas d'une cellule cancéreuse, ERR $\alpha$  devient le régulateur majeur de ces gènes et donc aussi des nombreuses signalisations cellulaires associées, ce qui en fait une cible thérapeutique intéressante.

Pour le cancer des ovaires les expressions d'ERR $\alpha$  et ERR $\gamma$  sont augmentées et correspondant respectivement à 57,6% et 48,5% des cas. La progression tumorale est moins rapide dans les cancers ERR $\gamma$ -positifs faisant de lui un biomarqueur favorable pour le diagnostic du cancer des ovaires. L'activité d'ERR $\gamma$  peut être plus facilement inhibée par les traitements anticancéreux comme avec la molécule 4-hydroxytamoxifen. ERR $\alpha$  est quant à lui un marqueur tumoral défavorable et est responsable de la progression tumorale très rapide, dans ce cas l'espérance de vie moyenne est plus courte (Sun et al., 2006).

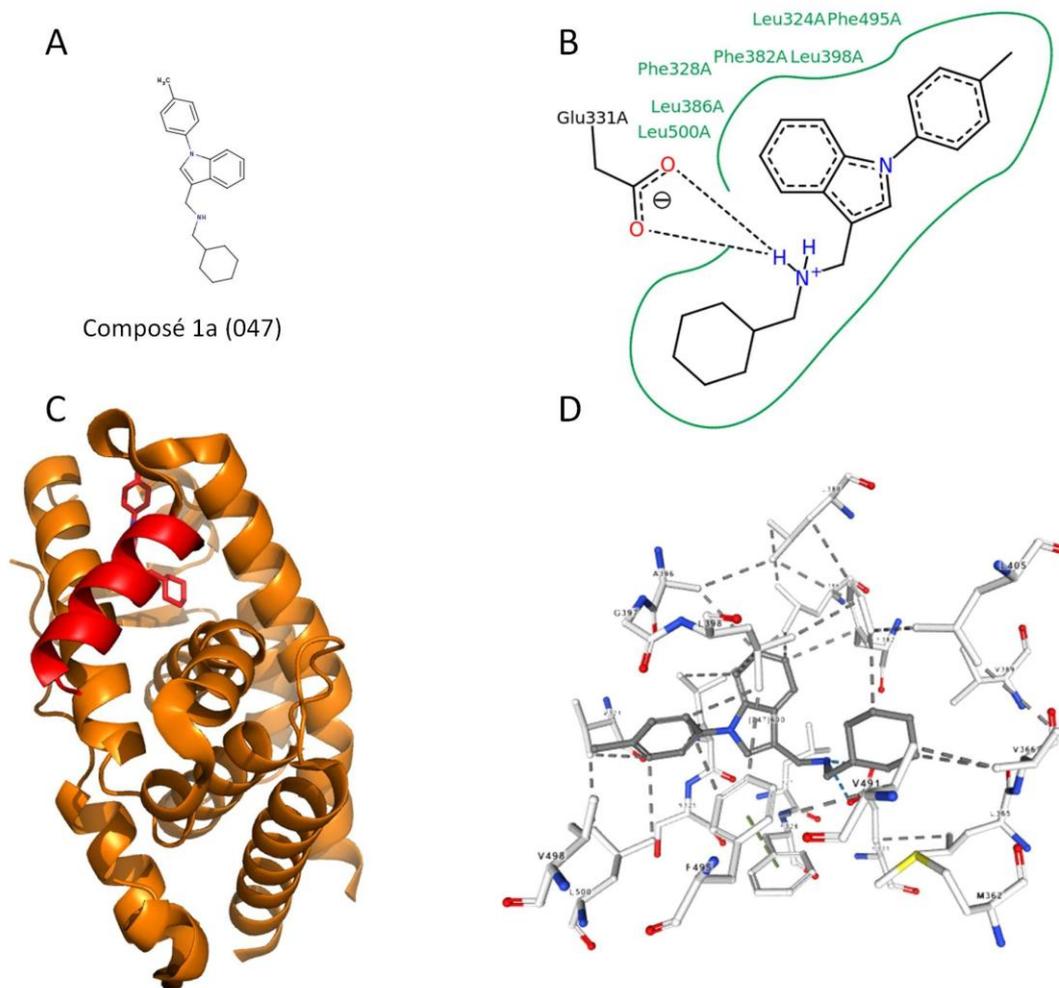
Cancers		ERR $\alpha$	ERR $\gamma$
Ovaire	Niveau d'expression	Haut	Haut
	Diagnostique	Défavorable	Favorable
Sein	Niveau d'expression	Haut	Moyen
	Diagnostique	Défavorable	Favorable
Endomètre	Niveau d'expression	Haut	/
	Diagnostique	Défavorable (invasif)	/
Prostate	Niveau d'expression	haut	bas
	Diagnostique	Défavorable	Effet neutre
Colon	Niveau d'expression	haut	bas
	Diagnostique	Défavorable	Effet neutre

**Tableau 1** : Niveau d'expression et diagnostique associé pour les cancers où les isotypes d'ERR sont impliqués.

## 2.5 Modulation d'ERR $\alpha$ par des ligands synthétiques.

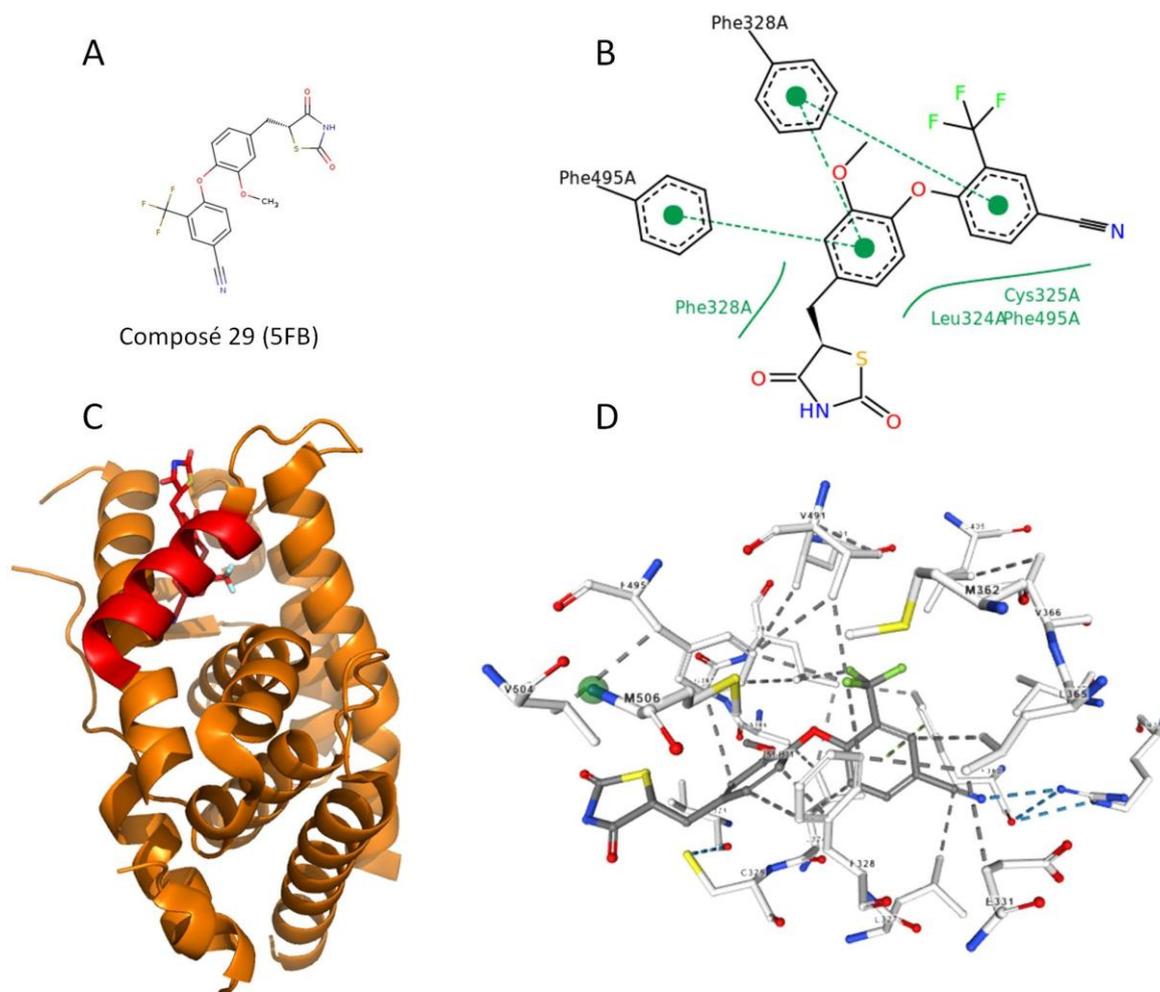
L'implication des deux isoformes ERR $\alpha$  et ERR $\gamma$  dans plusieurs cancers en font des cibles thérapeutiques de choix. Une autre particularité fait d'ERR $\alpha$  un récepteur intéressant pour la régulation par des ligands synthétiques. En effet la poche de liaison au ligand présent dans les LBD est plus étroite chez ERR $\alpha$  que chez ses homologues. La phénylalanine 328 d'ERR $\alpha$  est une alanine chez ERR $\gamma$ , ERR $\beta$  et ER $\alpha$ . Ce résidu réduit le volume de la poche de 230 Å<sup>3</sup> pour ERR $\gamma$  (Greschik et al., 2002) contre 100 Å<sup>3</sup> pour ERR $\alpha$  (Kallen et al., 2004). Cette taille réduite permet en général une plus grande sélectivité de ligand, car restreint fortement les possibilités de fixations. ERR $\alpha$  a une capacité réduite à fixer certains ligands synthétiques bien connus comme 4-OHT ou le BPA alors que ERR $\gamma$  en est capable. Plusieurs études de criblage à haut débit de banques de composés (chimiothèques) ont permis d'identifier des ligands antagonistes ciblant spécifiquement ERR $\alpha$  et qui sont incapables de se lier aux autres isoformes d'ERR. Plusieurs structures cristallographiques de LBD ERR $\alpha$  avec un ligand antagoniste sont d'ailleurs disponibles.

Par exemple le composé 1a (1-Cyclohexyl-N-{{1-(4-Methylphenyl)-1H-Indol-3-YL}Methyl}Methanamine) fixé à la poche de liaison du LBD de ERR $\alpha$  provoque le déplacement vers l'extérieur des parties N-terminales de l'hélice H3 et du C-terminal de hélice H11. La Phe328 adopte alors une nouvelle conformation qui provoque le déplacement de la Phe 510 de l'hélice H12 ce qui induit un déplacement de l'hélice H12 qui se place alors dans le sillon d'interaction des coactivateurs et ne permet plus leur liaison comme par exemple pour PGC-1 $\alpha$  (Kallen et al., 2007).



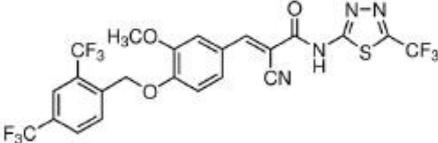
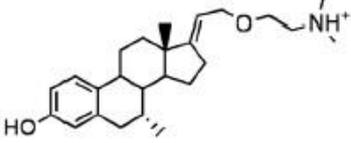
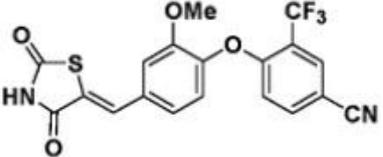
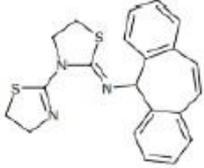
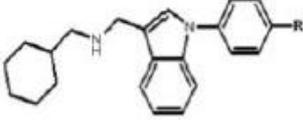
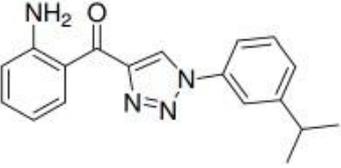
**Figure 29** : Le composé antagoniste 1a. (PDB 2PJL (Kallen et al., 2007)). A : Molécule du composé 1a. B : Représentation schématique du ligand dans la poche de liaison du ligand d'ERR $\alpha$ . Sa fixation induit le déplacement des phénylalanines de la poche dont Phe495 (H11) et Phe328 (H3) qui forment des interactions  $\pi$ - $\pi$  via leurs cycles aromatiques. C : L'hélice H12 (rouge) est placée en position antagoniste, empêchant ainsi la liaison des coactivateurs au LBD. D : Représentation en 3D du ligand et de l'ensemble des résidus de la poche qui forment des contacts, identiques à la version simplifiée en 2D de la figure B.

Le composé 29 (4-(4-(((5R)-2,4-dioxo-1,3-thiazolidin-5-yl)méthyl)-2-méthoxyphénoxy)-3-(trifluorométhyl)benzonnitrile) inhibe spécifiquement la liaison d'ERR $\alpha$  avec des protéines coactivatrices. La structure de ce composé avec les LBD d'ERR $\alpha$  permet de déterminer les bases moléculaires du mécanisme d'antagonisme par le composé 29 (Patch et al., 2011). Quand le composé 29 est lié à la poche du LBD, la boucle H11-H12 subit une distorsion ce qui permet de positionner l'hélice H12 dans une conformation antagoniste et donc empêche la liaison de coactivateurs protéiques.



**Figure 30** : Le composé antagoniste 29. (PDB 3K6P (Patch et al., 2011)). A : Molécule du composé 29. B : Représentation schématique du ligand dans la poche de liaison du ligand d'ERR $\alpha$ . Sa fixation induit le déplacement des phénylalanines de la poche comme pour le composé 1a. C : L'hélice H12 (rouge) est placée en position antagoniste, empêchant ainsi la liaison des coactivateurs au LBD. D : Représentation en 3D du ligand et de l'ensemble des résidus de la poche qui forment des contacts, identiques à la version simplifiée en 2D de la figure B.

Il y a d'autres composés connus ayant un effet antagoniste pour le récepteur nucléaire ERR $\alpha$ . Par contre il n'y a pas de structures disponibles actuellement. Il y a par exemple le composé XCT-790.

Ligands	Structures chimiques	IC <sub>50</sub>
XCT-790		0.500 μM
SR16388		0.700 μM
Composé 29 (PDB 3K6P)		0.040 μM
Composé A		0.170 μM
Composé 1a (PDB 2PJL)		0.190 μM
Composé 14n		0,021 μM

**Tableau 2** : Ligands antagonistes connus pour ERR $\alpha$ . La colonne IC<sub>50</sub> est la concentration inhibitrice médiane est une mesure de l'efficacité d'un composé donné pour inhiber une fonction biologique ou biochimique spécifique.

### 3. La microscopie électronique à transmission

#### 3.1 Les débuts et l'évolution au cours du 20<sup>ème</sup> siècle

Les débuts de la microscopie électronique à transmission (MET) ont été rendus possibles grâce à de nombreux progrès scientifiques en physique dans la deuxième moitié du 19<sup>ème</sup> siècle. Parmi ses travaux, ceux de J.J. Thomson sur les tubes cathodiques et ceux de Hans Busch sur l'optique électronique sont des contributions majeures. La découverte des propriétés des électrons est attribuée à J.J. Thomson qui a pu déterminer le rapport entre la charge et la masse des électrons, charge élémentaire d'électricité négative. Millikan, quant à lui en a mesuré la charge individuelle de  $1,6 \cdot 10^{-19}$  C en 1909.

C'est Max Knoll de l'Université Technique de Berlin et son étudiant Ernst Ruska qui sont les principaux contributeurs pour la construction des deux premiers microscopes électroniques en 1931. Ces microscopes avaient un grossissement d'environ 100 fois. Un des microscopes avait deux lentilles magnétiques, l'autre deux lentilles électrostatiques. Une des motivations de l'époque était de s'affranchir de la limite physique de la longueur d'onde liée à la microscopie photonique pour pouvoir aller plus loin en résolution. En effet, la limite de résolution est directement liée à la valeur de la longueur d'onde qui est comprise entre 400 et 700 nm pour la lumière visible contre 0.0025 nm pour un électron qui a une accélération de 200 kV. Le premier microscope électronique ayant une plus grande résolution que les microscopes photoniques est construit par ces derniers en 1933. 1939 est l'année de la première production industrielle de microscopes électroniques par Siemens. Dans les années 50 les microscopes électroniques se développent rapidement et sont produits par plusieurs groupes comme Zeiss, Philips (devenu FEI puis FEI Thermo Fisher Scientific), JEOL, Hitachi, etc. En parallèle aux améliorations techniques, les méthodes de préparation d'échantillons se développent aussi et permettent de faire les premières images de bactéries, des virus et de particules de carbone. Plus tard dans les années 80 une nouvelle méthode dite de cryo-microscopie électronique est mise au point par Jacques Dubochet (Nobel 2017) et ses collègues et est à la base des études d'objets biologiques à haute résolution (Adrian et al., 1984).

En 1998, un travail collaboratif entre Harald Rose, Maximilian Haider, Knut Urban et Johannes Buchmann permet de mettre au point un correcteur d'aberration sphérique. Ce correcteur permet de modifier le facteur caractérisant l'aberration sphérique (Cs) et améliore alors la résolution spatiale du microscope devenant inférieur à l'ångström. Ce correcteur fabriqué par CEOS GmbH et monté dans les microscopes de Zeiss, JEOL et FEI.

En un siècle, la MET a progressé significativement et a permis de passer d'une résolution de quelques dizaines de nanomètres à ses débuts, à une résolution atomique actuellement. Sous réserve de la nature de l'échantillon car les propriétés des échantillons dans l'étude des matériaux en physique et des complexes macromoléculaires en biologie sont très différents comme nous le verrons par la suite.

### 3.2 Le microscope électronique à transmission

La microscopie électronique est une méthode d'imagerie qui utilise un faisceau d'électrons accélérés par une différence de potentiel élevé (voltage élevé) pour produire un agrandissement conséquent de la matière. Son utilisation en biologie permet l'étude d'un échantillon biologique par irradiation avec un faisceau électronique cohérent. Sur le plan théorique, le fonctionnement d'un MET n'est pas très différent d'un microscope photonique. Il y a une source, où les électrons remplacent les photons, des lentilles électromagnétiques qui remplacent les traditionnelles lentilles en verre, et des diaphragmes qui bloquent les électrons diffusés hors du faisceau souhaité.

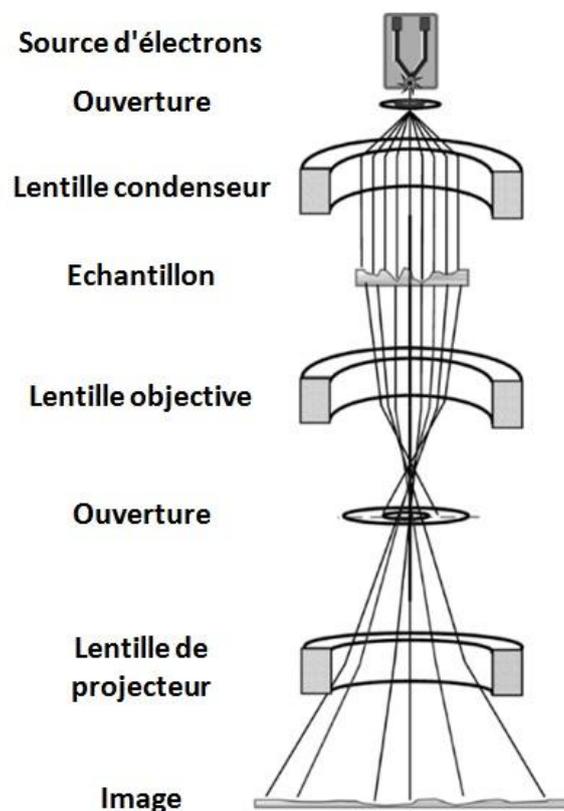
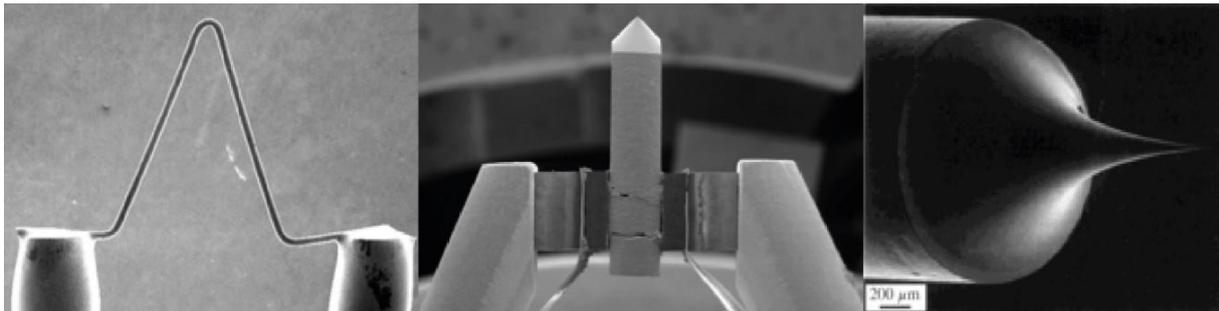


Figure 31 : Anatomie schématique d'un microscope électronique (adaptée de Orlova and Saibil, 2011).

### 3.2.1 Les sources d'électrons

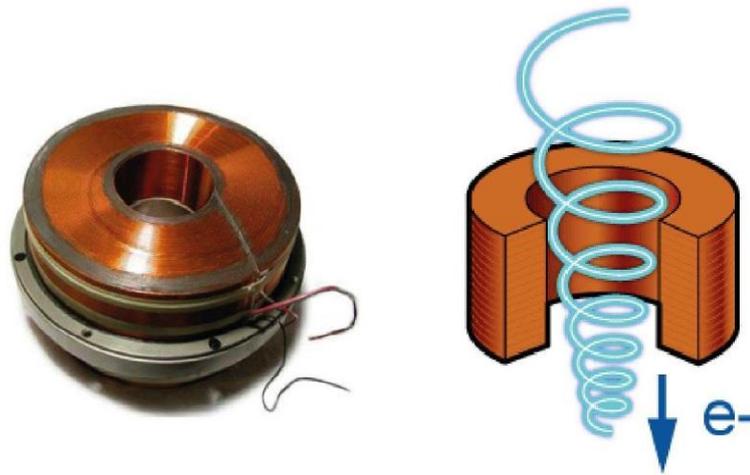
Les électrons sont obtenus grâce à un courant électrique circulant dans une pointe métallique. Ce courant va provoquer des collisions entre électrons voisins et si un électron possède assez d'énergie, il peut être extrait sous l'action d'un potentiel électrique dont la valeur dépend du métal utilisé. Au début de la MET, on utilisait un filament de tungstène en forme de V, dans ce cas le faisceau était émis lorsque le filament était chauffé à plus de 2500°C. Ensuite, on a utilisé un cristal d'hexaborure de lanthane, possédant une brillance supérieure et une zone d'émission des électrons réduite, il ne doit être chauffé qu'à 1500°C. Plus récemment a été développé un canon à émission de champs (FEG) qui est un cristal de tungstène chauffé à plus de 2000°C qui va émettre les électrons par une pointe. Ce dernier a plusieurs avantages dont une meilleure cohérence spatiale, une meilleure cohérence temporelle et une brillance plus performante du faisceau d'électron émis. Les microscopes électroniques pour la haute résolution sont équipés d'un FEG. Les électrons émis sont accélérés par une tension habituellement comprise entre 80 et 300 kV.



**Figure 32** : Les différentes sources d'électrons utilisées en microscopie électronique. De gauche à droite, un filament de tungstène en V, un cristal d'hexaborure de lanthane, un canon à émission de champ (FEG). (Adaptée de Williams and Carter, 2009 et nanoscience.com)

### 3.2.2 La formation de l'image

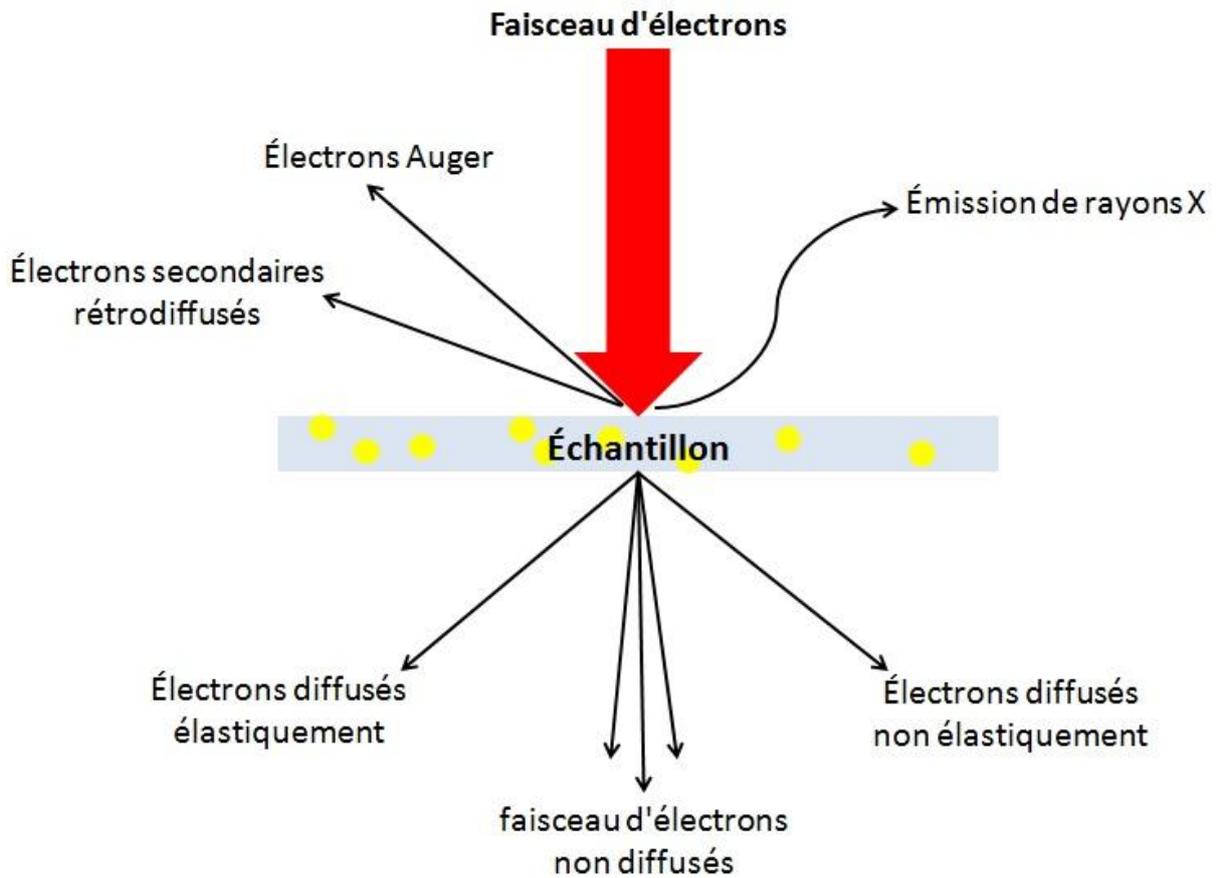
Après l'émission des électrons, ceux-ci sont dirigés vers un ensemble de lentilles électromagnétiques, qui sont physiquement des bobines de cuivres sous tension afin de générer un champ électromagnétique qui va permettre de modifier la trajectoire des électrons. Le but étant de reproduire les mêmes propriétés qu'aurait une lentille en verre pour des photons, notamment afin de focaliser le faisceau.



**Figure 33** : A gauche, une photo d'une lentille magnétique. A droite, le schéma du trajet parcourut par un électron à travers ce type de lentille (reproduit de ammrf.org.au).

Dans un MET, le faisceau est considéré comme un ensemble d'électrons possédant une longueur d'onde et une phase qui avance parallèlement par rapport à l'axe optique.

Les électrons passent ensuite par l'échantillon lui-même. Pour comprendre la formation du contraste de l'image, il faut comprendre comment se comporte l'électron. Il y a 3 cas de figure : première possibilité, l'électron n'interagit pas avec l'échantillon, il n'est pas dévié et ne perd pas d'énergie. Seconde possibilité, l'électron va subir une diffusion élastique lorsqu'il interagit avec un atome de l'échantillon et dévie de sa trajectoire mais ne perd pas d'énergie. La troisième possibilité est que l'électron va subir une diffusion inélastique lorsqu'il interagit avec un atome de l'échantillon et dévie de sa trajectoire initiale en perdant de l'énergie. Dans ce dernier cas, l'énergie perdue va induire la formation de radicaux libres conduisant à la dégradation de l'échantillon observé. Dans le cas d'un échantillon organique classique et une tension de 300 kV utilisé, les probabilités qu'un électron n'ait pas d'interaction avec un atome de carbone est de 80%, pour une diffusion élastique 5% et une diffusion inélastique 15%.

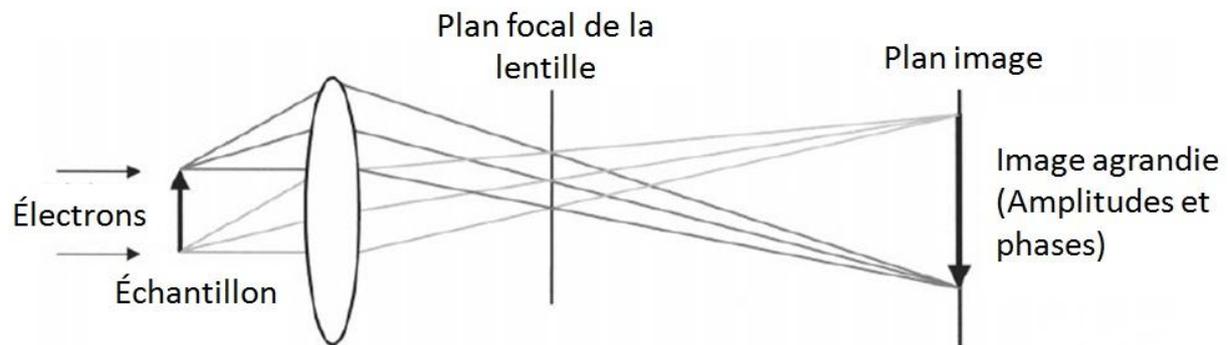


**Figure 34** : Schéma des différents types d'interactions entre les électrons et l'échantillon.

Ce sont ces interactions qui sont à l'origine du contraste observé. On peut distinguer 2 types de contraste. Il y a le contraste d'amplitude qui résulte de l'absorption d'une partie du faisceau dans l'échantillon. Un échantillon biologique étant principalement constitué d'atomes légers (H, O, N et C), le contraste d'amplitude est faible, mais il est amplifié avec le diaphragme objectif qui stoppe les électrons qui se diffusent latéralement avec de grands angles. Le second type est le contraste de phase, il résulte du décalage de phases entre les électrons non déviés et les électrons déviés lors d'une interaction élastique. Ce décalage dépend de la longueur du trajet des électrons qui peut varier grâce à la défocalisation de la lentille objective (c'est à dire que le point focal se trouve juste après l'échantillon).

La formation de l'image dépend de la diffusion des électrons sur l'échantillon et de la capacité des lentilles à refocaliser ces électrons. Après le passage des électrons par l'échantillon, ils passent par la lentille objective qui fait converger le faisceau d'électrons. Ainsi la formation de l'image peut se décomposer en deux étapes. Les électrons sont tout d'abord rassemblés au niveau du plan focal de la lentille, ceci formant le spectre de diffraction de l'objet. Ce spectre peut être décrit mathématiquement par la transformée de Fourier de l'objet. Par la suite, les électrons sont

recombinés pour former une image de l'objet. Le passage du spectre de diffraction à l'image peut être décrit mathématiquement par la transformée de Fourier inverse de l'objet.



**Figure 35** : Schéma du chemin optique de la formation de l'image (adapté de Saibil, 2000).

### 3.2.3 Les limites physiques rencontrées par la MET

La résolution en imagerie peut être caractérisée par l'angle ou la distance minimale qui doit séparer deux points contigus pour qu'ils soient correctement discernés. En MET, la résolution pour un échantillon de qualité est dépendante de la tension utilisée, de la qualité du système optique et de la caméra.

La résolution dépend de la longueur d'onde  $\lambda$ , de l'indice de réfraction du milieu traversé  $n$  et de l'angle d'incidence du faisceau  $\alpha$ , décrit par le critère de Rayleigh. Elle peut être exprimée selon l'équation 1 et simplifiée pour de petits angles à l'équation 2.

$$d = \frac{0.61 \times \lambda}{n \times \sin \alpha} \quad (1)$$

$$d = \frac{\lambda}{2} \quad (2)$$

La longueur d'onde est dépendante du voltage utilisé en MET. Pour un microscope qui est utilisé à un voltage  $V = 300 \text{ kV}$  on aura une longueur d'onde de  $\lambda = 0.022 \text{ \AA}$  d'après l'équation 3, la résolution théorique est donc de  $0.011 \text{ \AA}$  selon l'équation 2.

$$\lambda(\text{\AA}) = \frac{12.3}{\sqrt{V}} \quad (3)$$

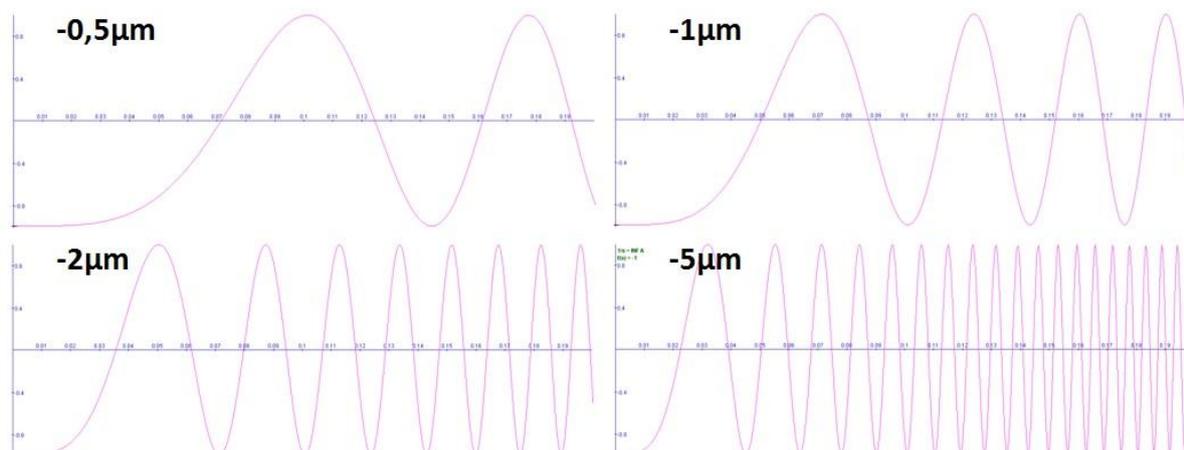
A cela s'ajoute le facteur de la caméra qui joue un rôle majeur dans la résolution des images enregistrées. Cette résolution est limitée par la taille du pixel du détecteur. Cette limitation est expliquée par le théorème d'échantillonnage (théorème de Nyquist-Shannon), qui énonce que l'échantillonnage d'un signal exige une fréquence d'échantillonnage supérieure au double de la fréquence maximale. Concrètement dans le cas de la MET, cela signifie que la résolution maximale est égale au double de la taille du pixel du détecteur, divisé par le grossissement utilisé, c'est à dire au double d'ångström par pixel.

Les lentilles présentent également des caractéristiques limitant la résolution. Tout d'abord, l'aberration sphérique ( $C_s$ ). Cette propriété définit que la déviation des électrons sera plus forte proportionnellement à leurs éloignements de l'axe optique. Ainsi un point sera représenté par un disque de taille proportionnel au coefficient d'aberration sphérique du microscope. Il y a également l'aberration chromatique ( $C_c$ ) qui définit que la déviation des électrons dépend de la longueur d'onde du rayon incident, c'est à dire que si les électrons n'ont pas tous la même énergie, ils ne convergeront pas tous au même point. Un autre défaut qui peut être rencontré est l'astigmatisme, qui est défini par un grossissement non uniforme dans toute les directions, c'est-à-dire qu'il y a aura une déformation des échelles de l'objet différentes en fonction de la direction.

La transmission de l'information par le MET n'est pas parfaite, en effet le microscope introduit des modifications de l'information qu'il est nécessaire de corriger. On appelle ces modifications fonction de transfert de l'instrument que l'on peut modéliser mathématiquement par la fonction de transfert de contraste (FTC). L'image de chaque objet sera donc affectée par la FTC du microscope. Il en résulte que la transformée de Fourier d'une image est la résultante du produit de convolution entre la transformée de Fourier de l'échantillon et de la FTC. La définition mathématique de la FTC est décrite dans l'équation 4 où  $C_s$  est le coefficient d'aberration sphérique,  $\Delta z$  la défocalisation,  $\kappa$  la fréquence spatiale et  $\lambda$  la longueur d'onde des électrons.

$$FTC = -2 \sin \left[ \pi \left( \Delta z \lambda \kappa^2 - \frac{C_s \lambda^3 \kappa^4}{2} \right) \right] \quad (4)$$

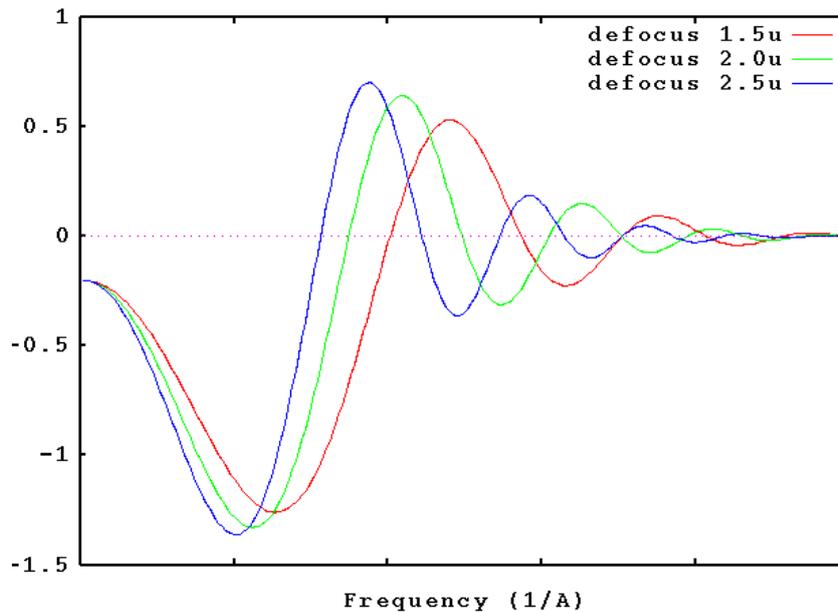
La fonction de transfert de contraste a une amplitude d'oscillation comprise entre 1 et -1 qui correspond à une inversion du contraste. La défocalisation fait fluctuer la fréquence d'oscillation. En effet plus la valeur de défocalisation augmente, plus la fréquence d'oscillation augmente.



**Figure 36** : Courbe de la fonction FTC théorique pour plusieurs valeurs de défocalisation. Les courbes sont obtenues avec le logiciel CTF simulation (Jiang and Chiu, 2001). Ce sont des courbes théoriques sans fonction enveloppe.

Si on utilise une forte défocalisation, le signal est plus fort, cependant, l'augmentation de la fréquence d'oscillation fera que le nombre de passage de la courbe par zéro sera également plus élevé. Une amplitude de zéro correspond à un signal nul et donc perdu, dans ce cas il n'y a pas de différence d'intensité entre les électrons déviés et non déviés. De plus lorsque l'on passe en valeurs négatives, il y a une inversion du contraste. Il est donc nécessaire de corriger cette inversion en rendant toutes les valeurs positives ("phase flipping"), et en plus d'avoir une gamme de défocalisations variée pour pallier à ce passage par zéro afin d'avoir du signal sur toute la gamme de fréquences spatiales.

L'ensemble des limites liée aux aberrations des lentilles et de la stabilité générale du microscope va affecter la FTC et diminuer progressivement le signal en fonction de la fréquence spatiale. Ce phénomène est décrit par la fonction d'enveloppe.



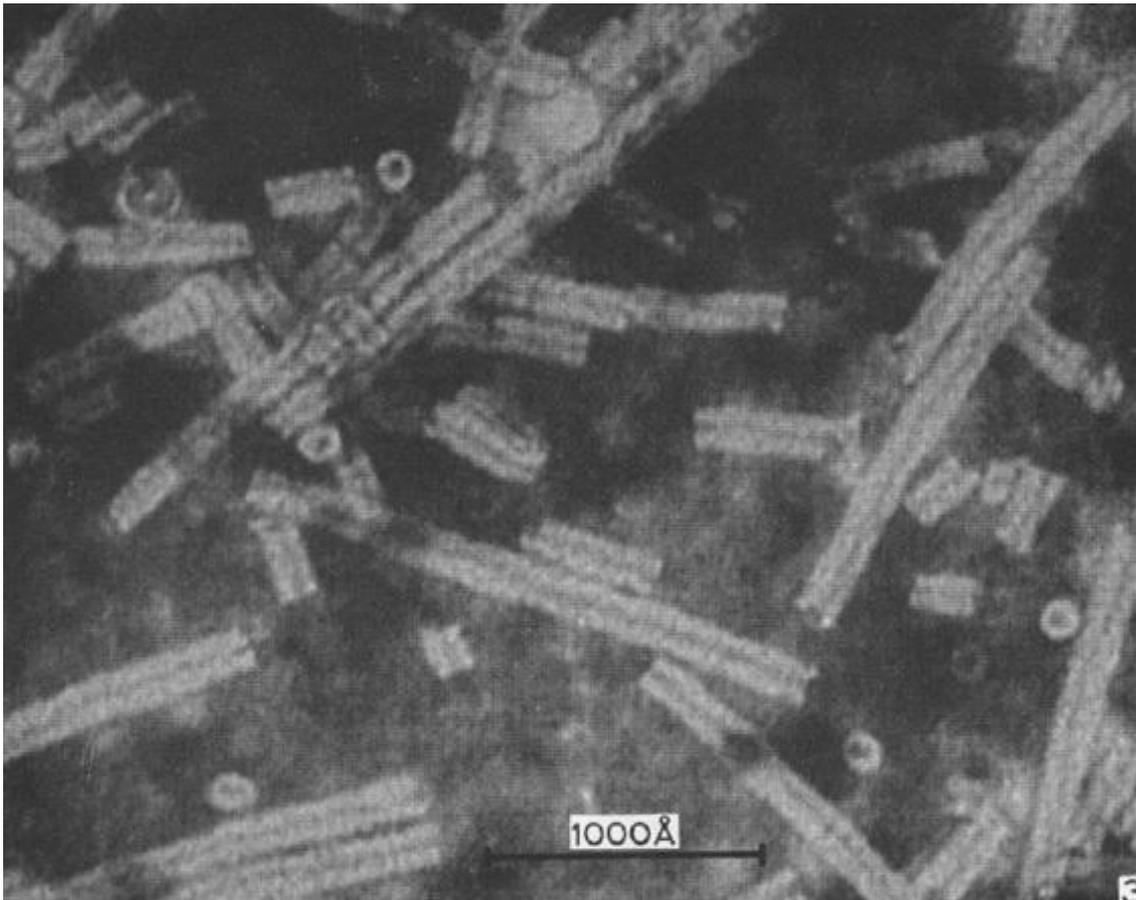
**Figure 37** : Effet de la fonction enveloppe sur la FTC théorique.

### 3.3 L'étude d'un échantillon biologique

L'utilisation d'un MET pour l'analyse structurale en biologie a soulevé plusieurs défis parmi lesquels la préservation de l'intégrité de l'échantillon observé afin de préserver la forme native du complexe d'intérêt ou encore le passage d'une information en 2 dimensions (2D) vers une information en 3 dimensions (3D). Il y a deux méthodes majeures pour ces observations, la coloration négative et la cryo-microscopie électronique. La coloration négative est rapide et permet de valider ou non une préparation d'échantillons (purification, tampon, concentration etc.), lorsqu'on a trouvé des paramètres adaptés à notre étude, commencent les étapes d'optimisation pour la cryo-microscopie électronique puis finalement en cas de réussite, l'acquisition permettant d'atteindre la haute résolution structurale dans une reconstruction 3D.

### 3.3.1 La coloration négative

Cette méthode est la plus simple et la plus rapide à mettre en œuvre pour observer des échantillons biologiques. Elle a été développée par Brenner et Horne en 1959 (Brenner and Horne, 1959), qui à l'époque utilisaient de l'acide phosphotungstique comme agent colorant sur des virus. Le principe de cette méthode est d'embrober l'échantillon avec une solution de sels d'atomes lourds tel que l'uranium qui permet de préserver en partie la structure et d'augmenter le contraste.



**Figure 38** : Exemple d'image de coloration négative de virus prise en 1959 (Brenner and Horne, 1959).

Avec cette méthode, les échantillons en solution sont adsorbés sur une grille métallique (généralement du cuivre et du rhodium) recouverte d'un film de carbone fin. Ce sont typiquement des complexes protéiques. L'excès d'eau est absorbé à l'aide d'un papier filtre. Une solution contenant un agent contrastant, tel de l'acétate d'uranyle, est ajoutée sur la grille pendant quelques secondes puis absorbée avec du papier filtre. On peut répéter cette opération 2-3 fois. L'agent contrastant va se fixer sur les particules adsorbées sur la grille. Puis on va laisser sécher la grille à température ambiante avant de pouvoir l'observer. De par sa forte masse atomique, le contrastant

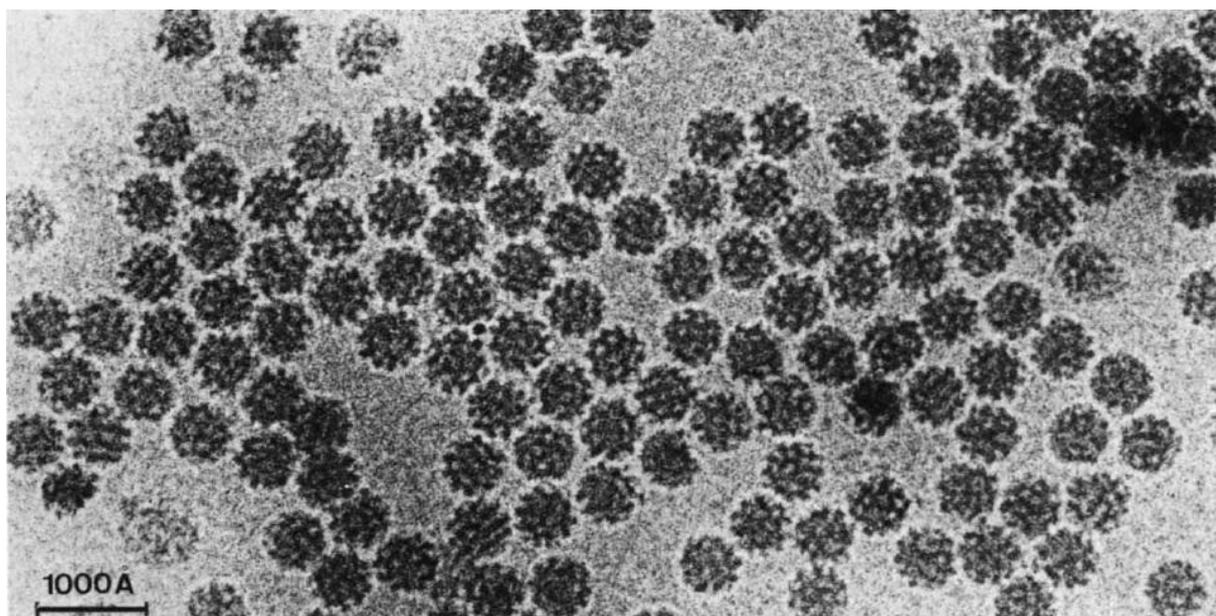
dévie les électrons dans le diaphragme objectif. Ainsi l'échantillon biologique apparaît plus clair que ce qui l'entoure, d'où le nom de coloration négative. Sur les micrographes l'échantillon apparaît blanc sur un fond sombre. Il faut noter qu'on n'observe pas directement l'échantillon mais la répartition du contrastant autour de l'échantillon qui se localise préférentiellement dans les zones hydrophiles de surfaces. De plus, l'échantillon étant recouvert de métaux lourds, il n'y a plus de molécules d'eau et par conséquent, le vide du microscope affectera moins l'intégrité de l'échantillon. Cette méthode rencontre néanmoins des limitations physiques, en effet l'adsorption sur la surface de carbone peut déformer l'objet, le séchage de l'échantillon peut entraîner son effondrement sur lui-même, la coloration est potentiellement que partielle et les agents de coloration peuvent former des agrégats de métaux lourds autour de l'échantillon. De ce fait, la limitation de la résolution est autour de 15 à 20 Å.



**Figure 39** : Photographie du CM120 (Philips) de l'institut IGBMC/CBI. Il s'agit du microscope pour la coloration négative.

### 3.3.2 La cryo-microscopie électronique

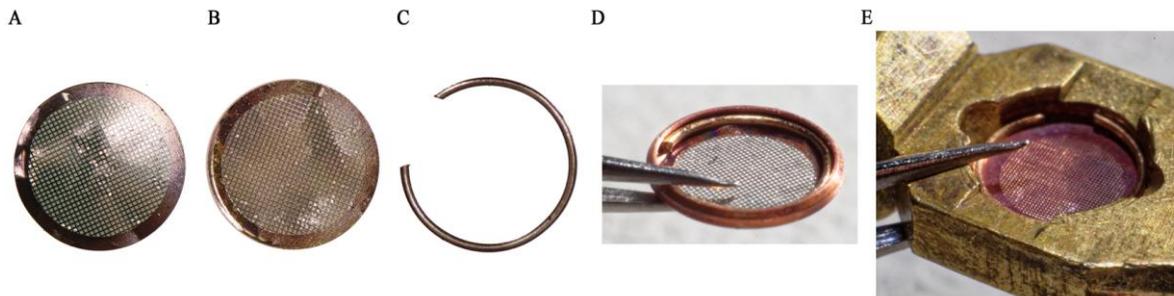
Cette méthode consiste à piéger le complexe biologique d'intérêt dans une fine couche de glace amorphe pour préserver l'état natif du complexe en solution. Cette méthode a été développée par Jacques Dubochet et Marc Adrian à l'EMBL de Heidelberg au début des années 80 (Adrian et al., 1984).



**Figure 40** : Exemple d'image de Cryo-microscopie électronique de virus prise début des années 80 (Adrian et al., 1984).

Pour ce type de préparation on utilise des grilles pour lesquelles le film de carbone est perforé ("grilles à trous"). Pour la préparation, l'échantillon est déposé sur une grille, l'excès d'échantillon est retiré par absorption avec du papier filtre, le but étant de n'avoir qu'une très fine couche d'échantillon formant un double ménisque en équilibre dans les trous du film de carbone. Une fois préparée, la grille est immédiatement plongée dans un bain d'éthane liquide qui est approximativement à  $-170^{\circ}\text{C}$ , lui-même refroidi dans un bain d'azote liquide qui est à  $-196^{\circ}\text{C}$ . La phase de congélation doit être très rapide, pour ce faire les propriétés du cryogène sont importantes afin d'assurer la formation de glace amorphe. La glace amorphe est de la glace sans structure cristalline classique, c'est-à-dire que les molécules d'eau ont refroidi suffisamment vite pour se figer sans avoir le temps de former des réseaux et donc des cristaux. Un refroidissement trop lent entraînerait la formation de glace hexagonale, tandis qu'un réchauffement de glace amorphe au-dessus de  $-140^{\circ}\text{C}$  entraînerait la formation de glace cubique. Dans ces deux derniers cas, l'échantillon ne sera pas observable avec un cryo-microscope électronique. Après congélation, il est donc

essentiel de toujours manipuler l'échantillon dans l'azote liquide avec des instruments préalablement refroidis à la même température.



**Figure 41** : Photographies de grilles. L'ensemble des objets présentés ici sont à l'échelle, les uns par rapport aux autres. Ils présentent quelques marques d'usures car ne sont pas neuf mais déjà utilisé pour mes manipulations décrites dans la partie résultat. A : Grille Quantifoil en carbone. B : Grille Quantifoil en or. C : Anneaux de clip (clip ring) permettant le maintien de la grille sur son support dans un microscope électronique. D : Grille montée sur sa monture (cartouche) et bloquée par un clip ring. Ce montage est destiné au cryo-microscope électronique Titan Krios. E : Grille montée sur sa cartouche et bloquée par un clip ring. Ce montage est destiné au cryo-microscope électronique Polara.

Cette méthode de préparation d'échantillon est vitale pour obtenir une meilleure résolution que celle obtenue en coloration négative. En effet, la congélation permet de travailler sous vide avec un échantillon hydraté car la sublimation est alors faible. Ici le principal avantage est que l'échantillon reste hydraté et est congelé dans son dernier état en solution, qui est de fait beaucoup plus proche de l'état physiologique de la protéine ou du complexe. La qualité d'un échantillon, dépendra de plusieurs facteurs. L'épaisseur de la glace est importante et demande des optimisations en fonction de l'échantillon. Une glace trop épaisse entrainera une diminution du contraste, et en fonction de la concentration pourrait entrainer une superposition des particules à plusieurs profondeurs dans la glace. Mais une glace trop fine, favorisera les orientations préférentielles des complexes voire l'impossibilité de faire des images, car elle risque de casser sous le faisceau électronique. La nature du tampon utilisé pour la congélation est également importante pour permettre un bon contraste et une bonne distribution des particules, encore une fois ces paramètres sont échantillons-dépendants. Dans tous les cas, l'utilisation de glycérol, de sucrose ou encore de concentrations élevées en sels est fortement déconseillée car réduisent le contraste. La concentration de protéines peut être environ 10 fois supérieure à celle utilisée pour de la coloration négative pour avoir une distribution homogène dans la glace, ce paramètre est aussi à optimiser en fonction des échantillons. Les

échantillons sont ensuite observés dans des cryo-microscopes électronique en faisant en sorte de garantir une chaîne du froid sans faille.



**Figure 42** : Photographie du Titan Krios (FEI) de l'institut IGBMC/CBI. Il s'agit d'un microscope pour la cryo-microscopie électronique.



**Figure 43:** Photographie du Polara (FEI) de l'institut IGBMC/CBI. Il s'agit d'un microscope pour la cryo-microscopie électronique.

### 3.4 Les conditions d'acquisition

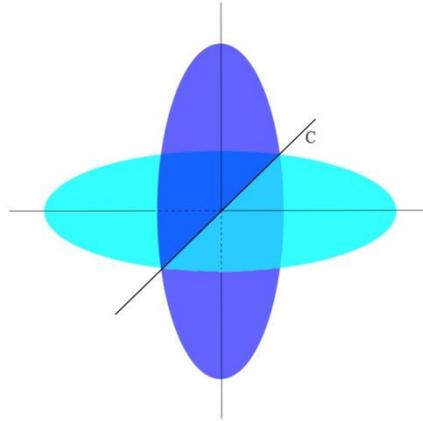
Les échantillons biologiques sont très sensibles aux dommages causés par l'irradiation du faisceau électronique. Une exposition trop longue ou trop forte induit des radicaux libres qui détruisent progressivement l'échantillon et les particules se diffusent progressivement car la glace amorphe conserve les propriétés physiques d'un fluide. Compte tenu de cette limitation, les acquisitions se font en mode "faible dose" qui permet de capturer des micrographes en contrôlant strictement la dose totale d'électrons utilisés. Pour y parvenir, la recherche de zones à photographier se fait à faible grossissement pour limiter fortement le nombre d'électrons par unité de surface. A grossissement d'acquisition il est nécessaire de faire plusieurs mesures avant la capture du micrographe comme par exemple l'autofocus pour déterminer la hauteur du focus. Pour ce faire, une zone dans du carbone

brut de la grille est utilisée à côté de chaque trou, afin d'avoir un défocus correct sans pré-exposer inutilement l'échantillon. Puis enfin vient la prise du ou des micrographes dans le trou en fonction du cas et du microscope utilisé. Ce mode de prise de vue est nécessaire au regard de la forte sensibilité de l'échantillon aux irradiations, mais il impose aussi un faible rapport signal/bruit présent dans les micrographes. Ce rapport faible nécessite d'être amélioré par moyennation d'images similaire lors des étapes de traitement d'images.

### 3.5 Principes généraux de la reconstruction 3D en microscopie électronique

Les micrographes enregistrés sont des projections 2D de l'échantillon observé. Un jeu de données correspond donc à un ensemble de projections très bruitées d'un objet prises dans des orientations idéalement variées et aléatoires. Lorsque l'on effectue la projection d'un objet tridimensionnelle en microscopie électronique à transmission, chaque projection correspond à la somme des densités le long de l'axe de projection. Il en résulte une image en 2 dimensions. Si on effectue l'inverse, c'est à dire une rétroprojection, sur un ensemble de projections 2D obtenues suivant des angles de projections déterminés, on peut reconstruire l'objet en 3 dimensions. C'est le principe de base utilisé ici pour la reconstruction 3D.

Le principal problème est que pour effectuer une rétroprojection, il est nécessaire de connaître pour chaque projection 2D l'orientation de l'objet correspondante. Pour se faire, une première méthode utilisée est basée sur le concept des lignes communes des projections. Par ailleurs, le théorème de la section centrale décrit que dans l'espace de Fourier, chaque projection 2D correspond à une section centrale de la transformée de Fourier 3D de l'objet. L'orientation de cette section est directement liée à l'orientation de l'objet 3D qui a donné la projection 2D considérée. Ainsi, deux projections 2D de l'objet correspondront à deux sections de la transformée de Fourier 3D de l'objet, il y aura donc une ligne commune qui correspond à l'intersection des deux sections. Cette méthode est par exemple utilisée dans le logiciel IMAGIC (van Heel and Keegstra, 1981; van Heel et al., 1996). Et permet d'attribuer les angles d'Euler relatifs. Ces angles correspondent à un ensemble de 3 angles pour décrire l'orientation de l'objet. Cette méthode a l'avantage de ne nécessiter aucune information préalable, notamment pas besoin de connaître une structure similaire qui pourrait servir de référence.

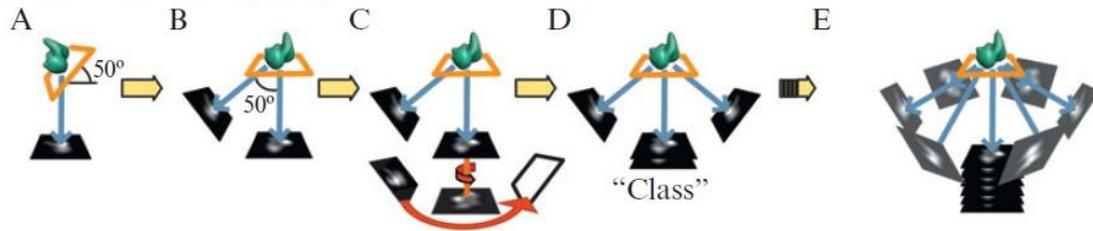


**Figure 44** : Schéma d'une ligne commune (C) de deux projections dans l'espace de Fourier d'une même particule.

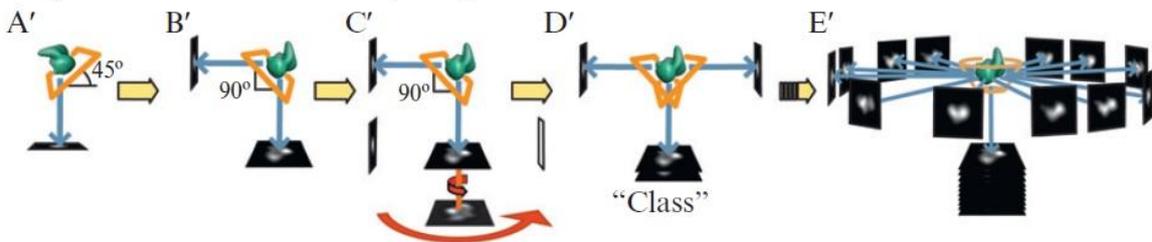
Il y a une seconde méthode pour laquelle il faut confronter les images en deux dimensions à une structure 3D ou un modèle 3D qui est une structure à basse résolution de l'objet d'étude. Cette structure de référence sera projetée dans toutes les directions générant ainsi un ensemble de projections 2D qui seront comparées avec les projections 2D expérimentales. Ces projections 2D sont classifiées pour regrouper les objets qui ont la même orientation ensemble et ainsi, en les moyennant permettent d'améliorer le rapport signal/bruit.

Le choix de la structure de référence est très important en raison du bruit important présent dans les projections, car les images vont être alignées par rapport à cette référence, même si elles présentent du bruit au lieu de vraies particules, ce qui peut mener à des analyses erronées (Henderson, 2013; Subramaniam, 2013; Van Heel, 2013). On aura alors l'illusion que la reconstruction progresse alors que le "signal" ne proviendra que du bruit contenu dans les images et de la référence. La structure de référence peut être obtenue grâce à plusieurs méthodes. Il y a l'inclinaison conique aléatoire et la reconstruction inclinée orthogonale qui utilisent toutes les deux des paires inclinées de particules. Dans le cas de l'inclinaison conique aléatoire (Radermacher et al., 1987) les angles sont typiquement de  $50^\circ$ , alors que pour l'inclinaison orthogonale (Leschziner and Nogales, 2006) les angles sont de  $90^\circ$  (avec une paire tilté de  $\pm 45^\circ$ ). La reconstruction avec une inclinaison orthogonale est plus récente et présente l'avantage de ne pas avoir de cônes manquant, contrairement à l'inclinaison conique aléatoire.

Random Conical Tilt (RCT)

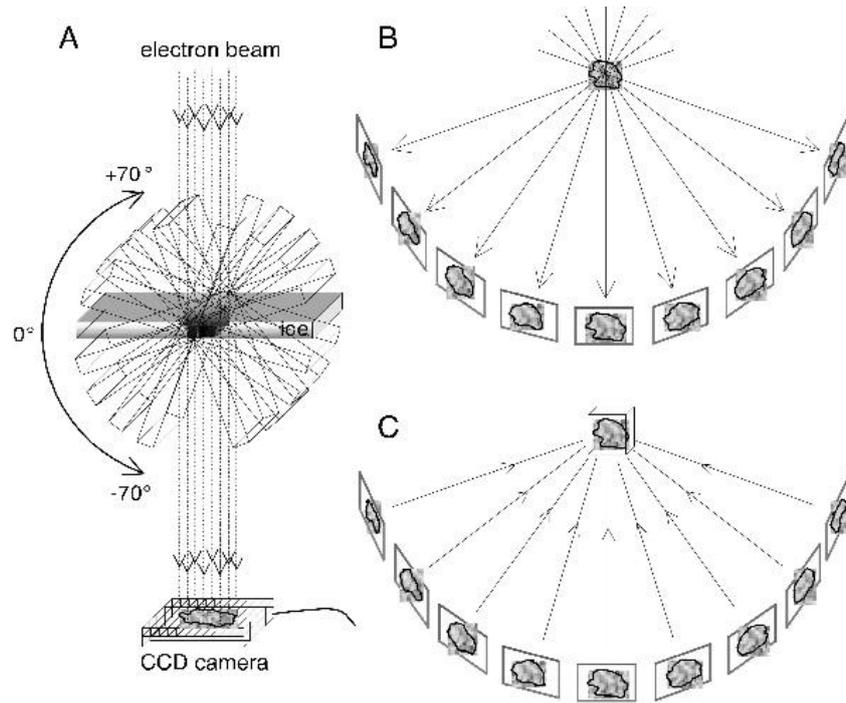


Orthogonal Tilt Reconstruction (OTR)



**Figure 45** : Schéma de l'inclinaison conique aléatoire et de la reconstruction inclinée orthogonale. A et A' : Acquisition de la première image avec l'angle sélectionné. B et B' : Acquisition d'une seconde image après inclinaison pour créer des paires inclinées. C et C' : Les premières images sont alignées entre elles si elle représentent la même vue. Les secondes images sont repositionnées en fonction de la rotation que les premières particules ont subie pendant l'alignement. D et D' : Les images alignées sont regroupées en classes de particules ayant la même orientation. E et E' : La densité est reconstruite si le nombre de vues le permet. (Reproduit de Leschziner, 2010).

Enfin, nous pouvons également utiliser une méthode où tous les angles de rotations sont connus, il s'agit de la cryo-tomographie électronique qui permet de faire une série d'images de la même particule avec des angles généralement compris entre  $-60^\circ$  et  $+60^\circ$  avec un incrément de quelques degrés seulement. Cette série d'images sera ensuite alignée et combinée pour obtenir une reconstruction 3D. Cette approche est utilisée principalement pour des études cellulaires, et n'est utilisée que depuis récemment pour l'étude de particules isolées car le rapport signal/bruit limite sensiblement la résolution que l'ont peu obtenir.



**Figure 46** : Schéma du principe de tomographie. A : Le porte échantillon est incliné de manière incrémentielle autour d'un axe perpendiculaire au faisceau d'électrons. B : Les images projetées par un spécimen à des angles d'inclinaison successifs. C : L'objet imagé est reconstruit dans une carte de densité 3D (appelée tomogramme) par une procédure de rétroprojection pondérée. (Reproduit de Steven and Belnap, 2005).

Lors de l'affinement d'une structure en cryo-ME, les données d'orientation obtenues à partir de la référence, par rétroprojection des données expérimentales, sont utilisées pour une nouvelle reconstruction 3D. Cette nouvelle structure sera à son tour utilisée comme référence en créant des projections selon des angles définis pour continuer d'affiner les orientations des projections 2D de manière itérative. Ce processus se fait dans deux sous-jeux de données équivalents divisés aléatoirement. A chaque itération, la corrélation entre les deux reconstructions est calculée dans l'espace de Fourier (FSC pour "Fourier Shell Correlation") et permet ainsi de suivre l'amélioration de la résolution. Lorsque la FSC ne varie plus, les données sont fusionnées et une reconstruction finale est réalisée avec l'ensemble des images de départ.

Lors de l'interprétation de la reconstruction finale, il y a des différences avec la cristallographie liées à la méthode elle-même qu'il faut prendre en compte. En effet lorsqu'on obtient une carte par cristallographie aux rayons X il s'agit d'une carte de densité électronique. En cryo-ME on utilise des électrons pour la mesure, la carte obtenue ne sera donc pas une carte de densité électronique mais une carte du électrostatique, ainsi les acides aminés chargés positivement présentent une densité plus

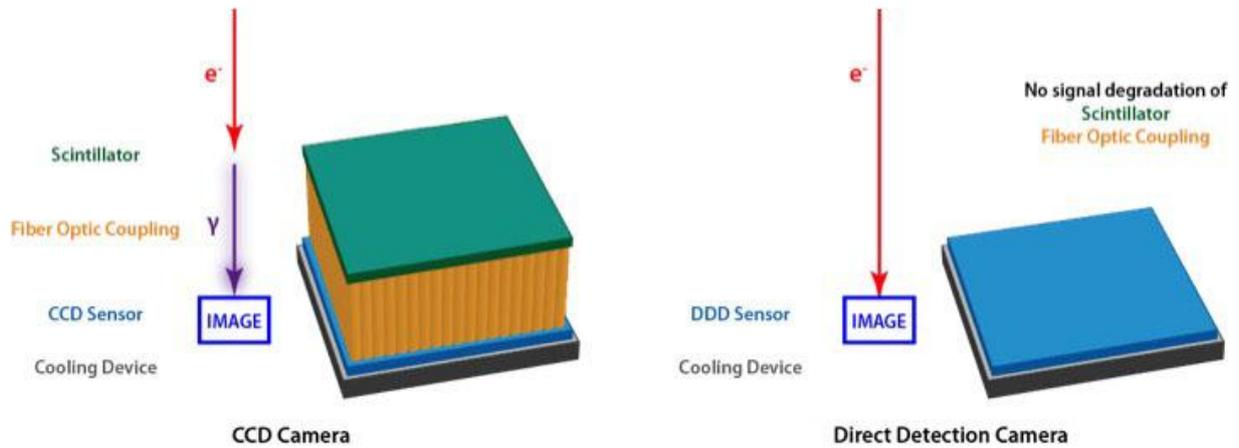
importante qu'en cristallographie, à l'inverse les charges négatives sont minimisées. Ce comportement peut être une source d'erreur d'interprétation de la carte pour les hautes résolutions

### 3.6 La révolution de la résolution

Au cours des dernières années, de nombreux développements ont permis de franchir un pas important en microscopie électronique, parfois qualifié de "révolution de la résolution" (Kühlbrandt, 2014). Avant ce bond technologique, la microscopie électronique à transmission était décrite comme une approche structurale à basse résolution par rapport à la cristallographie aux rayons-X. En effet les limites de résolution se situaient autour de 5 à 10 Å et ne permettaient donc pas d'avoir des informations précises sur l'organisation atomique, mais seulement sur la topologie des objets étudiés. Ces données étaient principalement utilisées pour la topologie de grands complexes pour lesquels on ne parvenait pas à avoir des cristaux et dans lesquels on intégrait des données de haute résolution de domaines plus petits obtenues grâce à l'une des deux autres méthodes citées. La cryo-microscopie électronique a finalement franchi la barrière de la résolution quasi atomique dès 2008 avec des structures de virus résolues entre 3,5 et 4 Å de résolution (Jiang et al., 2008; Yu et al., 2008; Zhang et al., 2008). Depuis les années 2013, les innovations techniques ont permis de résoudre des structures par cryo-ME en dessous de 5 Å de résolution de façon relativement routinière pour divers types de complexes, allant même jusqu'à 1,8 Å pour la glutamate déshydrogénase (Merk et al., 2016).

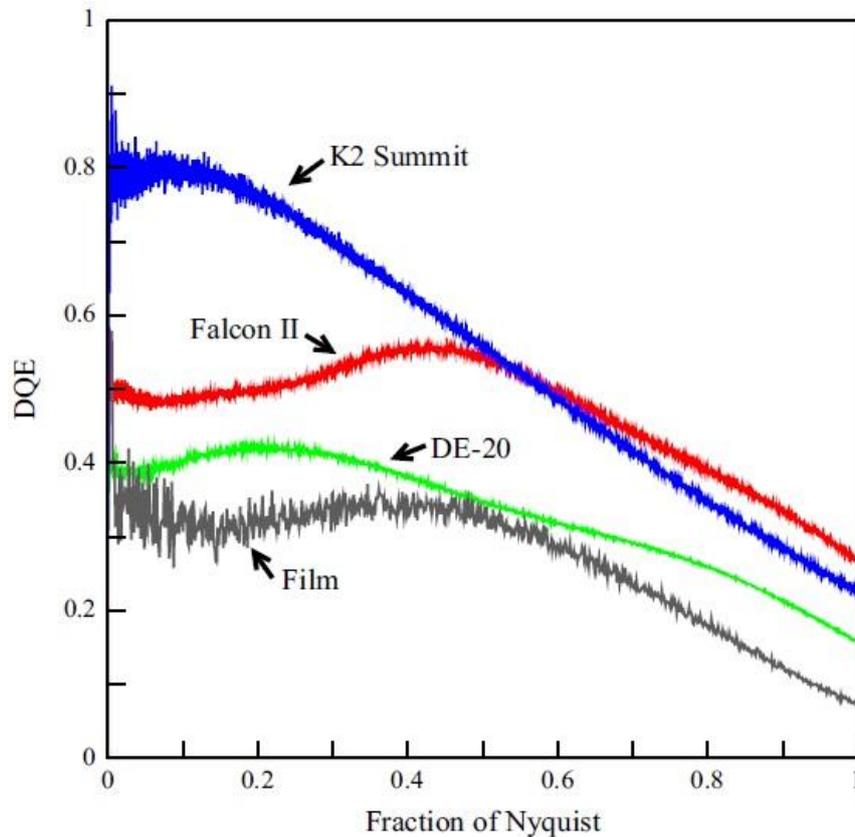
#### 3.6.1 Des caméras à détections directes

Un élément majeur pour atteindre la haute résolution est la qualité des dernières générations de caméras à détection directe. En effet, auparavant la détection était indirecte, les électrons étaient d'abord convertis en photons par un scintillateur pour amplifier le signal, puis transmis au détecteur CCD (Charge-Coupled Device; dispositif à transfert de charge en français) par un réseau de fibres optiques plus petits que les pixels eux-mêmes. Aujourd'hui ces nouveaux détecteurs utilisent un détecteur CMOS (Complementary Metal-Oxide-Semiconductor; oxyde métallique semi-conducteur complémentaire) et détectent les électrons directement lorsqu'ils traversent la fine couche de silicium du détecteur.



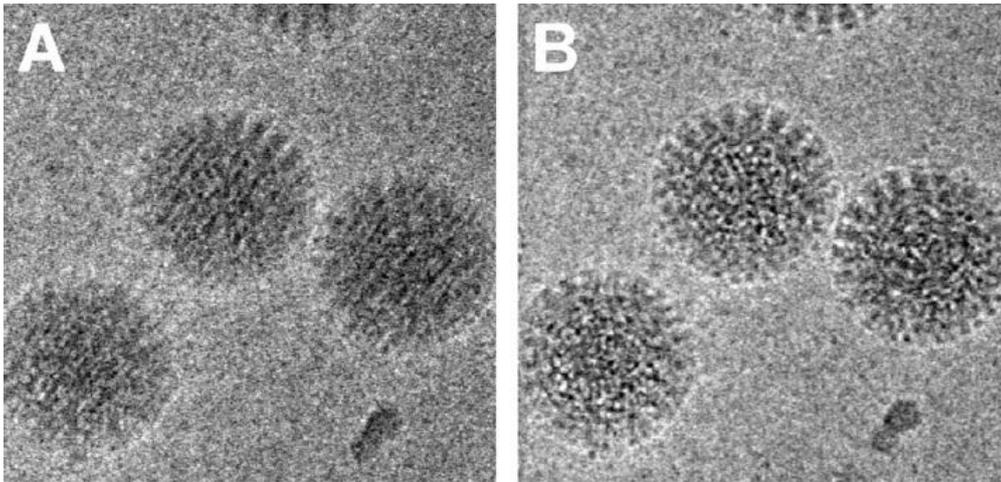
**Figure 47** : Comparaison des deux systèmes de détections pour la cryo-microscopie électronique. A gauche, schéma de la structure d'un capteur CCD. A droite, schéma de la structure d'un capteur à détection directe. Reproduit de directelectron.com.

Ces caméras (DE-series (Direct Electron); Falcon I, II et III (FEI); K2 et K3 (Gatan)) offrent un meilleur rapport signal/bruit car il n'y a pas de dégradation du signal due à la conversion des électrons en photons par le scintillateur. Pour mesurer la performance d'un détecteur on utilise l'efficacité quantique de détection (Detective Quantum Efficiency; DQE) qui permet de mesurer la quantité de signal détecté par rapport au bruit pour une fréquence spatiale donnée.



**Figure 48** : Comparaison DQE entre plusieurs caméras et films photographiques. A noter, à titre de comparaison supplémentaire que les films négatifs sont plus sensibles que les caméra CCD à détection indirecte (Reproduit de McMullan et al., 2014).

Il y a aussi une vitesse d'acquisition plus élevée. Cette vitesse permet d'ajouter le facteur temps à la prise de vue grâce à un mode film. Au lieu d'avoir un micrographe unique, on a une série d'images sur un endroit de l'échantillon. Les avantages sont de pouvoir réaligner les sub-micrographes entre eux et ainsi en corrigeant le mouvement des particules durant un long temps d'exposition comme expliqué précédemment. Cette méthode permet également de contrôler la dose totale après acquisition, en excluant les premières images qui peuvent présenter quelques défauts à cause des effets de charge mais aussi les dernières images pour lesquelles l'échantillon peut être trop détérioré par l'irradiation. On peut également pondérer chaque vue les unes par rapport aux autres en fonction de la dose totale reçue pour améliorer la qualité du signal. Ce mode permet d'augmenter considérablement la qualité des micrographes. Différents logiciels existent pour corriger l'alignement des films, comme MotionCorr (Li et al., 2013) et Unblur (Grant and Grigorieff).



**Figure 49** : Effet de l'alignement des images d'un film. Ce film de 60 micrographes n'est pas aligné et moyenné à gauche, et est aligné et moyenné à droite. On constate l'augmentation de détails structuraux du rotavirus après alignement.

En plus de ces nouvelles innovations de caméra, ont été développés les caméras Falcon III et Gatan K2 summit et Gatan K3 summit disposant de nouveaux modes d'enregistrement. Il y a le mode classique, le mode intégré qui, lorsqu'une charge électronique est diffusée sur plusieurs pixels, elle collecte les charges de chaque pixel individuellement. Le mode comptage où la charge électronique est enregistrée uniquement sur le pixel où elle est la plus forte. Enfin le mode super-résolution qui est propre aux caméras de Gatan, où chaque pixel est divisé en 4 pixels virtuels grâce à ce que détectent les pixels voisins. Ainsi le micrographe final aura 4 fois plus de pixels enregistrés que ce que le détecteur possède physiquement.

### 3.6.2 Le traitement d'images

Le traitement d'images et la reconstruction 3D représentent un autre aspect important pour la révolution de la résolution de la cryo-ME moderne car il permet d'obtenir des informations 3D très détaillées de l'objet d'intérêt. Lorsque l'on débute un processus de reconstruction 3D, l'hypothèse de départ est que les images décrivent le même objet, c'est à dire avec une conformation structurale homogène. Il est facile de valider cette hypothèse dans le cas d'un tomogramme, car l'ensemble des images sont enregistrées sur un seul et unique objet. Cependant c'est rarement le cas lorsque des techniques de moyennation sont utilisées pour une reconstruction à partir de particules isolées. L'hétérogénéité des échantillons peut rendre l'interprétation des cartes 3D difficile, voire impossible dans certain cas (Orlov et al., 2017). L'effet concret observé sera alors une limite significative de la résolution atteignable. Pour cette raison, les méthodes de classification 3D permettant le tri des

structures deviennent un outil essentiel pour l'analyse à haute résolution des données cryo-ME. On peut ainsi analyser simultanément plusieurs structures en équilibre les unes avec les autres en solution. Pour une telle étude, plusieurs approches sont possibles. On peut se baser sur une analyse de corrélation croisée utilisant des structures de références (Gao et al., 2004). On peut également se baser sur une analyse statistique multivariée (MSA), en incluant une analyse de variance locale dans les images de particules (Klaholz et al., 2004; Orlova and Saibil, 2010; White et al., 2004), une méthode de ré-échantillonnage et de bootstrap pour identifier des régions flexibles dans un complexe macromoléculaire et effectuer des classifications 3D (Fischer et al., 2010; Klaholz, 2015; Liao et al., 2015; Penczek et al., 2006; Simonetti et al., 2008). Les méthodes basées sur le MSA et le maximum de vraisemblance (ML) sont les deux méthodes à présent couramment utilisées car elles se révèlent plus robustes lors du traitement d'images cryo-EM de structures hétérogènes.

---

### 3.6.3 Des microscopes plus stables et plus performants

Depuis quelques années, la société FEI/TFS commercialise une série de microscopes de nouvelle génération tels que le Titan Krios et le Talos Artica. Ces microscopes sont très stables mécaniquement et électroniquement et sont capables de faire des acquisitions de qualités automatisées. Parmi les avantages de ces microscopes on retrouve notamment le système de cartouche qui permet de charger jusqu'à 12 échantillons (6 dans le cas du Polara) et de faciliter la récupération, voir même de changer l'échantillon de microscope facilement (ce qui n'était pas possible avec le Polara). La stabilité du microscope permet de faire des acquisitions de plusieurs jours avec seulement quelques corrections quotidiennes, notamment la re-calibration du correcteur Cs et du GIF (filtre d'énergie) s'ils sont présent. Un système de lentilles condenseurs qui permet d'avoir un faisceau parallèle de faible intensité sur une petite zone de 500 nm<sup>2</sup> (Nanoprobe). Pour pouvoir facilement automatiser l'acquisition, il est nécessaire d'utiliser des grilles trouées de manière régulière. L'acquisition en mode particules isolées ou en mode tomographie est possible grâce, soit au logiciel EPU, qui est un logiciel propriétaire fournis par le constructeur, soit au logiciel libre SerialEM (Mastrorade, 2005). Ces deux derniers logiciels sont ceux que j'ai utilisés au cours de ma thèse, mais il existe une dizaine de logiciels d'acquisition différents (Tan et al., 2016).

Il y a actuellement quatre équipements particuliers présents dans le microscope Titan Krios de l'Institut: un Cs correcteur, un filtre d'énergie (GIF) avec la caméra K2 summit, une caméra Falcon III et un volta phase plate (VPP). La correction d'aberration sphérique permet de corriger le défaut

optique transformant l'image d'un point en un petit disque qui impose une limite de résolution. Sa correction permet d'améliorer la résolution, cependant cette amélioration se fait principalement ressentir à des échelles de résolution très petites, inférieures à 2 Å (Hosokawa et al., 2003). Le GIF intégré à la caméra K2 summit modifie le chemin optique et y ajoute un angle droit. Ainsi, de manière similaire à un prisme de verre pour les photons, les électrons sont déviés en fonction de leurs vitesses et il est possible de sélectionner les électrons contribuant à la formation de l'image et donc de faire le tri entre les électrons élastiques et inélastiques (Saunders and Shaw, 2014). Le VPP est une fine plaque de carbone qui est chauffée à environ 200°C par un courant électrique. Lorsque le faisceau électronique traverse cette plaque, il se crée un décalage de phase électronique qui a pour effet d'augmenter le contraste. Il est ainsi possible de voir de petits échantillons, même avec un voltage d'accélération de 300 kV, et il devient même possible de voir ses particules sans appliquer de défocus. Le décalage de phase varie progressivement avec les irradiations répétées de 0° à 180°. Lorsque le décalage de phase est proche de zéro le contraste est faible et lorsqu'il est trop élevé on perd les hautes fréquences. Il faut rester dans des valeurs comprises environ entre 35° et 145° pour une étude biologique. Ainsi on doit changer de position régulièrement la phase plate, lorsqu'une zone est trop irradié (Danev and Baumeister, 2016; Danev et al., 2014).

### 3.6.4 Des échantillons plus reproductibles

Nous avons aujourd'hui des systèmes semi-automatiques permettant une meilleure reproductibilité des échantillons. Ces systèmes sont appelés des plongeurs et sont développés par différents fournisseurs.

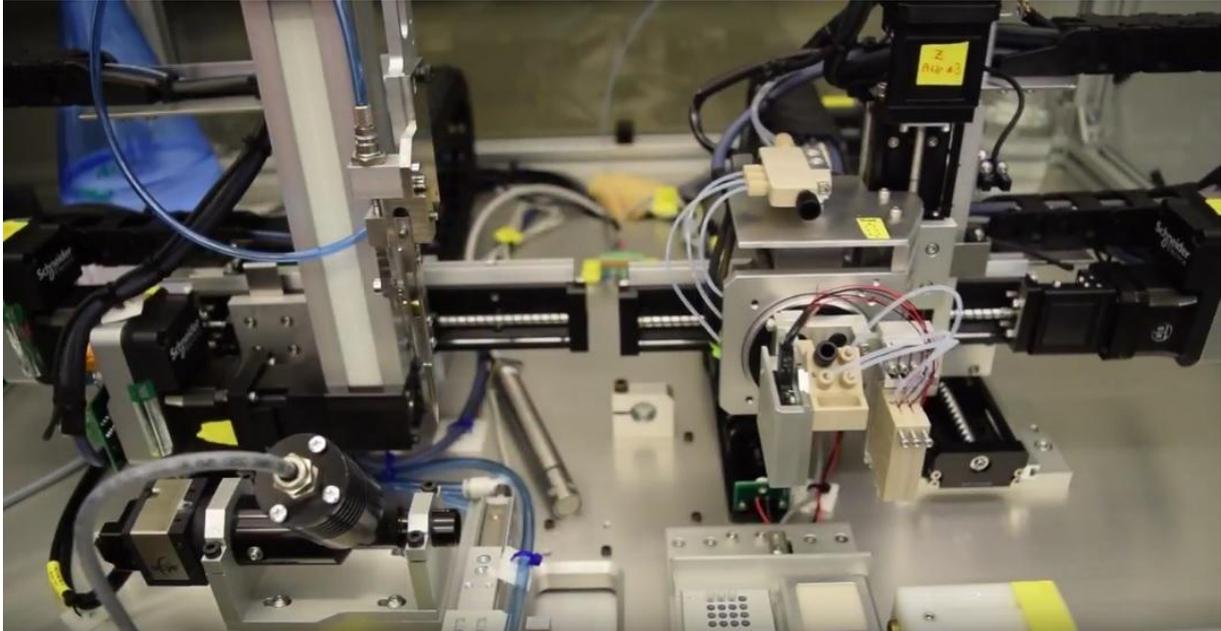


**Figure 50** : Les différents plongeurs actuels. De gauche à droite : Vitrobot Mark IV (FEI), EM GP (Leica), CP3 (Gatan).

Ces systèmes permettent de préparer l'échantillon en le déposant sur la grille (2 à 3  $\mu\text{L}$ ) dans une chambre à température et humidité contrôlées. L'excédent d'échantillon est absorbé sur papier filtre suivant les paramètres de temps et de pression choisis puis plongé dans de l'éthane liquide très rapidement par un bras mécanique.

De très récents développements vont dans le sens de l'automatisation totale, c'est le cas avec le robot dédié Spotiton actuellement développé dans l'équipe de B. Carragher (Dandey et al., 2018). De la même manière que pour un automate semi-automatique où un scientifique fait une partie des étapes, cette machine fait en totale autonomie le dépôt de l'échantillon sur les grilles, le plongeon dans le cryogène et le transfert de la grille dans la boîte de rangement dans l'azote liquide. Cette automatisation permet de travailler avec de très petits volumes, de l'ordre de 2.5 à 16 nL. Le papier filtre est remplacé par un système automatique d'adsorption qui est le résultat du recouvrement des grilles de nanofibres traitées à l'ammonium persulfate qui vont aspirer l'excès d'eau. Une

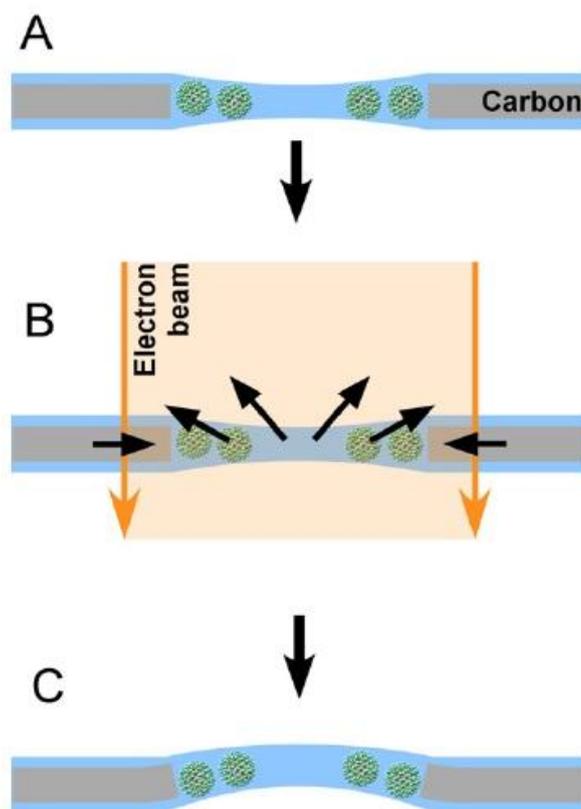
automatisation complète du processus permettrait une forte amélioration de la reproductibilité des grilles en contrôlant plus finement les différents paramètres, notamment temporels.



**Figure 51** : Plongeur Spotiton V1.0. (Dandey et al., 2018).

### 3.6.5 De nouvelles grilles plus performantes

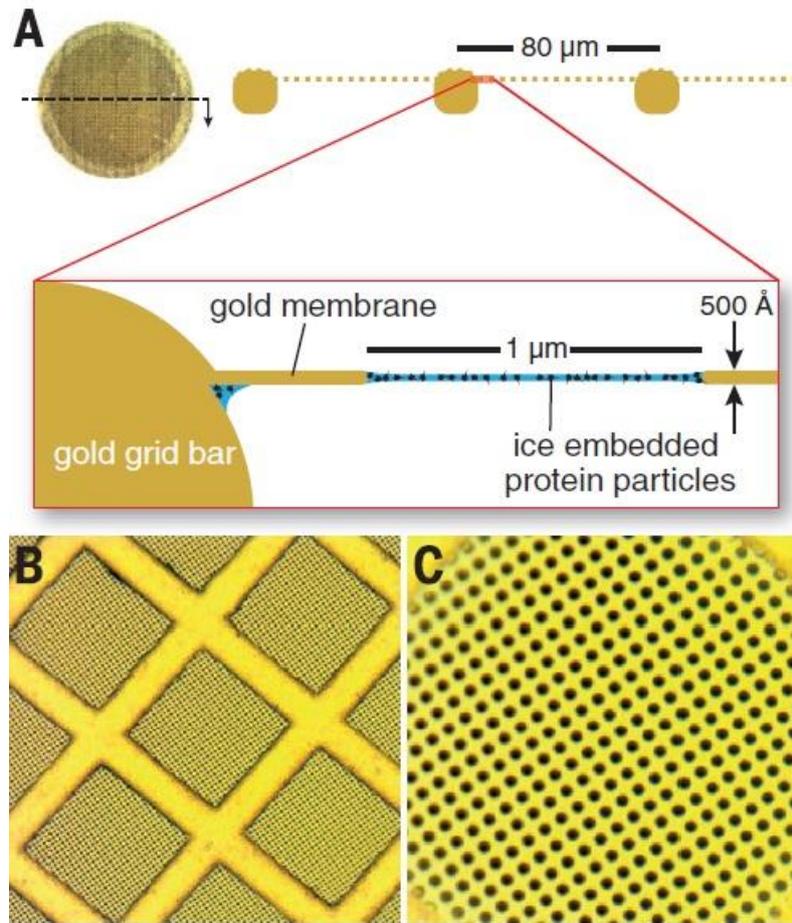
Lorsqu'une acquisition d'image est en cours, les particules peuvent se déplacer. Ce mouvement est dû à plusieurs facteurs, l'énergie absorbée, le fait que la glace amorphe se comporte comme un fluide visqueux. Le carbone de la grille peut se déformer et se contracter sous l'effet des différences de températures, mais il y a également des radicaux libres formés lors de l'exposition qui augmentent la pression dans la couche de glace (Brilot et al., 2012).



**Figure 52** : Mouvements de l'échantillon dû à l'irradiation du faisceau d'électrons. A : Echantillon avant irradiation. B : Durant l'irradiation, il y a un mouvement des particules lié à la déformation du carbone et la formation de radicaux libres. C : Après irradiation, les modifications sont persistantes. (Reproduit de Brilot et al., 2012).

Une exposition longue stabilise la zone, cette observation suggère que les mouvements du carbone sont les plus importants durant la première seconde puis diminuent au fur et à mesure de l'acquisition, ce sont ensuite les radicaux libres qui provoquent des perturbations.

Pour tenter de résoudre en partie ces problèmes, un nouveau type de support entièrement en or a été développé (Russo and Passmore, 2014). Ces grilles sont plus adaptées à la haute résolution. En effet elles permettent de fortement diminuer la déformation de la grille induite par le faisceau d'électrons lors de l'acquisition. Une comparaison a été faite en prenant comme sujet d'étude l'apoferritine, deux acquisitions ont été faites, l'une avec une grille conventionnelle en carbone et l'autre avec une grille en or. L'acquisition faite avec une grille en or montre une réelle progression de la résolution.



**Figure 53** : Grille en or. Le support en or ultrastable comprend un disque de maille d'or de diamètre 3 mm A : Photographie et schémas de la structure d'une grille d'or. B et C: Micrographies optiques d'une grille en or.à moyen grossissement pour la figure B et à fort grossissement pour la figure C. Un trou illustré sur la figure C mesure 1.2μm. (Reproduit de Russo and Passmore, 2014).

D'autres approches existent, comme par exemple la pré-exposition des grilles avec un faisceau électronique directement dans un microscope (équipe Hong Zhou).

# Chapitre 1

## Origine et évolution des récepteurs nucléaires

---

### 1. Introduction

Les récepteurs nucléaires sont apparus dans le clade des métazoaires. Il est couramment admis que HNF-4 est le premier récepteur nucléaire et est présent dès le taxon des Poriferas (spongiaires) qui représente le taxon basal des métazoaires. A partir de leurs apparitions, on les retrouve dans l'ensemble des taxons descendants au sein des métazoaires. On note également l'augmentation de la diversité des récepteurs nucléaires au fil de l'évolution et l'augmentation progressive de la complexité des organismes. La diversification des récepteurs nucléaires a permis de moduler avec plus de précisions et plus de variété l'expression et la régulation des gènes. On peut ainsi augmenter le nombre de voies métaboliques d'un organisme. Les premiers récepteurs nucléaires sont homodimériques et font tous partie des récepteurs nucléaires dit de classe I. Dans un second temps, il y a eu l'émergence de récepteurs nucléaires dit de classe II. Ces derniers forment des hétérodimères exclusivement avec le récepteur nucléaire de classe I RXR. RXR peut néanmoins également former des homodimères en solution. Les deux classes, I et II peuvent être différenciées grâce à des marqueurs de classes. Ce sont des acides aminés, strictement présents dans une classe et strictement absents dans l'autre classe, exceptés pour certains qui sont strictement présents dans les deux classes (Brelivet et al., 2004).

D'un point de vue structural, nous observons que HNF-4 et RXR ont une particularité structurale identique dans la structure secondaire de leur hélice H7. Il s'agit d'un  $\pi$ -turn (tour  $\pi$ ). C'est un tour

d'hélice comportant 5 acides aminés au lieu de 4 pour une hélice  $\alpha$  classique. Cette structure secondaire n'est pas banale et est énergétiquement défavorable. Sa présence et sa conservation stricte suggère donc qu'elle est importante pour le fonctionnement de ces deux récepteurs, d'autant plus que ce sont les deux seuls récepteurs nucléaires ayant cette modification dans le repliement conservé des LBD. Ce  $\pi$ -turn est accompagné d'un motif RxxxE au sein de H7 qui permet son maintien structural. L'observation est d'autant plus surprenante que ces deux récepteurs sont des exceptions, chacun à leur manière. En effet HNF-4 a deux acides aminés marqueurs de classe absent, tandis que RXR est le seul récepteur nucléaire de classe I à former des hétérodimères.

Nous discutons de la capacité différentielle de RXR et HNF4 à hétérodimériser. Nous recherchons également l'origine de cette particularité.

## A Structural Signature Motif Enlightens the Origin and Diversification of Nuclear Receptors

### Abstract

According to phylogenetic analyses, HNF-4 is at the base of the NR family, whereas RXR is the common partner of the large group (subclass) of receptors that function as heterodimers. Both share a unique structural feature, a  $\pi$ -turn or  $\pi$ -bulge, within helix H7, one of the 11 conserved  $\alpha$ -helices of the unique LBD fold. This peculiar secondary structure is further characterized by the conserved signature motif RxxxE. Since HNF-4 and RXR were present in the most ancestral metazoans for which gene sequences are available and since the motif is conserved throughout evolution, we reinvestigated the origin and evolution of the superfamily in the light of sequence and structure conservation of the  $\pi$ -turn. We show that a few other NRs encompass the RxxxE sequence motif, but that only HNF-4 and RXR feature a  $\pi$ -turn conformation at this location. The structural analysis of these receptors suggests that the  $\pi$ -turn is stabilized by conserved intra-molecular interactions at both ends of helix H7, maintaining an otherwise unstable  $\pi$ -helical conformation at the RxxxE location. We further discuss the differential capability of RXR and HNF-4 to heterodimerize. Indeed, both NRs are present in Porifera (sponges) and their biological activity is likely to rely on the homodimeric form. Later during evolution, RXR was selected as the heterodimerization partner, in part because of the presence of the  $\pi$ -turn, but also because of subtle differences in amino acid composition with respect to HNF-4 which favoured complementary interactions between RXR and their NR partners. The use of the  $\pi$ -turn can thus be considered as an exaptation (that is a trait that originate for a given function and subsequently evolved another one) that led to the new crucial NR function of heterodimerization, eventually leading to the complexification in transcriptional response and the tight and specific regulation of gene expression.

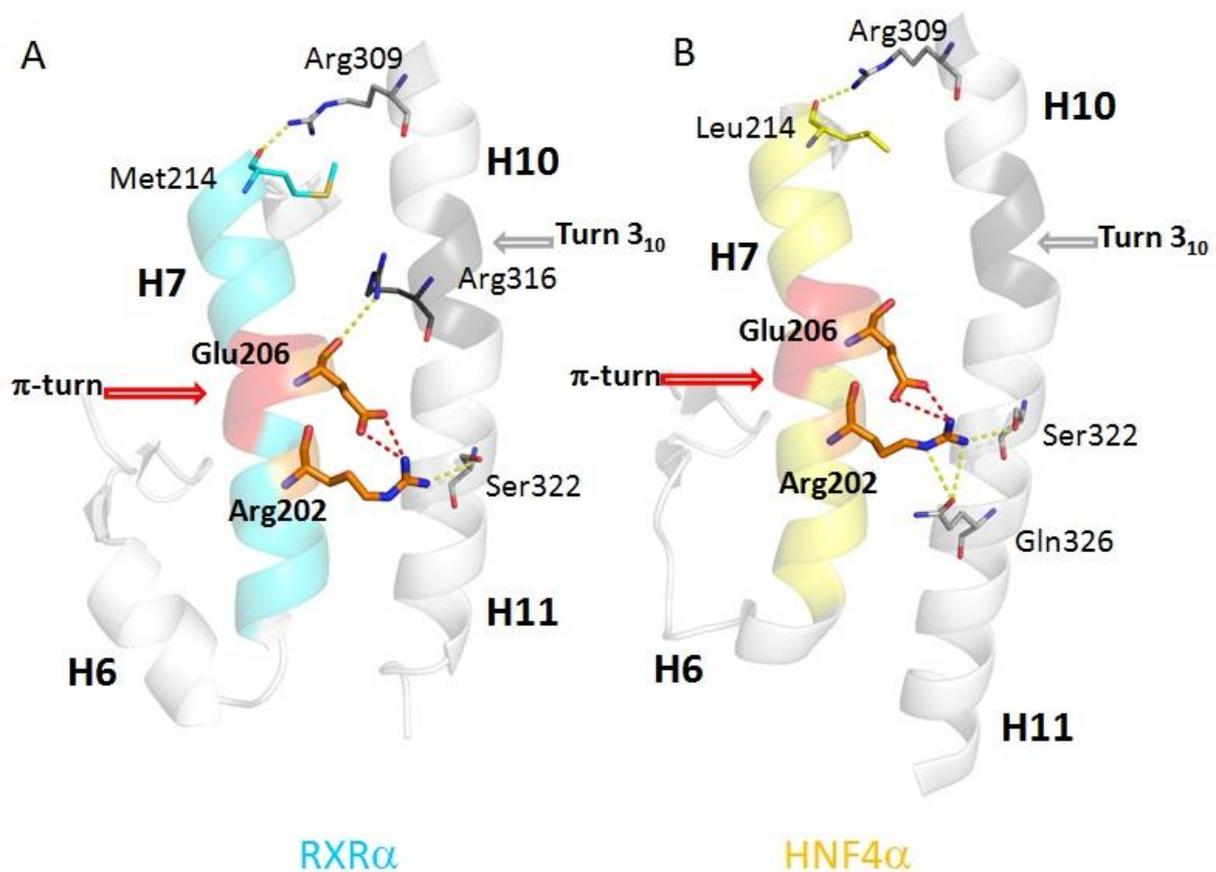
## Results

### Structural analysis

#### A specific structural feature characterizes helix H7 of RXR-USP and HNF-4

A search in the protein data bank reveals that all crystal structures of RXR-USP (some refs) and HNF-4 LBDs (some refs) feature the presence of a  $\pi$ -helix loop insertion in the central part of  $\alpha$ -helix H7 (refs) (see Figure 1). In a  $\pi$ -helical loop, also called  $\pi$ -turn, the N+4 classical hydrogen bonds of the  $\alpha$ -helix are replaced by N+5 hydrogen bonds. A conserved RxxxE motif, where the two invariant residues R and E form an intra-helical salt bridge further characterizes this specific conformation. The  $\pi$ -helical geometry results in the protrusion of the E residue out of the axis of helix H7 with the two polar residues, E and R, closer to the helices H10-H11. Their side-chains form intricate inter- and intra-molecular interactions, stabilizing the *per se* energetically unfavorable  $\pi$ -helical conformation. First, the presence of the glutamate residue allows the formation of an intra-molecular salt bridge with the conserved arginine residue of the motif. Further, the arginine residue helps connecting helix H7 to helices H10-H11, by binding to a conserved serine residue in helix H11 (S322 on the alignment, S427 in RXR $\alpha$ HS). An additional hydrogen-bond is observed between the  $\pi$ -turn and H10-H11, but is not identical in RXR-USP and HNF-4. For RXR-USP, the H-bond is formed between E206 (E352 in hRXR $\alpha$ ) and R316 in H10 (R421 in hRXR $\alpha$ ), as seen in Fig. 1A. For HNF-4 the H-bond is formed between R202 (R267 in h HNF-4 $\alpha$ ) and Q326 in H11 (Q350 in HNF-4 $\alpha$ ), as depicted in Fig. 1B.

The  $\pi$ -bulge induced shift of residues affects only the N-terminal part of H7. The C-terminal side is anchored by conserved bond between residues M/L214 (H7) and R309 (H10). A similar type of interaction pattern prevails for both receptors and leads to strong interactions between H7 and H10-H11. These helices are, together with the loop H8-H9 and H9, the main contributors to the canonical NR LBD dimerization interface. Another interesting observation should be mentioned here: in RXR the contact between main chain of E206 (H7,  $\pi$ -turn), and the side chain of R316 (H10), occurs at a place where the  $\alpha$ -helical conformation of H10 is locally changed to a  $3_{10}$  helix. This peculiar  $3_{10}$  conformation of H10 is observed for all know NRs structures, except for PXR and SF1 that have classical  $\alpha$ -helices (e.g. PXR, PDB: 1ILG, Watkins et al 2001; SF1, PDB: 4QJR, Blind et al 2014).



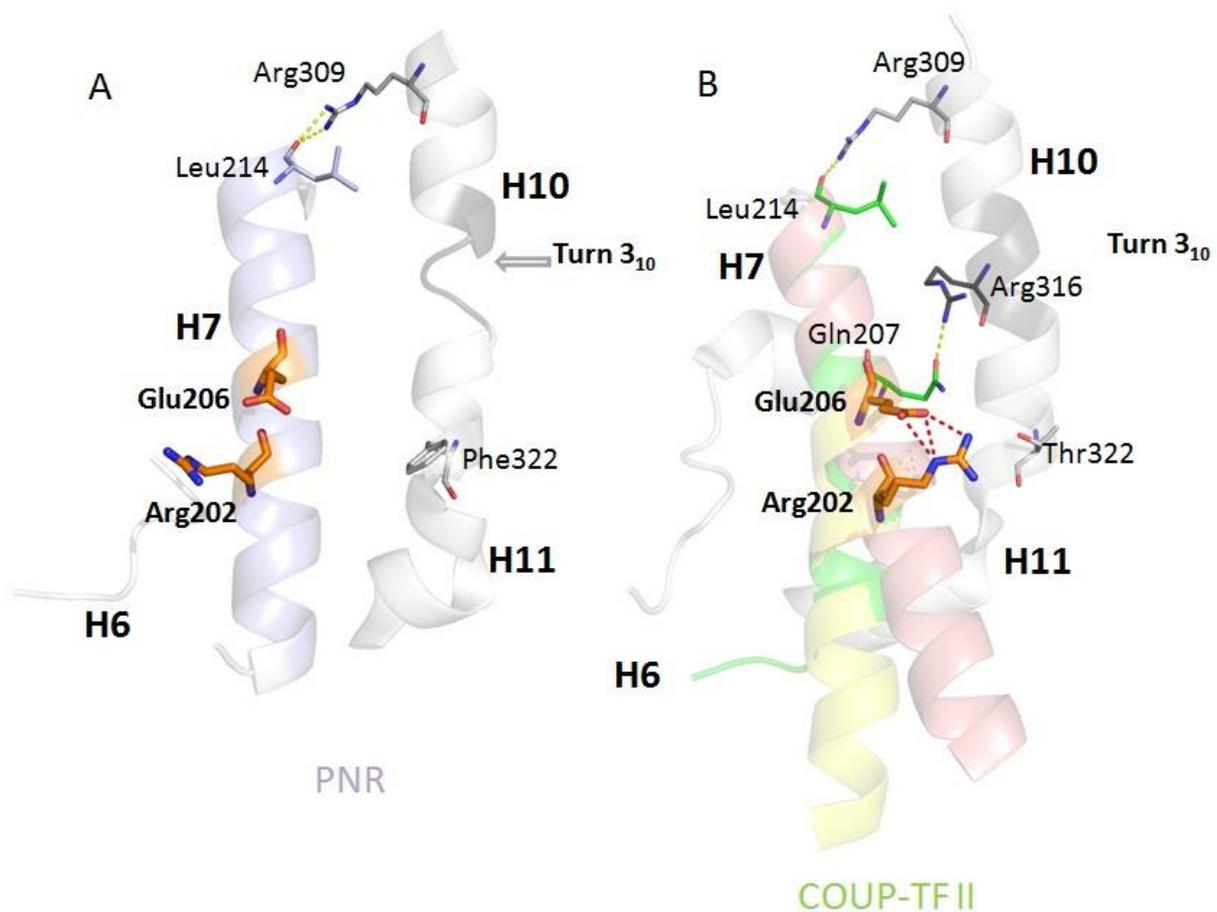
**Figure 1. The environment of the  $\pi$ -bulge in helix H7 in (A) RXR $\alpha$  and (B) HNF-4 $\alpha$  LBDs.** The  $\pi$ -turn is shown in red, the  $3_{10}$  helix at the H10-H11 junction in dark gray. Helix H7 is shown in blue and yellow ribbon representation for (A) RXR $\alpha$  and (B) HNF-4 $\alpha$ . In both cases, the C-terminal part of helix H7 is anchored to the N-terminal part of H10 by a conserved arginine (R309). The side chains of the signature motif residues, R202 and E206, and S322 of H11 form a triad of H-bonds. R316 at the H10-H11 junction in RXR $\alpha$  and Q326 in H11 of HNF-4 $\alpha$  complete the set of conserved bonds. The figures depicted here are based on the PDB structures 1DKF for RXR $\alpha$  and 4IQR for HNF-4 $\alpha$ .

### The RxxxE motif in H7 is not necessarily correlated to the presence of a $\pi$ -helix

To correlate the presence of the RxxxE motif with the occurrence of a  $\pi$ -turn in H7, we carried out a structure-sequence analysis focused on H7. We analysed over 800 crystal NR structures from the protein data bank (PDB) and more than 15,000 protein sequences distributed among 63 different NRs. For 14 NRs, only the sequence was available. For the remaining 49 NRs, at least one crystal structure is known. The RxxxE motif in H7 was found to be present in the NR2F group (COUP-TF, SVP46/7-UP, EAR-2) as well as PNR from the subfamily NR2E. While no crystal structure is available for SVP and EAR-2 LBD, crystal structures were reported for COUP-TFII (PDB ID 3CJW) and PNR (PDB

ID4LOG). In neither of these two structures is a  $\pi$ -turn conformation seen nor a salt bridge between R and E residues of the motif.

In PNR LBD, H7 exhibits a canonical  $\alpha$ -helical conformation with no visible distortions, as shown in Fig. 2A. No intra-molecular interactions are seen between the motif residues R and E. The serine residue observed in RXR H11 (S322) that is important for the stability of the  $\pi$ -turn is replaced by F322 in PNR. This residue would generate a steric clash with a  $\pi$ -turn conformer. Due to the absence of  $\pi$ -turn, the interaction network between H7 and neighboring regions of the receptor is weaker. The N-terminal part of H7 interacts weakly with H11 through a bond between E200 and R327, but not with H6 (not visible on the electron density map), whereas the C-terminal end of H7 interacts with residues at the C-terminal of H5. If a  $\pi$ -helix would be present in PNR, the offset induced by the bulge would change the position of E200 which would point into the direction of H5-H6, more specifically in a hydrophobic region composed of several leucine residues that would not favour interaction (Fig. 2A).



**Figure 2.** The environment of the  $\pi$ -turn for (A) PNR and (B) COUP-TFII. The  $\pi$ -turn motif residues (R202 and E206) are shown in orange, the  $3_{10}$  turn at the H10-H11 junction is shown in dark gray. In both cases, the C-terminal part of H7 is held

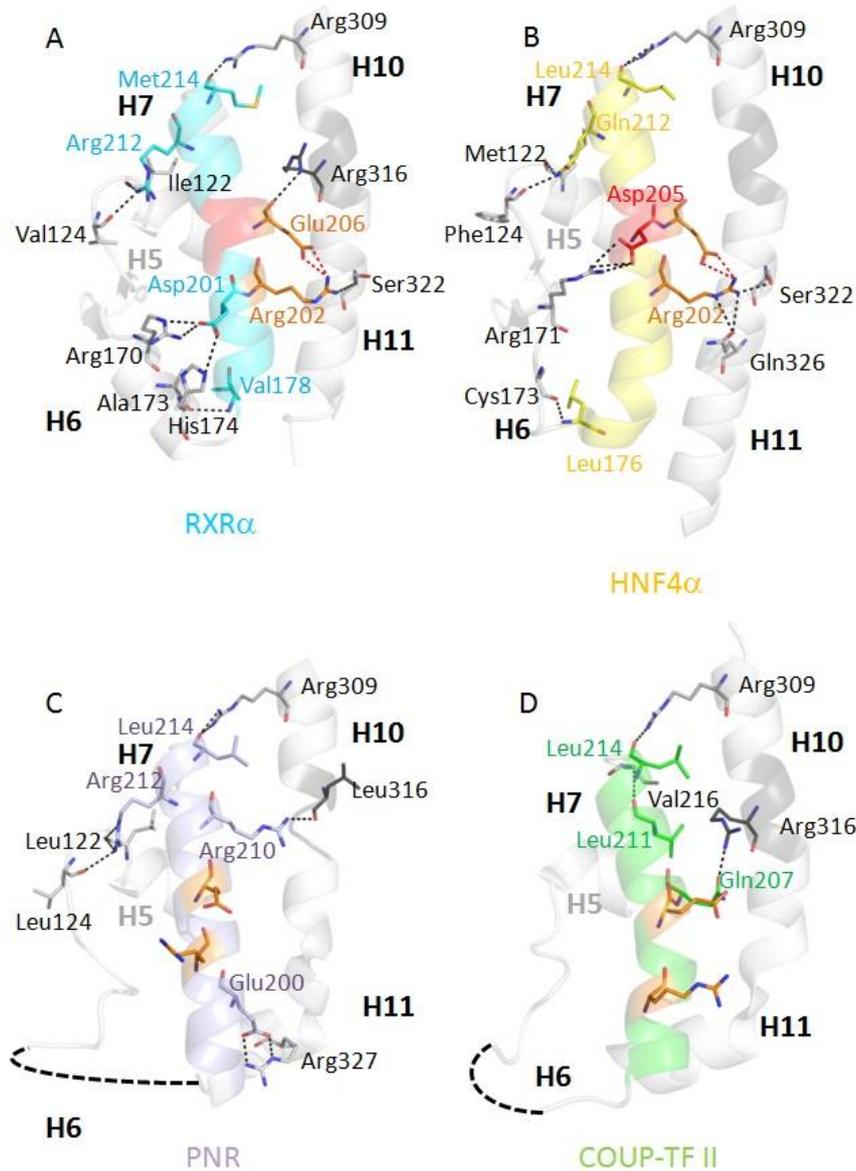
by arginine 309 at the N-terminus of H10, as observed in RXR $\alpha$  and HNF-4 $\alpha$ . For (A) PNR and (B) COUP-TFII, H11 is unstructured and the amino acid shift due to the absence of  $\pi$ -turn only affects the N-terminus of H7, while no major changes occur at the C-terminal part, in line with the structural alignment of the corresponding residues. In (A), PNR is depicted with H7 in light violet. No contacts between H7 and H10-H11 are seen only at the C-terminus of H7. A phenylalanine residue (F322) in H11 replaces S322 of RXR H11 that interacts with the  $\pi$ -turn. In (B), the original conformation of COUP-TFII is shown in green, the re-refined H7 conformers (see S.I.) are shown in salmon and yellow for the straight and curved helical conformation, respectively. A threonine residue (T322) in H11 replaces S322 of RXR H11 that interacts with the  $\pi$ -turn. The figures depicted here are based on the PDB structure 4LOG for PNR and on the PDB structure 3CJW and our re-refined structure for COUP-TFII.

A careful analysis of the crystal structure of COUP-TFII and its corresponding electron density map reveals that large portions of the electron density in the region of H7 could not be interpreted. The main problems are located between H5 and H7, with the absent  $\beta$ -sheet, H6 with no visible electronic density and a disordered Nter of H7. For the latter a closer inspection to the electron density map led to the hypothesis that the N-terminal part of H7 could adopt different conformations. Since this region was critical for our analysis of the RxxxE motif and the structural features associated to it, we decided to further improve the protein structure around this location by iterative building of residues in the non-interpreted electron density map followed by a crystallographic refinement using Phenix (See Suppl. Information). This work resulted into better crystallographic quality factors R and Rfree and to a more confident interpretation of the electron density map with two main observations.

After crystallographic re-refinement, we observed in the crystal packing helix H7 can adopt two helical structures, as shown in Fig. 2B as a salmon and a yellow ribbon, together with a lengthening of helix H7 at its N-terminal side as compared to the original helix of the PDB structure (green ribbon in Fig. 2B). The two novel conformations seen in the re-refined structure (salmon and yellow ribbons in Fig. 2B) correspond to a regular straight and a curved  $\alpha$ -helix that is bent at the level of the  $\pi$ -turn. The C-terminal parts of the two helices overlap nicely, while their N-terminal ends are located over 6Å apart. These conformations are in equilibrium in the protein crystal, alternating between nearest neighbour molecules to ensure optimal packing and are likely to be the natural conformations. The dynamics of H7 resulting from the absence of a stabilizing H11 promotes the adaptability to packing constraints with a subsequent disorder of this subdomain. In fact, the lengthening of the original single helix H7 to the size of the re-refined one would lead to steric clashes between crystallographic dimers (See Sup Mat).

The second important observation is the absence of the  $\pi$ -turn, even though the intra-helical salt bridge between the side chains of the arginine and the glutamic acid of the motif is maintained as suggested by the electronic density (see Suppl. data). However, this intra-helical salt bridge is rotated to a position such that no interaction between the motif and H10-H11 can take place. The shift induced by the absence of the  $\pi$ -turn prevents E206 from binding R316, instead the connection is made with its neighboring residue Q207 (Q298 in hCOUP-TFII). The conserved serine residue of RXR H11 that stabilizes the  $\pi$ -turn in RXR-USP and HNF-4 is replaced by a threonine residue, but without interacting with H7 residues (Fig 2B). Furthermore, no interactions are seen between H7 and H5-H6. Of note, helix H7 after refinement does not exhibit a  $3_{10}$  helical turns as suggested in the original structure.

Altogether, the structure-sequence analysis of the NRs that exhibit a motif RxxxE in H7 indicates that RXR-USP and HNF-4, and only them, possess a peculiar  $\pi$ -helical geometry. Strong interactions between both the N- and C-terminal parts of H7 and neighboring regions of the receptor are seen to hold together the peculiar H7 conformation and help maintaining an otherwise intrinsically unstable  $\pi$ -helical secondary structure conformation. This contrasts with COUP-TF and PNR, where no  $\pi$ -turn is observed despite the presence of the RxxxE motif, in line with a much scarcer and weaker interaction network observed between H7 and neighbouring regions of the receptor. The peculiar topological  $\pi$ -helical feature seen in RXR-USP and HNF-4 allows intricate interactions with helices H10-H11 which are crucial elements of the dimerization interface.



**Figure 3. The interaction network of helix H7 and the  $\pi$ -turn with surrounding structural elements of the receptor.** The environment of H7 and the  $\pi$ -turn in (A) RXR $\alpha$ , (B) HNF-4 $\alpha$ , (C) PNR and (D) COUP-TFII LBDs. Helix H7 is shown as a blue (A), yellow (B), violet (C) and green (D) ribbon representation. The  $\pi$ -turn is highlighted in light orange in red and the  $3_{10}$  helix at the H10-H11 junction in dark gray. The side chains of the RxxE signature motif residues, R202 and E206, are shown in a stick representation with carbon atoms in orange and the H-bonds connecting them as red dashed line. H-bonds between residues of H7, including the  $\pi$ -turn, are shown by black dashed lines. Notice the lack of interactions between H7 and H5-H6 for PNR (C) and COUP-TFII (D) compared to what is observed for RXR $\alpha$  (A) and HNF-4 $\alpha$  (B), consistent with the flexibility of these elements that are poorly visible in the electron density. The figures depicted here are based on the PDB structures 1DKF for RXR, 4IQR for HNF-4 $\alpha$ , 4LOG for PNR and 3CJW for COUP-TFII.

## RXR-USP and HNF-4 are distinct in the conservation of the class I specific residues

Previous published work identified specific residues which are strictly conserved in some NRs and strictly absent in other receptors (Brelivet et al, 2003). This led to the classification of NRs into two classes, that eventually correspond to homodimeric and monomeric NRs for class I and to NRs that form heterodimers with RXR for class II. For class I NRs, the class specific residues define an interaction pattern that links H1 to H8 and further H8 to H10, and can be thought as linking the ligand binding pocket to the dimerization interface. RXR-USP, COUP-TFII, PNR and HNF-4 belong to class I NRs. However, careful examination of class specific residues indicates that in contrast to the three first aforementioned NRs, HNF-4 is an outlier. In fact, two class I invariant residues, W109 (W40) and R321 (R105), which are important for our subsequent evolutionary analysis, are not conserved in HNF-4 (Table1). W40 is located at the junction of H4-H5, which is a highly conserved structural feature of the NRs family and interaction point with ligands for all LBDs. It is thought to be involved as a ligand sensor (Brelivet et al 2003) and in the ligand-dependent allosteric mechanism in RXR-TR (Kojetin et al, ncomms 2015). R321 (R426 in hRXR $\alpha$ ) is an important residue at the dimerization interface and is mainly conserved for all nuclear receptors. This class I invariant residue in H10 is replaced by a glutamine residue (Q345 in h HNF-4 $\alpha$ ) which interacts with Q321 of the other subunit.

Secondary structure	H4-H5	H5	H7					H8	loop H8-H9			H10			H11	
Class specific	I (W)	II (E,D)	$\pi$ -turn					I ; II (E)		II (R)	I (R)		I ; II (R)			
Alignment residue	109	111	202	206	207	210	214	220	262	263	309	316	321	322	326	
Brelivet numbering	40	42	-	-	-	-	-	50	61	62	93	100	105	106	-	
hRXR $\alpha$	W305	E307	R348	E352	L353	K356	M360	E366	D379	S380	R414	R421	R426	S427	K431	
hHNF4 $\alpha$	A224	E226	R267	E271	L272	P275	L279	E285	D298	A299	R333	L340	Q345	S346	Q350	
hCOUP-TFII	W249	E251	R293	E297	Q298	K301	L305	E311	D324	A325	R359	R366	R371	T372	S376	
hPNR	W257	E259	R301	E305	T306	R309	L313	E319	E332	T333	R367	L374	R379	F380	E384	
hRAR $\alpha$	C265	D267	D307	A311	F312	Q315	L319	E325	D338	R339	M373	K380	R385	S386	K390	
hTR $\beta$	C309	E311	D351	D355	L356	S359	F363	E369	D382	R383	F417	K424	R429	M430	C434	

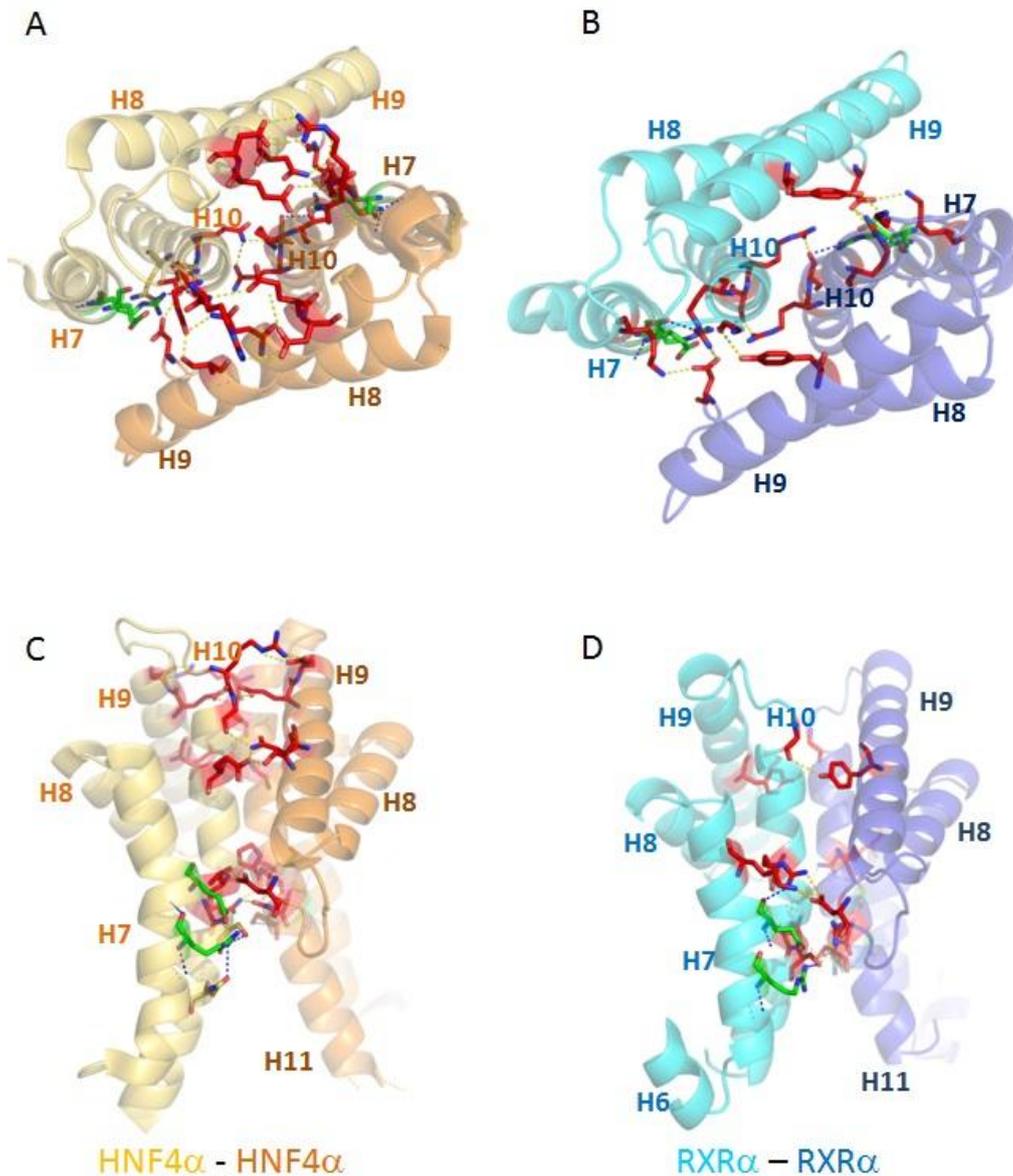
**Table 1. Amino acid residue mapping for the specific nuclear receptors considered in this study.** Secondary structure elements are specified by H for the helices. Alignment residue is the generic numbering used in this study. The class specific residues are specified by I and by II for class I and class II and the boxes are colored in blue and green, respectively, together with the corresponding residue numbering (ref Brelivet). The H7 column with yellow boxes specifies residues of the  $\pi$ -turn.

## H7 and the dimer interfaces, functional correlations

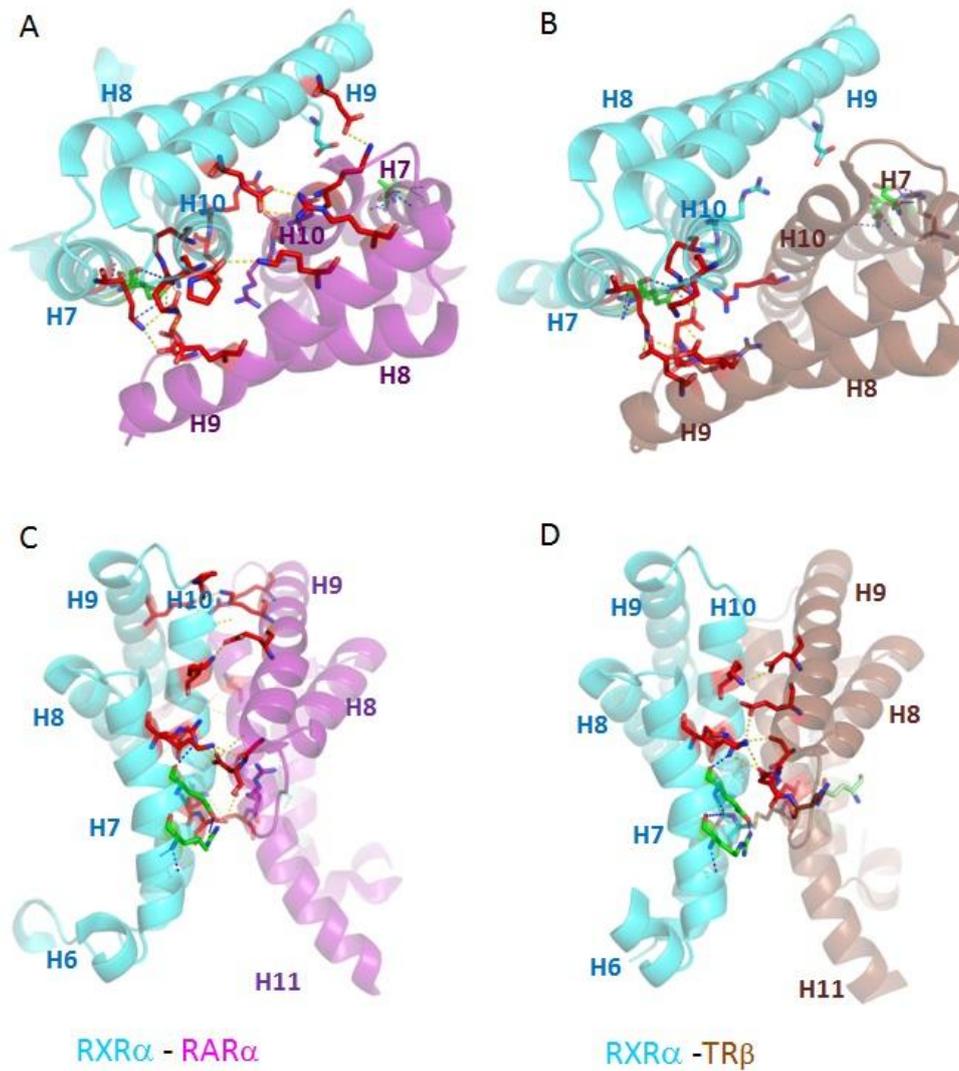
The different bonds involving  $\pi$ -turn residues are important for dimerization (refs). Functional studies reported by Eeckhoutte and coworkers on HNF-4 indicated that the mutation of the conserved glutamic acid and arginine of the motif led to the disruption of the dimers (2003 Eeckhoutte).

A comparative analysis of the crystal structures of the homodimer of HNF-4 $\alpha$  with that of RXR $\alpha$  indicates that the majority of the contacts at the dimer interface are different (see Figure 4). Differences are in the nature of amino acid residues at the interface as well as in their polarity. For example, in RXR $\alpha$ , K312 (K417 in hRXR $\alpha$ ) and R316 (R421 in hRXR $\alpha$ ) in H10 are found to be E312 (E336 in h HNF-4 $\alpha$ ) and L316 (L340 in h HNF-4 $\alpha$ ) in HNF-4 $\alpha$ , respectively. Moreover, for HNF-4 $\alpha$ , a large contribution to the stability of the dimer interface of HNF-4 $\alpha$ , comes from the stacking of the tryptophan residue W325 (W349 in h HNF-4 $\alpha$ ) in H10 from each of the two subunits. This large residue is instead a leucine residue L325 (L430 in hRXR $\alpha$ ) in RXR $\alpha$ . With respect to the  $\pi$ -turn of HNF-4 $\alpha$ , R202 is not involved in the dimerization interface, whereas E206 that makes a weak intramolecular bond to T319, and forms an inter-molecular stacking interaction to D262 (D298 in h HNF-4 $\alpha$ ) in the loop H8-H9 of the other subunit. Strikingly, a subsequent interface interaction network can be seen in HNF-4 $\alpha$  between H9 and H10 of the other protomer and between the H10 helices of both subunits (Figs 4A-B). These interactions represent the major interface contacts in the HNF-4 $\alpha$  homodimer and strongly help stabilizing the latter. Much less contacts can be seen in RXR $\alpha$  homodimer (see Figs 4C-D), suggesting a less stable complex, consistent with biophysical and structural data (refs). Nevertheless, the  $\pi$ -turn of RXR $\alpha$  is crucial in the stabilization of the homodimer, generating an intricate network of interactions. In fact, in contrast to HNF-4 $\alpha$ , R202 and E206 of RXR $\alpha$  are both involved in the dimerization interface, and link together S322 (H11) of the same monomer to R321 (H10) of the other protomer. This arginine residue is peculiar, since it is an invariant of the whole NR family but absent in HNF-4 $\alpha$ . It is further involved together with K210, located at the C-ter of the  $\pi$ -turn, in polar interactions to the aspartic residue of the loop L8-9 of the other protomer. This D residue is the same as that seen to contribute to the HNF-4 $\alpha$  homodimer interface and is present in most class I and class II NRs.

In contrast to HNF-4 $\alpha$ , RXR $\alpha$  can form heterodimers with NR partners, such as RAR $\alpha$ , TR $\alpha$ , or VDR. In all cases, the heterodimers are distinct from the homodimer by several features (Figure 5). First, from the topological point of view, the heterodimer is asymmetric in the relative positioning of the subunits, with RXR H7 helix being closer from the loop H8-H9 of the partner than the reverse. Second, an intricate network of complementary interactions can be observed at the interface between RXR and its partner that spans the entire dimer interface, with interactions between the elements H7, loop H8-H9, H9 and H10 of both subunits. This is in line with the observation that the heterodimers are more stable than RXR $\alpha$  homodimers, as shown for the heterodimer RXR $\alpha$ /RAR $\alpha$  (ref). The consequence of the asymmetry in the interface directly impacts on the number and types of interactions on both ends of the interface (with respect to the helices H10-11). For example H7 of one of the protomers interacts with the loop L8-9 of the other protomers, but not with the same strength. Taking RXR $\alpha$ /RAR $\alpha$  (PDB code 1DKF) as an example, only very few bonds are seen between the loop H8-H9 of RXR $\alpha$  and H7 of RAR $\alpha$ , in particular a single and weak H-bond is observed between D262 (loop H8-H9 of RXR $\alpha$ ) and Q210 (H7 of RAR $\alpha$ ) REF Bourguet 2000. However, when looking at the interactions between H7 of RXR $\alpha$  and the loop H8-H9 of RAR $\alpha$ , more interactions, including H-bonds, can be observed which in particular involve the  $\pi$ -turn residues R202, T205, E206 and the residue K210 at the C-ter of the  $\pi$ -turn. These residues interact in a direct or a water-mediated manner to D262, Q3264, D265 in the loop H8-H9 of RAR  $\alpha$ . Another illustration of the heterodimeric situation is provided by the crystal structure of RXR $\alpha$ /TR $\alpha$  (PDB code 4ZO1) with similar features observed with some variations around the themes and with the strong involvement of residues of the  $\pi$ -turn (see Figure 5B-D). In both cases, an interesting observation can be made. The loop H8-H9 contains an arginine residue next to the aspartic acid residue D262 (R263) which is a class II signature residue found in all class II and strictly absent in all class I NRs. This residue makes a conserved H-bond to E111 in H5, which itself is also a class II conserved residue and present in most class I NRs. This strongly suggests that an interacting network is created between the  $\pi$ -turn of RXR, and the loop H8-H9 and the helices H4-H5 (and thus the pocket) of the partner NR. This is consistent with previous report that indicates that the salt bridge between E(or/D)111 in H5 and R263 in the loop H8-H9 favor heterodimeric interactions (Brelivet et al 2004). Altogether, our analysis suggest that the  $\pi$ -turn of RXR plays a key role in the heterodimerization mechanism by creating a strong interaction networks with the partner loop H8-H9 and helices H9 and H10 and favoring heterodimer rather than homodimer formation.



**Figure 4.** The dimerization interfaces of (A, B) HNF-4 $\alpha$  homodimers and (C, D) RXR $\alpha$  homodimers are not similar. Two different views are depicted for both receptors. Amino acid residues that are involved in dimerization through hydrogen bonds are shown in red. Amino acid residues 202 to 206 of the  $\pi$ -turn are shown in green. The major region of dimerization interface for RXR $\alpha$  corresponds to H10 helices of both subunits, whereas for HNF-4 $\alpha$  it comprises H10 of one subunit and H9 of the other subunit. More contacts are seen at the interface for homodimeric HNF-4 $\alpha$  compared to homodimeric RXR $\alpha$ . The figures depicted here are based on the PDB structures 1LBD for RXR $\alpha$  and 4IQR for HNF-4 $\alpha$ .



**Figure 5.** The dimerization interfaces of (A,B) RXRα-RARα and (C, D) RXRα-TRβ heterodimers are asymmetric. Two different views are depicted for both heterodimers. Amino acid residues that are involved in dimerization through hydrogen bonds are shown in red. Amino acid residues 202 to 206 of the  $\pi$ -turn are shown in green. RXRα makes an intricate network of complementary interactions with its partners, resulting in better dimer stability compared to its homodimeric form. The figures depicted here are based on the PDB structures 4ZO1 for RXRα-TRβ and 1DKF for RXRα-RARα.

## Evolutionary Considerations

	Residues Class I						Other Residues	
	E5	W40 (305)	E50	KR55	R93	R105	W16 (282)	$\pi$ -turn
HNF4 $\alpha$	E	A	E	K	R	Q	W	RILDE
HNF4 $\gamma$	E	A	E	K	R	Q	W	RILDE
HNF4 $\beta$	E	A	E	K	R	Q	W	RVLDE
RXR $\alpha$	E	W	E	R	R	R	W	RVLTE
RXR $\beta$	E	W	E	R	R	R	W	RVLTE
RXR $\gamma$	E	W	E	R	R	R	W	RVLTE
USP	E	W	E	K	R	R	W/y	RVLSE
COUP-TFI	E	W	E	K	R	R	W	RIFQE
COUP-TFII	E	W	E	K	R	R	W	RIFQE
NR1_AmpME		W	E	K	R	R	W	RVSHE
NR2_AmpMe		V	E	K	R	Y	W	VTLTK

**Table 2** : Class I marker and  $\pi$ -turn residue.

The origin of nuclear receptors is commonly linked to the Porifera taxon (sponges).

Two nuclear receptors have been identified in sponges, NR1 and NR2. It is thought that these receptors are ancestral forms of HNF-4 and that the first nuclear receptor was an HNF-4 that subsequently duplicated and diverged (Bridgham et al 2010). Interestingly, if we focus our analysis on a restrained part of the sequence that encompasses the class I markers reported by Brelivet and coworkers (2004), we note that HNF-4 can clearly be differentiated from RXR and the rest of class I NRs.

Analysis of the available sequences of NR1 and NR2 from *Amphimedon queenslandica* shows that NR1 possesses all of the class I markers, while NR2 exhibits HNF-4-like features (table 2). W40 is replaced by V40 and R105 is replaced by Y105. Note that in both HNF-4 and NR2, W40 is replaced by smaller non-polar amino acid residues, such as alanine in the case of HNF-4 and valine in the case of NR2. Arginine 105 (R105), a positively charged polar amino acid residue, is replaced by Glutamine (Q105) for HNF-4 and Tyrosine (Y105) in NR2, both being uncharged polar amino acids. This would

suggest that NR2 is an ancestral sequence of HNF-4, whereas NR1 is an ancestral sequence of a canonical class I, RXR-like nuclear receptor. According to this view the RXR/ HNF-4 split would be a fundamental dichotomy at the origin of the NR superfamily.

NR1 and NR2 are the only two nuclear receptors known in Porifera, they are paralogous and therefore very similar. We cannot determine which one is the first but the present analysis would suggest a plausible scenario for the evolution of the NRs family. Considering the overall degree of sequence identity, HNF-4 is the most similar to both NR1 and NR2. Other class I nuclear receptors, in particular the family of COUP-TF (NR2F), PNR (NR2E) and RXR (NR2B) are also very close to NR1, their most probable common ancestor. NR1 has progressively moved away from HNF-4 to form the class I nuclear receptors known today.

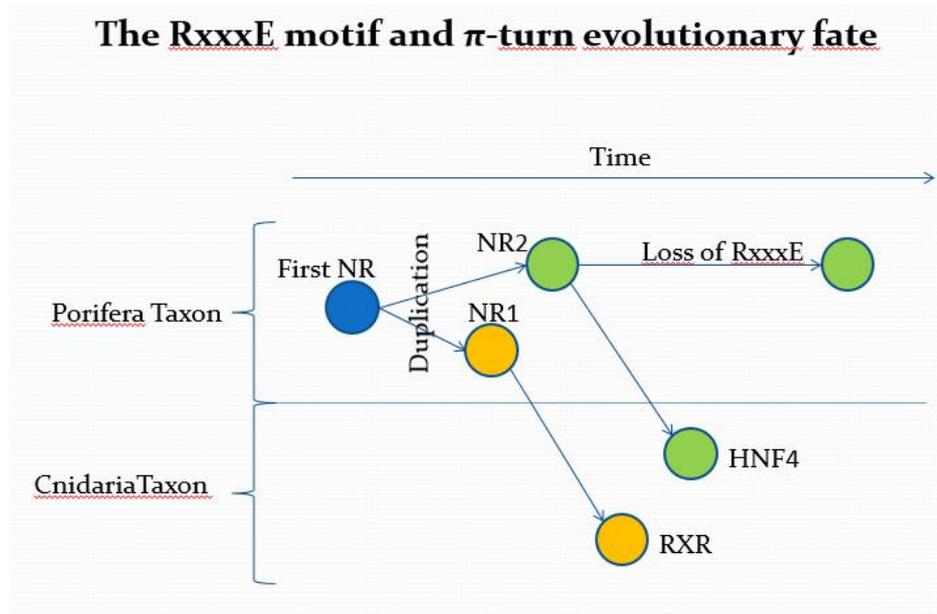
			Cnidaria			Mammalia (human)						
ACA04755.1 <i>Amphimedon</i>	NR2A(NR1)	HnF4a	57,05%	59,30%	58,89%	50,92%	56,85%	57,05%	56,03%	57,46%	50,51%	49,89%
P_001266240.1 <i>Amphimedon</i>	NR2A(NR1)	HnF4a	57,05%	59,30%	58,89%	50,92%	56,85%	57,05%	56,03%	57,46%	50,51%	49,89%
ADK78987.1 <i>Amphimedon</i>	NR2A(NR2)	HnF4a	41,51%	44,17%	43,96%	39,67%	40,49%	40,89%	39,87%	45,19%	38,65%	38,85%
P_001266221.1 <i>Amphimedon</i>	NR2A(NR2)	HnF4a	41,51%	44,17%	43,96%	39,67%	40,49%	40,89%	39,87%	45,19%	38,65%	38,85%
			XP_015763968.1_A croptora	XP_015751727.1_A croptora	XP_015751728.1_A croptora	Q9YSX4_Homo	P10589_Homo	P24468_Homo	P10588_Homo	P41235_Homo	P19793_Homo	P48443_Homo
						COUP-TF I	HnF4a	HnF4a	PNR	COUP-TF I	COUP-TF II	EARF-2
						NR2F	NR2A	NR2A	NR2E3	NR2F1	NR2F2	NR2F6
												NR2A1
												NR2B1
												NR2B3

Figure 6 : Percentage of sequence identity.

The RxxxE motif is present in the sequence of Porifera NR1 at the same location that the  $\pi$ -turn in HNF-4 and RXR. In absence of experimental molecular structures, the sequence RVSHE alone does not allow to predict of the presence of a  $\pi$ -turn. Nevertheless one can notice the absence at the position 4 of a glutamine residue that would disturb the formation of a  $\pi$ -turn.

In other Porifera species such as *Syconciatum* (RLVDE) NR1 sequences are very similar to those of both RXRa (RVLTE) and HNF-4 (RILDE), but surprisingly the RxxxE pattern is not present in H7 helix of NR2. Indeed in *Amphimedon queenslandica*, the Arginine (R) is replaced by a Valine and Glutamate





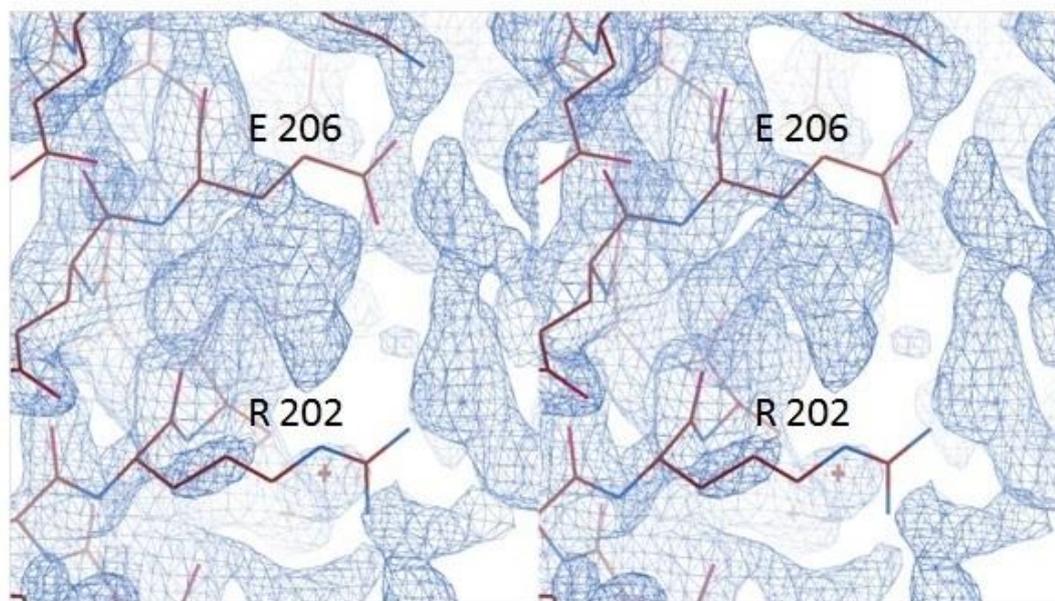
## Discussion Bullet points

- $\pi$ -helices are found in 15% of the protein structures
- $\pi$ -helices are the same as structures known as  $\alpha$ -bulges,  $\pi$ -bulges, and looping outs, and are evolutionarily derived by the insertion of a single residue into an  $\alpha$ -helix. Their evolutionary origin explains both why  $\pi$ -helices are cryptic, being rarely annotated despite occurring in 15% of known proteins, and why they tend to be associated with function.
- Role of H5-H6 (anchoring points for H7) ; C-ter stabilized for all NRs and N-ter not
- Pi-turn stabilization in HNF-4 and RXR compared to other NRs
- Functional implications of pi-turn
- Evolution of heterodimerisation and the existence & stabilization of pi-turn!
- Insertion or deletion of pi-turn?
- Hetero- vs Homo: asymmetric vs symmetric ; complementarity H9 and H10;
- RXR as heterodim partner not HNF-4; structural reasons
- The NRs which encompass the RxxxE motif are orphan NRs ( HNF-4 and RXR, PNR and COUP-TF); ERR then appeared without RxxxE then later SRs
- Orphan character of ancestral NRs is consistent with absence of circulatory system in sponges which are the most ancestral organisms of metazoan taxon
- Based on our analysis, cannot clearly state that HNF-4 is the first NR, in sponges 2 NRs one more RXR like one more HNF-4 like which has further evolved in sponges

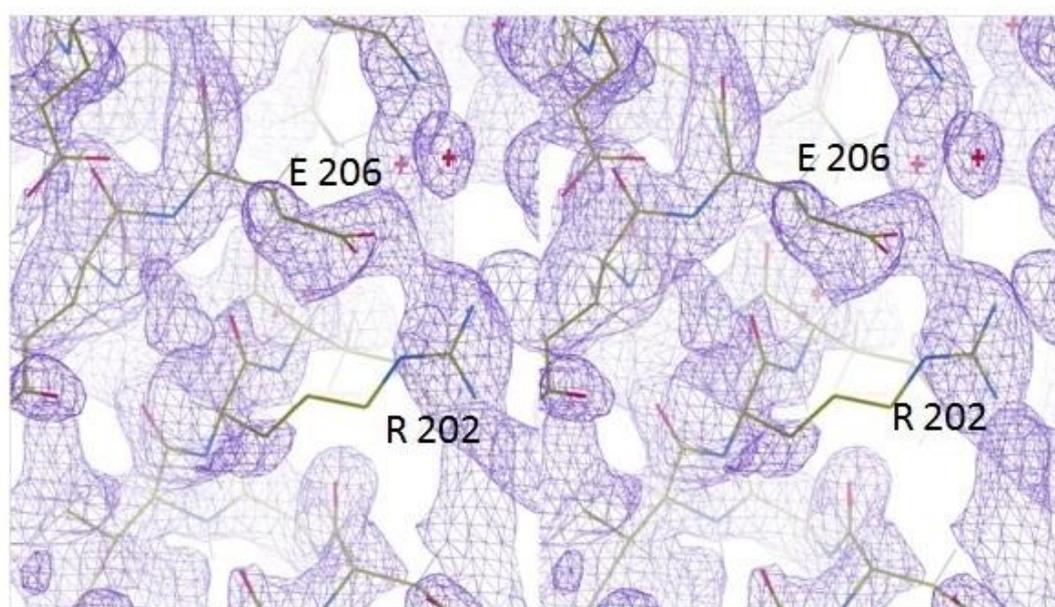
## **Figures Sup Mat**

	Original	Rerefined
Wavelength		
Resolution range	10.0 - 1.48 (1.53 - 1.48) 6 - 1.48 (1.53 - 1.48)	29.16 - 1.481 (1.534 - 1.481)
Space group	C 1 2 1	C 1 2 1
Unit cell	97.85 47.76 43.13	97.85 47.76 43.13
	90 100.87 90	90 100.87 90
Total reflections	238208	
Unique reflections	32366 31940 (2991)	32452 (3056)
Multiplicity		
Completeness (%)	98.8 (91.7) 97.70 (93.53)	99.37 (94.50)
Mean I/sigma(I)	35.3 (2.1)	
Wilson B-factor	24,02	24,17
R-merge (%)	4.7 (73.6)	
Reflections used in refinement	31940 (2991)	32452 (3056)
Reflections used for R-free	1575 (142)	1603 (147)
R-work	0.168 (0.180) 0.168 (0.288)	0.1470 (0.2204)
R-free	0.238 (0.302) 0.238 (0.286)	0.1825 (0.2967)
Number of non-hydrogen atoms	1636	2245
macromolecules	1636	2104
ligands		19
solvent		122
Protein residues	207	217
RMS(bonds)	0,037	0,018
RMS(angles)	2,9	1,64
Ramachandran favored (%)	93,03	97,1
Ramachandran allowed (%)	2,49	2,9
Ramachandran outliers (%)	4,48	0
Rotamer outliers (%)	9,73	4,88
Clashscore	26,17	16,56
Average B-factor	34,49	35,09
macromolecules	34,49	34,23
ligands		53,07
solvent		47,24

Table 1 : Original and re-refined statistics. For the original structure, the red values are the original values of the publication, the bleu values are the value calculated by Phenix. When the values are black, it's the same values in both cases.

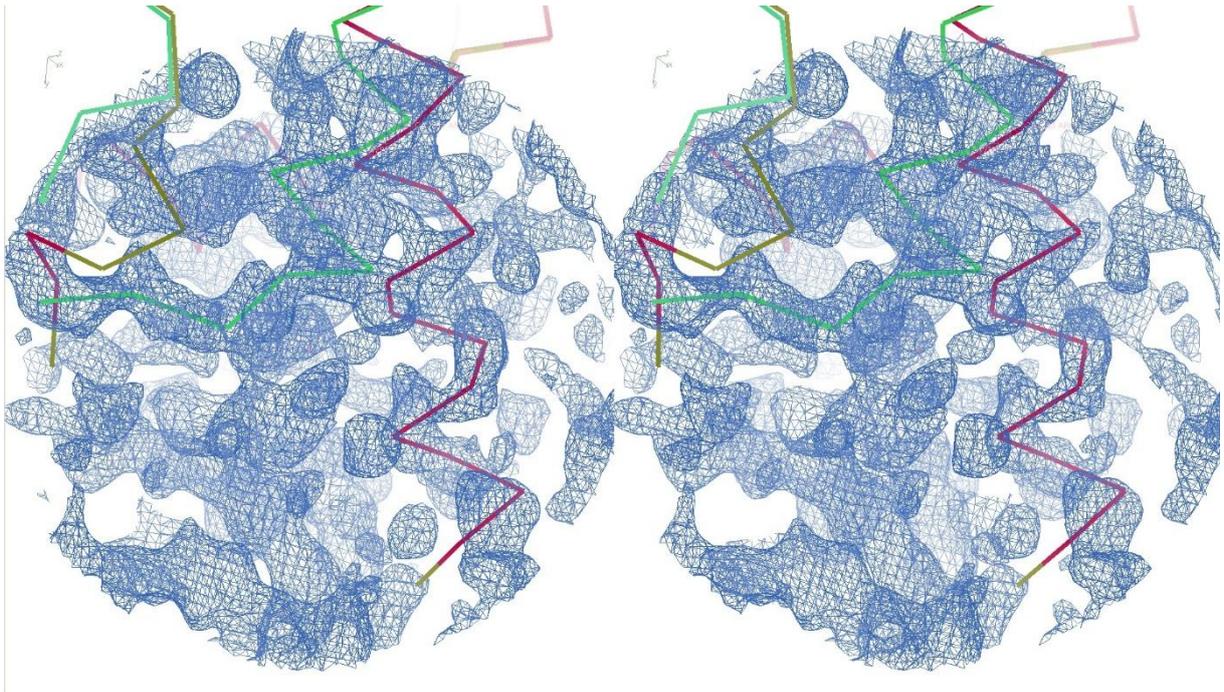


Original structure and density



Re-refined structure and density

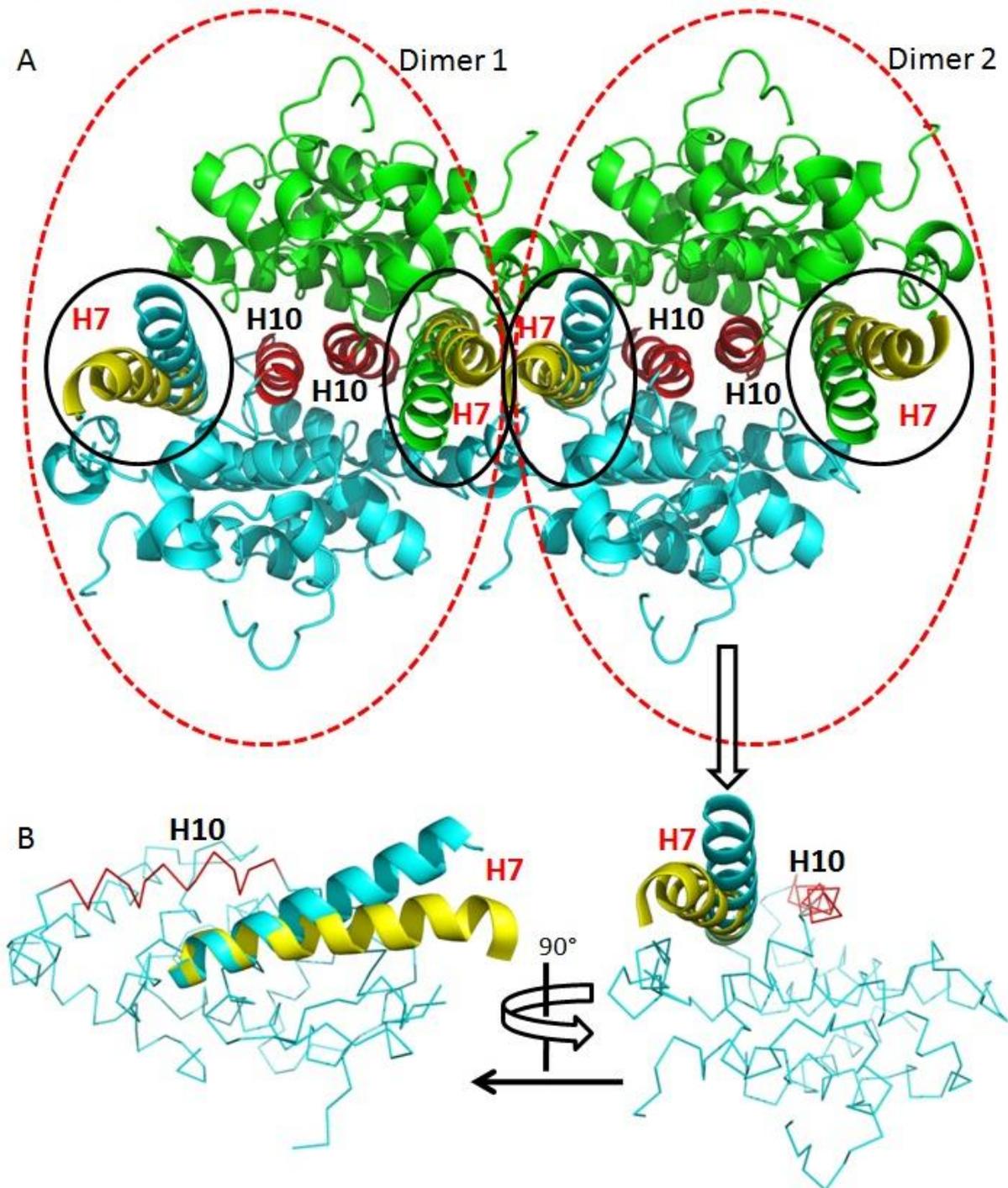
**Figure 1 :** stereo pairs of residues R202 and E206 before and after re-refined (H7 straight ).



**Figure 2** : stereo pairs of H7 helix before and after re-refined.



Figure 3 : Reference sequence alignment for this article.



**Figure 4** : COUP-TFII, structure 3CJW re-refined : crystal packing effect on H7. 2 conformations for H7, a straight conformation in green (monomer at the top of part A) and blue (monomer at the bottom of part A) and a curved conformation in yellow. H10 is red to materialize the dimerization interface.

A) Crystal packing of 2 homodimers. If there are 2 curved H7 helices (yellow) there will be a steric clash. The curved and straight form coexists in equal proportion in the crystal.

B) Front and side views of both H7 conformations.

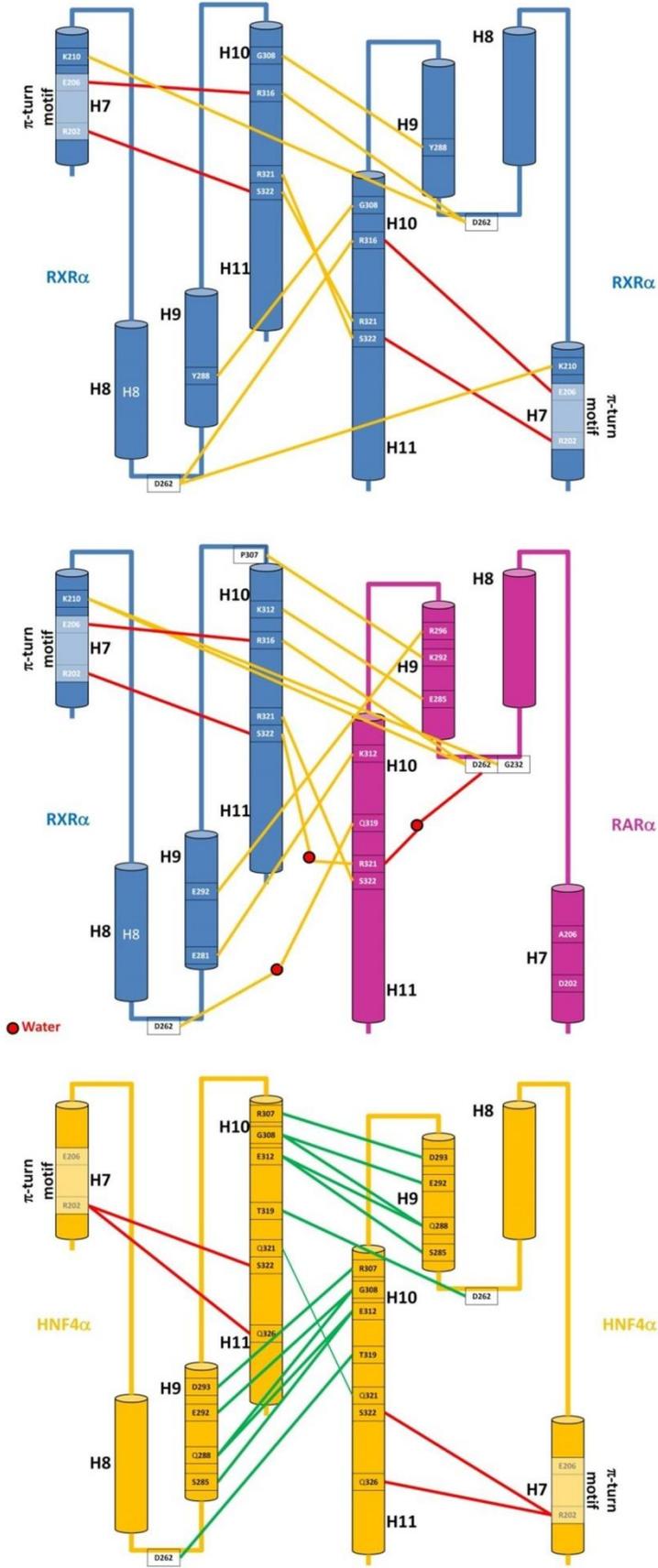


Figure 5 : Dimerization scheme of different nuclear receptors.

### 3. Conclusion et perspectives

Cette étude a permis d'apporter des précisions au sujet de l'apparition des récepteurs nucléaires.

Nous avons ainsi découvert que HNF-4 et RXR étaient présents tous les deux chez les métazoaires les plus ancestraux et non uniquement HNF-4. Pour ces deux récepteurs nucléaires, des séquences de gènes sont disponibles et le motif RxxxE est conservé tout au long de l'évolution, nous avons donc réexaminé l'origine et l'évolution de la superfamille à la lumière de la conservation en séquences et en structure du  $\pi$ -turn.

Nous montrons que quelques autres récepteurs nucléaires proches de RXR ont le motif de séquence RxxxE, mais seuls HNF-4 et RXR présentent une conformation de  $\pi$ -turn à cet endroit. L'analyse structurale de ces récepteurs suggère que le  $\pi$ -turn est stabilisé par des interactions intramoléculaires conservées aux deux extrémités de l'hélice H7, maintenant une conformation intra-hélicoïdale autrement instable à l'emplacement RxxxE.

Plus tard, au cours de l'évolution, RXR a été sélectionné comme partenaire d'hétérodimérisation, probablement en partie à cause de la présence du  $\pi$ -turn, mais aussi à cause de différences subtiles dans la composition en acides aminés par rapport à HNF-4. Ces différences favorisent les interactions complémentaires entre RXR et ses récepteurs nucléaires partenaires. L'utilisation du  $\pi$ -turn peut donc être considérée comme une exaptation (c'est-à-dire une adaptation sélective opportuniste, privilégiant des caractères qui sont utiles à une nouvelle fonction), ceci ayant probablement conduit à la nouvelle fonction cruciale de l'hétérodimérisation et la régulation stricte et spécifique de l'expression des gènes. En effet, les deux récepteurs nucléaires sont présents dans le taxon des Porifera (spongiaires) et dans ce contexte, en absence de récepteurs nucléaires de classe II, leur activité biologique est susceptible de dépendre de la forme homodimérique.

Pour la suite de cette étude, il est envisagé de continuer de suivre l'évolution des récepteurs nucléaires, à savoir que la plupart des grandes familles de récepteurs nucléaires sont apparus dès l'embranchement des bilatériens. Cependant des premières observations demandant investigations laissent penser que la famille des récepteurs stéroïdiens aurait commencé à apparaître déjà chez les Cnidaires sous la forme d'un ERR-like.

## Chapitre 2

# Détermination structurale d'ERR $\alpha$ par cryo-microscopie électronique

---

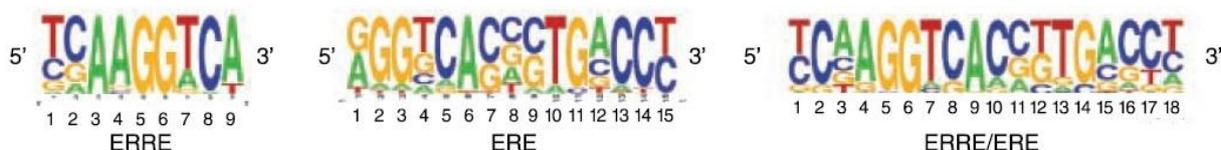
### 1. Introduction

Le récepteur nucléaire ERR (Estrogen Related Receptor) est une protéine suscitant de nombreux intérêts aussi bien pour des raisons de connaissances fondamentales que pour des applications en biomédecine.

ERR $\alpha$  est apparenté aux récepteurs des œstrogènes (ER $\alpha$  et ER $\beta$ ) et se lie à des séquences similaires à celles liées par les ERs sous forme d'homodimères. Vu la meilleure solubilité d'ERR lors de sa purification biochimique observé lors de tests de solubilités et d'optimisation mis en place dans l'équipe, son utilisation comme système modèle des Récepteurs nucléaires homodimériques nous permet d'accéder plus facilement à la connaissance de l'architecture moléculaire et de l'organisation topologique des récepteur nucléaire stéroïdiens homodimériques, le but étant d'analyser le complexe entier en complexe avec l'ADN, et en présence ou en absence de coactivateurs.

Actuellement il n'y a aucun récepteur nucléaire stéroïdien dont la structure serait connue, comprenant le DBD, le hinge et le LBD, fixé sur son ADN cible. Seuls les domaines DBD (Schwabe et al., 1993) et LBD (Gangloff et al., 2001; Greschik et al., 2002, 2008; Tanenbaum et al., 1998) individuels sont connus. De plus, il n'y a à l'heure actuelle aucune structure à haute résolution, tous récepteurs nucléaires confondus, comprenant un LBD lié à un grand fragment de plusieurs centaines d'acides aminés d'un co-activateur.

La singularité du récepteur ERR se traduit également par son mode d'interaction avec l'ADN. En effet, la quasi-totalité des Récepteurs nucléaires se lient à des éléments de réponse dimériques (2x6 nucléotides séparés). ERR est également homodimérique, en raison des fortes interactions entre les LBDs, cependant il peut se lier sur un élément de réponse ADN comprenant un demi-site étendu en 5' (ERRE-embedded ER). La raison moléculaire de ce mode de fixation n'est pas connue et constitue l'un des points d'intérêts de l'étude structurale. Cette spécificité d'interaction pourrait aider à cibler précisément ERR et son interface avec l'ADN.



**Figure 54** : Eléments de réponses reconnus par ER et ERR. La tailles des lettres est proportionnelle à la fréquence de distribution des nucléotides. ERRE correspond à l'élément de réponse d'ERR, ERE correspond à l'élément de réponse de ER et ERRE/ERE correspond à la combinaison des deux éléments de réponses. (Reproduit de Deblois et al., 2009).

Le coactivateur choisi pour former un complexe avec ERR $\alpha$  est PGC-1 $\alpha$ . Ainsi au cours de cette étude, plusieurs complexes sont formés : le complexe ERR $\alpha$ -ADN, le complexe ERR $\alpha$ -PGC-1 $\alpha$ -ADN et différentes variantes du complexe ERR $\alpha$ -nucléosome. PGC-1 $\alpha$  est un régulateur crucial du métabolisme cellulaire et de l'homéostasie énergétique, agissant de manière fonctionnelle avec les récepteurs liés aux œstrogènes (ERR $\alpha$  et ERR $\gamma$ ) dans la régulation des réseaux de gènes mitochondriaux et métaboliques. La dimérisation des ERR est nécessaire pour les interactions avec PGC-1 $\alpha$  et d'autres coactivateurs, permettant ainsi son activité transcriptionnelle (Takacs et al., 2013). Le nucléosome (NCP) est un complexe de 8 protéines (octamère d'histones) avec un segment d'ADN d'environ 146 nucléotides enroulé autour. Chez les eucaryotes, le nucléosome constitue l'unité de base d'organisation de la chromatine. Il représente le premier niveau de compaction de l'ADN dans le noyau (Luger et al., 1997).

L'étude de tels complexes demande d'aller aux limites techniques de la cryo-microscopie électronique. C'est un défi technologique en soi. En effet, le récepteur nucléaire étudié, ERR, dont le poids moléculaire comprenant un homodimère lié à son ADN cible ne fait que 100 kDa et le poids moléculaire comprenant un homodimère lié à son ADN cible et un coactivateur PGC-1 $\alpha$  fait 170 kDa. Le défi réside aussi bien dans la préparation de l'échantillon sur grille de cryo-microscopie électronique que dans le traitement des données afin d'obtenir une structure à moyenne résolution, voir à haute résolution.

Des études précédentes de l'équipe publiées en 2012 et 2014 ont déjà porté sur une étude structurale par cryo-microscopie électronique de récepteurs nucléaires comprenant DBD, LBD et ADN. La technologie disponible à l'époque ne permettait pas d'accéder à la haute résolution avec des complexes aussi petits. Les acquisitions avaient été faites sur un microscope Polara avec une caméra Falcon I, technologie la plus avancée à ce moment. Les cartes finales obtenues à partir de la cryo-ME dans lesquelles ont été positionné et recalé les structures atomiques de domaines obtenues par cristallographie ont une résolution d'environ 10 Å (Maletta et al., 2014; Orlov et al., 2012). Il s'agit, pour ces deux complexes, de récepteurs nucléaires hétérodimères, RXR/VDR et USP/EcR liés à leur fragment d'ADN promoteur. L'objectif de mon projet était idéalement d'obtenir une structure à résolution quasi atomique d'un complexe de récepteur nucléaire qui nous permettra de construire un modèle moléculaire sans avoir besoin d'aligner des domaines obtenus par cristallographie. Comme décrit par la suite, ceci a demandé beaucoup d'optimisation et le traitement des données est encore en cours au moment de la rédaction.

Les complexes n'avaient pas un comportement idéal sur grilles de cryo-ME, et les grilles étaient difficilement reproductibles, d'où la nécessité de rechercher de manière systématique les bonnes conditions de congélation et la composition idéale de tampon afin d'obtenir des grilles avec une bonne distribution des particules, une bonne concentration et une non-dissociation du complexe, notamment au regard de la présence de l'ADN dans le complexe. A partir d'échantillons purifiés dans l'équipe (par Kareem Mohideen-Abdul et Isabelle Billas), j'ai pu optimiser la composition du tampon pour la congélation des grilles tout en maintenant la stabilité du complexe durant la purification. J'ai effectué la congélation à l'aide d'un appareil de congélation, un Vitrobot, grâce auquel j'ai réalisé au cours de ma thèse plusieurs centaines de grilles en faisant varier les paramètres physiques (température, humidité, pression, temps...). Chaque grille a été d'abord vérifiée sur le microscope Polara en condition cryo, puis en cas d'obtention d'images avec une bonne distribution de particules, un jeu de données a été acquis sur le Titan Krios.

## 2. Matériel et méthodes

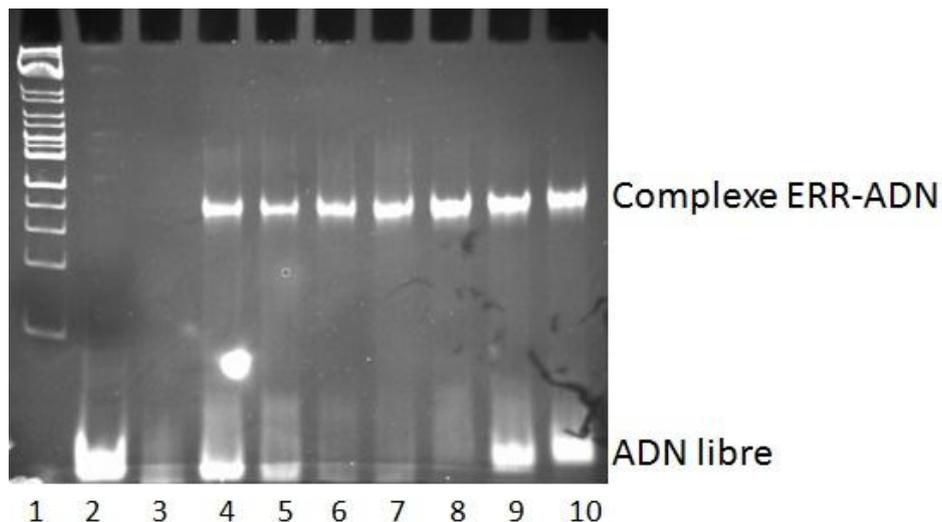
L'expression et la purification du récepteur nucléaire ERR $\alpha$  et du coactivateur PGC-1 $\alpha$  ainsi que la formation des différents complexes ont été optimisées au sein de l'équipe pour les études structurales de cryo-ME. Je me suis donc focalisé sur la préparation et l'optimisation des grilles puis sur le traitement d'images du complexe. Après une optimisation des conditions adéquates pour la

production de grilles de cryo-ME qui a nécessité de très nombreux tests, j'ai réussi à produire de façon reproductible des grilles de bonne qualité (bonne concentration, distribution homogène des complexes, stabilisation du complexe lors de l'étape de cryogénéisation, conditions de cryogénéisation). Ceci nous a permis d'acquérir des micrographes de bonne qualité sur les microscopes électroniques à haute résolution Polara et Titan Krios de l'institut (CBI-IGBMC).

Devant les difficultés rencontrées lors du traitement de données pour ces petits complexes asymétriques, j'ai combiné différents logiciels car un seul ne permettait pas de mener à bien une reconstruction 3D dans mon cas d'étude. J'ai utilisé par exemple pour le prétraitement, MotionCor2, Gctf, CTFind4, et pour le traitement, cisTEM, RELION, EMAN-2, IMAGIC, Sphire et cryoSPARC.

## 2.1 La production et purification des complexes

La partie purification et formation des complexes est faite au sein de l'équipe par Kareem Mohideen-Abdul et Isabelle Billas. Les protéines sont produites dans des cellules d'insectes SF9 via un système baculovirus. Après sonication des cellules, la première étape de purification est une chromatographie d'affinité sur une colonne nickel suivi par une chromatographie d'exclusion stérique sur colonne de filtration sur gel. Dans un second temps, pour le complexe  $ERR\alpha$ -ADN, il y a formation des complexes en ajoutant l'ADN au dimère d' $ERR\alpha$ . Le complexe  $ERR\alpha$ -ADN-PGC-1 $\alpha$  a d'abord été reconstitué à partir de  $ERR\alpha$  et PGC-1 $\alpha$  exprimés séparément dans des cellules SF9, mais la production de PGC-1 $\alpha$  seul avait un rendement très faible, le problème a été résolu en co-exprimant les protéines  $ERR\alpha$ -PGC-1 $\alpha$  ensemble, puis en ajoutant l'ADN après purification du complexe  $ERR\alpha$ -PGC-1 $\alpha$ . Pour les complexes contenant des nucléosomes, la purification est faite de la même manière pour  $ERR\alpha$  ou  $ERR\alpha$ -PGC-1 $\alpha$  le nucléosome est produit à part selon des protocoles mis en place par Kareem Mohideen-Abdul dans l'équipe, puis on reconstitue le complexe  $ERR\alpha$  ou  $ERR\alpha$ -PGC-1 $\alpha$  et nucléosome en respectant les rapports molaires des composants du complexe.



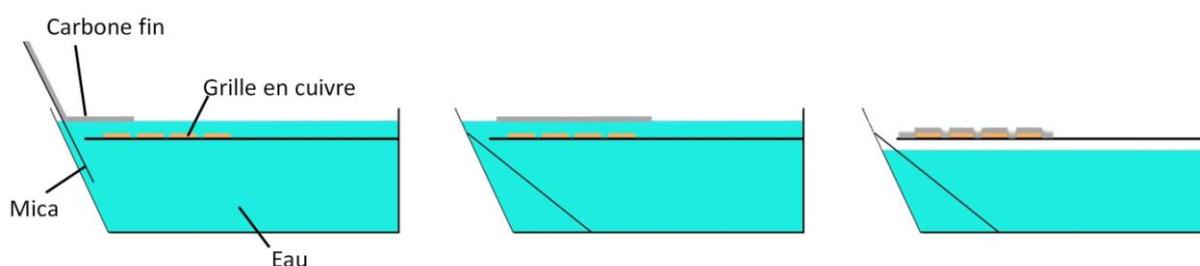
**Figure 55** : Gel natif 5% EMSA d'un complexe ERR $\alpha$ -ADN. La colonne 1 est le marqueur de poids moléculaires. La colonne 2 est l'ADN seul, la colonne 3 est ERR seul. Les colonnes suivantes sont une série de rapports molaires croissants entre ERR et ADN.

## 2.2 La congélation de l'échantillon

### 2.2.1 Les grilles

J'ai effectué une optimisation approfondie des échantillons par rapport à la distribution de particules et à l'état d'agrégation des objets en cryo-microscopie électronique, notamment en utilisant différents types de grilles commerciales et faites maison, dont des grilles Quantifoil qui sont des grilles en cuivre et en rhodium. Du côté du cuivre de la grille, il y a une feuille de carbone pré-perforée d'environ 10 à 12 nm d'épaisseur qui est formé par évaporation de carbone. Pour ce faire, il y a initialement une couche de plastique qui sert de matrice à la couche de carbone évaporé. A cause de ce procédé, il peut rester des résidus de plastique après fabrication. Ils sont très mauvais pour l'imagerie en microscopie électronique car cela crée une instabilité de l'échantillon. Il est conseillé de laver ses grilles avec de l'acétate d'éthyle avant toute utilisation. Il y a de nombreuses variantes de ces grilles qui existent sur le marché et j'en ai utilisé 3 types, les Quantifoil R1.2/1.3, les Quantifoil R2/2 et les Quantifoil R3.5/1. Le premier chiffre correspond à la taille du trou perforé et le second chiffre correspond à l'espacement entre les trous en  $\mu\text{m}$ . Similairement, j'ai également utilisé la version en or (Quantifoil UltraAuFoil) avec une géométrie R1.2/1.3. Pour ces grilles entièrement en or, la feuille perforée fait environ 50 nm d'épaisseur. Le principal avantage de ces grilles est leurs stabilités lors de l'acquisition et sa bonne résistance aux écarts de températures. J'ai utilisé des grilles

C-flat provenant d'un autre fournisseur, ce sont des grilles similaires au Quantifoil avec une feuille de carbone. La principale différence est la méthode de fabrication (nano perforation), qui n'utilise pas de matrice en plastique, elles sont donc directement utilisables sans lavage préalable. Enfin pour les grilles faites maison, je suis parti de grilles commerciales Quantifoil R3.5/1 pour préparer des grilles de carbone continue avec la méthode du carbone flotté. Ces grilles permettent de ne pas avoir de différences sur le support tout au long de la grille, ce qui permet dans certain cas une meilleure distribution des particules. Il y aura néanmoins une légère perte de contraste. Au début du projet j'ai également utilisé et préparé des grilles de coloration négative, l'agent de coloration utilisé étant l'acétate d'uranyle.

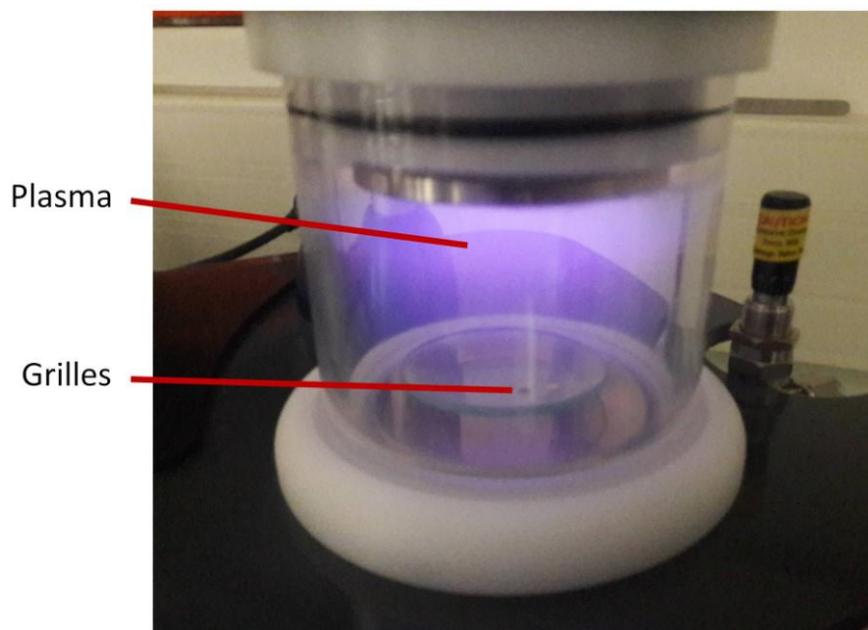


**Figure 56** : Schéma de la préparation de grilles avec du carbone flotté. Après le dépôt d'une fine couche de carbone par évaporation sur une plaque de mica, on plonge la plaque dans de l'eau pour faire flotter le carbone à la surface de l'eau. Il suffit ensuite de faire baisser le niveau de l'eau en dessous du niveau où se trouve les grilles puis de laisser sécher.

### 2.2.2 Système à décharge lumineuse

La décharge lumineuse (Glow Discharge) est un procédé ayant pour but de rendre le support carbone des grilles plus hydrophiles pour une bonne adsorption de l'échantillon sur la grille. J'ai utilisé deux modèles de machines différentes pour réaliser cette action, un Elmo cordouan et un Fischione model 1070 NanoClean plus récent qui donne une meilleure reproductibilité des résultats. Les grilles sont placées dans une chambre connectée à un générateur électrique qui permet d'ioniser l'air ambiante dans la chambre qui est placée en dépression. Dans le cas de la première machine, la manipulation se fait avec la composition de l'air atmosphérique en dépression (200 mbar), par contre avec la seconde machine on peut choisir sa composition en mélangeant de l'oxygène et de l'argon. L'ionisation du gaz interne va créer du plasma, qui aura pour conséquence l'accumulation de charges négatives à la surface du carbone qui devient alors hydrophile. Dans le cas où on utilise du graphène qui a une structure cristalline contrairement au carbone amorphe, la décharge lumineuse va faire

passer le carbone d'une forme SP3 à SP2. L'effet diminue progressivement pendant les jours qui suivent.



**Figure 57** : Photographie d'une décharge lumineuse avec une machine Elmo cordouan.

### 2.2.3 Cryogénisation des échantillons

La cryogénisation des échantillons a été réalisée avec un Vitrobot (MARK IV) qui est le plongeur de FEI. Après l'étape de décharge, on prépare un compartiment en métal rempli d'éthane liquide au centre d'un compartiment plus grand, rempli d'azote liquide pour maintenir le premier au froid. On accroche ensuite une grille sur la pince du plongeur qui l'amène dans une chambre qui est maintenue à température et à hydrométrie constante. Dans mon cas, l'hydrométrie est toujours comprise entre 95 et 100% d'humidité. La plage de température testée va de 4 à 20 degrés Celsius et en fonction du complexe utilisé, la température la plus adaptée n'est pas toujours la même. Pour le dépôt d'échantillons sur la grille, j'ai utilisé une variation allant de 2 à 4  $\mu\text{L}$ . Cependant, il faut noter que 3  $\mu\text{L}$  est une quantité adaptée à la grande majorité des cas et est devenue ma valeur par défaut. Ensuite on doit choisir le temps d'incubation de l'échantillon avant absorption par le papier filtre, le temps de l'absorption par papier filtre et la force avec laquelle on plaque le papier contre la grille de part et d'autre. Dans certains cas, j'ai aussi procédé à la mise en place de plusieurs gouttes et plusieurs absorptions sur la même grille afin de reconcentrer l'échantillon sur grille. Après absorption la grille

est immédiatement plongé dans l'éthane liquide pour former de la glace amorphe contenant l'échantillon. Les conditions optimales pour chaque complexe seront décrites dans la partie résultat.

### 2.3 Acquisition des images

Tous les microscopes utilisés sont présents au sein de l'institut où j'ai effectué ma thèse (CBI-IGBMC). Chaque grille est produite en double exemplaire lors de la phase d'optimisation. Le premier exemplaire est toujours observé sur un microscope Tecnai Polara de FEI équipé d'un porte échantillon de Gatan, et d'une caméra Falcon I au début et au milieu de ma thèse et depuis peu d'une caméra Falcon II. Cette phase de criblage vise à optimiser les conditions de préparations de l'échantillon et n'a pas nécessairement pour but de faire une acquisition pour obtenir des résultats à haute résolution. Quand une grille est de bonne qualité, elle est difficilement récupérable pour faire une acquisition avec le Titan Krios de FEI. Par conséquent, je réalise une petite acquisition sur le Polara afin de réaliser un premier traitement d'images et si possible d'obtenir une première structure à basse résolution permettant d'avoir une première idée de la qualité et l'homogénéité de l'échantillon. Avec ce microscope je travaille avec une tension de 100 kV afin d'augmenter le contraste (en comparaison à 300 kV), un grandissement compris entre 56 000 et 93 000 fois et une dose totale comprise entre 15 et 40 électron par  $\text{Å}^2$  ( $\text{é}/\text{Å}^2$ ). La plage de défocalisation utilisée est comprise entre -1,5 et -4,5  $\mu\text{m}$  par pas de 0,3  $\mu\text{m}$ . Le criblage est manuel et les acquisitions sont faites par le logiciel d'automatisation EPU (logiciel propriétaire de FEI).

Lorsqu'une grille de bonne qualité est identifiée, sa réplique est stockée dans l'azote liquide en attendant d'avoir une session sur le microscope Titan Krios. Si l'échantillon à l'origine de cette grille est encore frais, une série de 3 autres grilles identiques est congelée (le stockage se fait par boîtes de 4 grilles). Les acquisitions en vue d'obtenir une structure à résolution quasi atomique voire atomique sont faites sur le microscope Titan Krios. Au début de ma thèse, il était équipé d'un correcteur Cs et d'une caméra Falcon II. En cours de thèse il a eu des mises à jour comme l'installation d'une phase plate, d'une caméra Falcon III et d'une caméra Gatan K2 Summit. Ce microscope est uniquement destiné à l'acquisition et est utilisé à une tension de 300 kV. La plage de défocalisation est dépendante des conditions d'imagerie. En absence de phase plate, la gamme standard est comprise entre -1,5 et -3,0  $\mu\text{m}$ . Avec la phase plate, on vise une défocale fixe de -0,5  $\mu\text{m}$ . Les 3 caméras que j'ai utilisées sur ce microscope sont capables de produire des images en mode film, comme décrit dans la partie introduction. Les logiciels d'automatisation de l'acquisition utilisés sont EPU et SerialEM.

### 2.4 Le traitement d'images et reconstruction 3D

Une fois l'acquisition terminée, sur le Polara ou le Titan Krios, viennent les différentes étapes de traitement d'images afin d'obtenir une structure du complexe acquis. La résolution structurale nécessite une quantité de données et un temps de calcul très important afin de retrouver l'orientation de chaque particule du jeu de données très bruité, puis de reconstruire l'objet en 3D et affiner la structure.

#### 2.4.1 Le prétraitement

Pour les acquisitions faites sur le Titan Krios en mode film, il est nécessaire dans un premier temps d'aligner les images d'un même micrographe entre elles pour former le micrographe final moyenné qui sera utilisé pour la suite du traitement. Le but ici est d'éliminer les images qui auront un impact négatif sur le micrographe final, à savoir les 2 premières images à cause de l'instabilité de la caméra pour les premiers électrons reçus et les effets de charge qui se produisent sur l'échantillon. On peut aussi choisir à ce moment la dose totale de l'échantillon en supprimant les dernières images de la pile d'images. Chacune des images sera pondérée les unes par rapport aux autres en fonction de la dose totale reçue, ceci permettant d'améliorer la qualité du signal. L'alignement en lui-même permet d'avoir des micrographes de meilleure qualité, car les caméras ont une plage de réponse linéaire optimale à la dose reçue par unité de temps. Il est par conséquent nécessaire de faire varier le temps de l'acquisition jusqu'à atteindre la dose totale désirée en restant dans la fourchette préconisée par le constructeur et non pas augmenter la dose par unité de temps. Il faut faire une longue acquisition de 5 à 10 secondes pour la caméra K2 summit et une acquisition de 50-60 secondes avec la caméra Falcon III. Ce temps de pause est trop long et ne permet pas d'éviter le déplacement de l'échantillon pendant la prise de vue, mais l'alignement permet de contrer cet effet négatif. Pour ce faire j'ai utilisé le logiciel MotionCorr2 qui s'utilise en ligne de commande (Command Line Interface : CLI). Une fois les images de chaque micrographe moyenné, on peut calculer la FTC (Fonction de Transfert de Contraste; CTF) de chaque image. Ce faisant on détermine la valeur de défocalisation précise à une dizaine de nanomètres de chaque micrographe, l'astigmatisme, et la courbe théorique de la FTC. Cette valeur est très importante pour les logiciels de traitement. En effet s'il y a un très léger décalage entre le spectre de puissance expérimentale et le spectre de puissance théorique, les résolutions correspondantes (les hautes résolutions) ne pourront pas être récupérées et exploitées lors du traitement d'images. Pour cette estimation de la FTC, j'ai utilisé deux logiciels, CTFFind4 et

gCTF. En plus de la FTC, lorsque l'on fait une acquisition avec une phase plate, il faut prendre en compte le déphasage des électrons qui évolue dans le temps. Le déphasage est compris entre 0 et 180° et augmente progressivement au fur et à mesure que la phase plate se fait irradier par le faisceau d'électrons. Le déphasage conservé pour le traitement est compris entre 35 et 145°, les 145° étant atteint en environ 2H d'irradiation. Par conséquent durant l'acquisition, il est nécessaire de bouger la position de la phase plate toutes les 2H environ. Cette valeur de déphasage est également calculée par CTFFind4 et gCTF si on leur met les paramètres adaptés. Ces deux logiciels s'utilisent avec une CLI ou via une interface intégrée à un logiciel de traitement.

---

#### 2.4.2 Le tri des micrographes

Après ces deux étapes, chaque micrographe est visualisé et conservé ou éliminé manuellement. Je prends en compte plusieurs paramètres pour prendre cette décision. Il faut que le micrographe semble visuellement correct, c'est-à-dire qu'il y ait absence de grande surface de contaminants telle que de la glace, absence de fracture dans la glace, absence de dérive de l'échantillon (provoque un flou de bougé comme en photographie classique), une glace amorphe pas trop épaisse et de bonne qualité, la présence de particules avec une distribution correcte et absence de zones importantes d'agrégations. Les paramètres suivants sont vérifiés : la valeur de l'astigmatisme et si elle existe, si elle est détectée par le logiciel ou non ; la bonne concordance des anneaux de Thon du spectre de puissance théorique avec ceux de la FTC expérimentale pour assurer une bonne estimation de la FTC, et ce surtout pour les hautes résolutions (anneaux de Thon les plus en périphérie) ; la valeur du déphasage qui doit être comprise entre 35 et 145°. Si l'une de ces règles n'est pas respectée, le micrographe est éliminé du jeu de données. Le tri des micrographes peut se faire sur plusieurs logiciels de traitement. Pour cette étape j'ai notamment utilisé au cours de ma thèse les logiciels Relion, cisTEM, EMAN2 et cryoSPARC v2 (en version bêta durant l'été 2018).

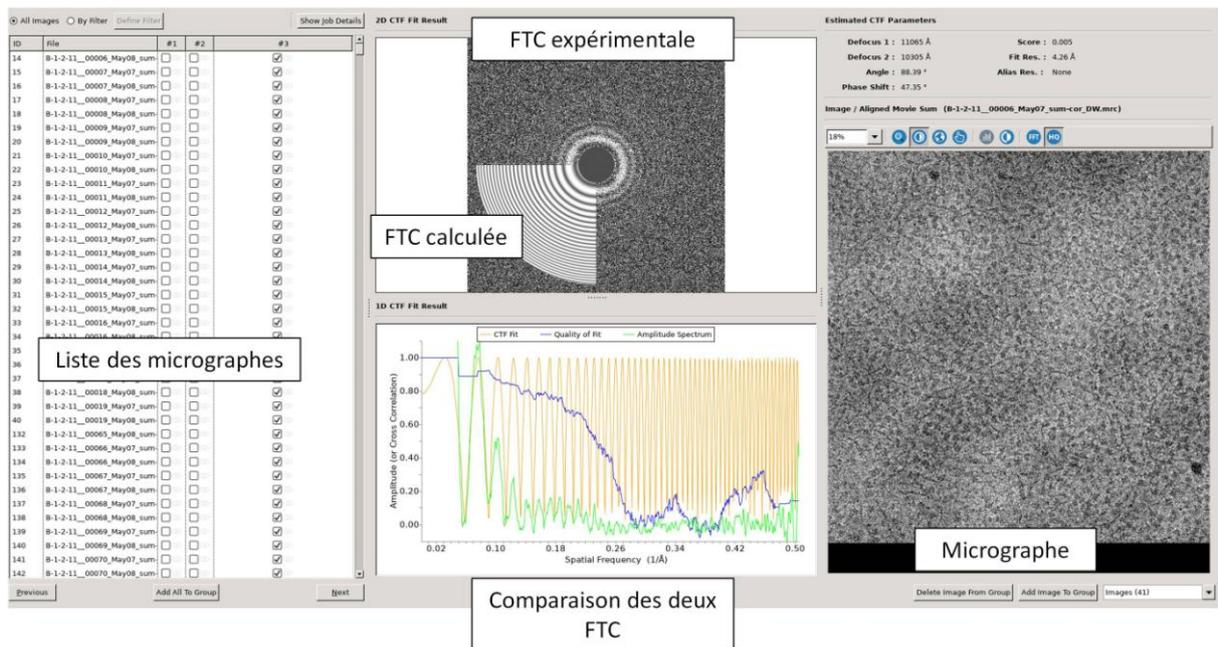
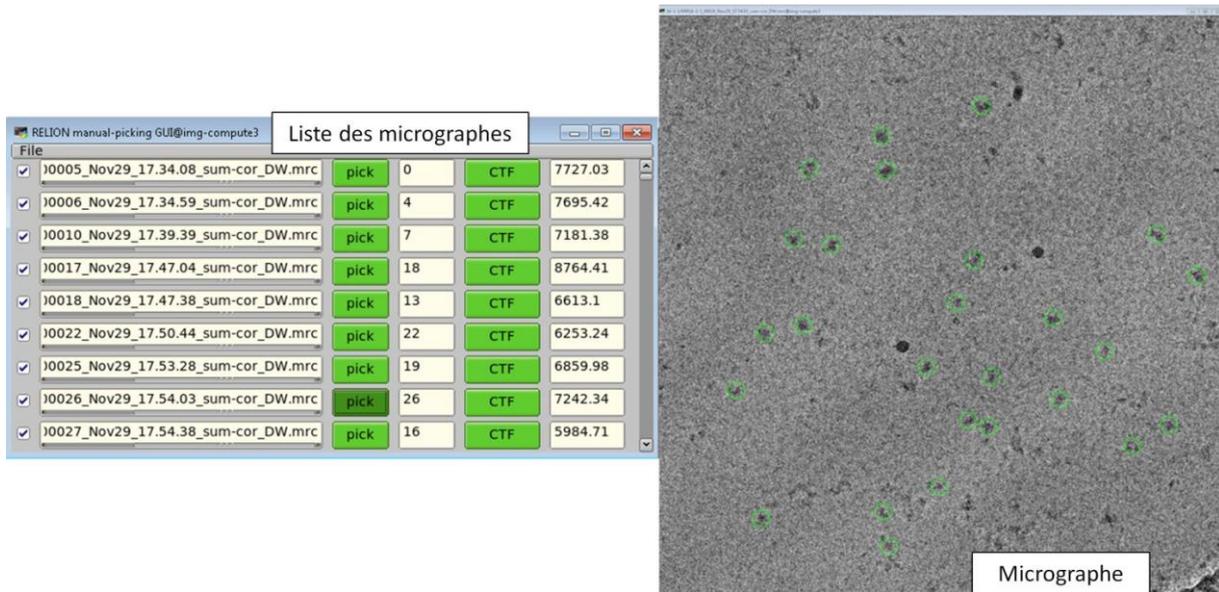


Figure 58 : Exemple d'une interface de trie des micrographes (cisTEM).

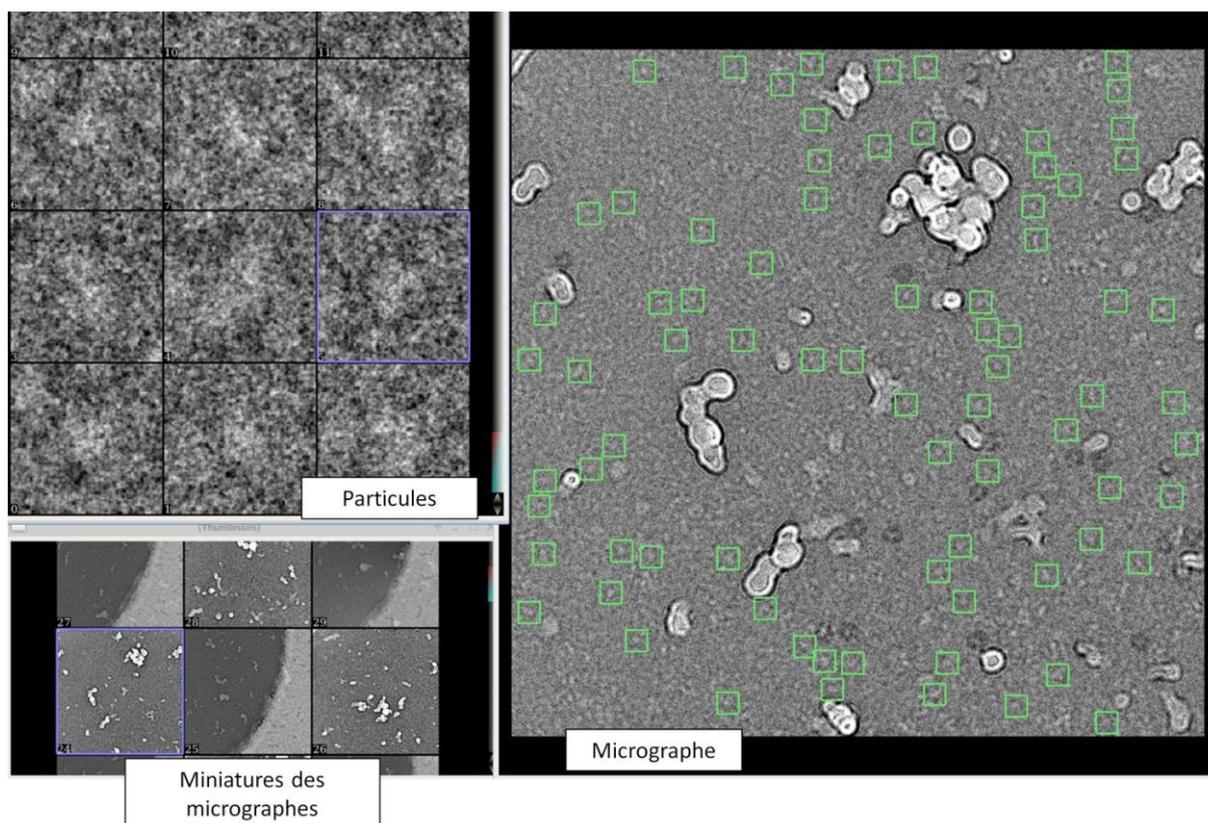
### 2.4.3 La sélection des particules

Pour cette étape il y a plusieurs méthodes et stratégies pouvant être utilisées, chacune d'entre-elles étant adaptées ou non en fonction de la nature de l'échantillon. J'ai utilisé pour cette étape les mêmes logiciels que pour l'étape du tri des micrographes. Le principe ici est de sélectionner les coordonnées (X, Y) du centre de chaque particule présente sur les micrographes conservés. Le bon centrage de la particule est un détail important pour la suite. Dans le cas de Relion, l'approche est basée sur une approche automatique, elle-même basée sur une référence obtenue manuellement. Il s'agit ici de sélectionner à la main un ensemble d'environ 1000 à 2000 particules. A partir de ces particules on peut faire une première classification 2D (CF paragraphe suivant sur la classification 2D) et isoler 3 à 5 classes de bonne qualité qui représentent l'échantillon. Puis basé sur cette référence, on démarre alors une sélection automatique sur l'ensemble des micrographes. Cette étape peut prendre du temps, plus de 24H sur un serveur de calcul s'il y a beaucoup de données, c'est pourquoi il est préconisé de d'abord optimiser la sensibilité de la sélection automatique sur un petit nombre de micrographes (4 par exemple).



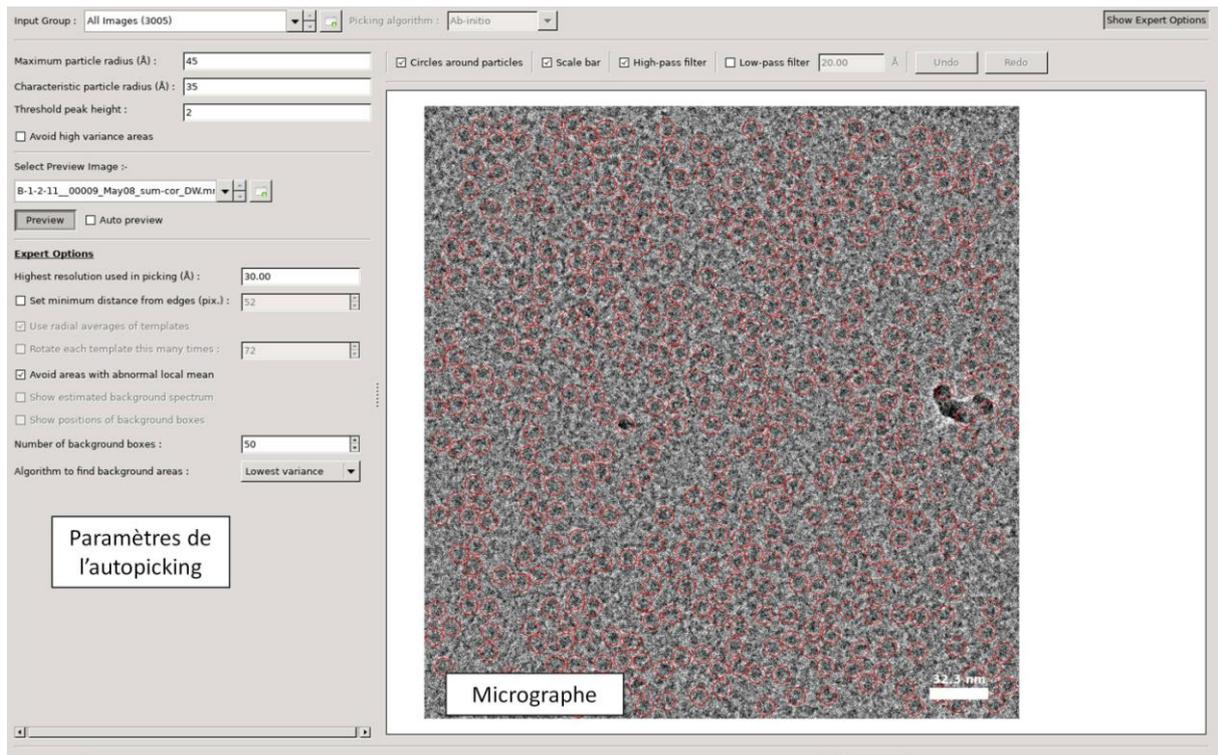
**Figure 59** : Exemple d'une interface de sélection de particules manuelle (Relion).

Le logiciel Eman2 quant à lui a un mode semi-automatique de sélection des particules. Le logiciel apprend au fur et à mesure ce qui est recherché. Sur le premier micrographe, on commence par sélectionner à la main les premières particules, le logiciel prend cette sélection comme référence et fait une proposition de sélection complémentaire sur le même micrographe. L'utilisateur doit après continuer de sélectionner les particules non détectées et supprimer les faux-positifs. Lorsque l'on passe sur les micrographes suivants, une proposition de sélection est faite et s'affine au fur et à mesure que l'utilisateur sélectionne ou désélectionne manuellement certaines particules.



**Figure 60** : Exemple d'une interface de sélection de particules semi manuelle (EMAN2).

Pour cisTEM, la stratégie est différente, il ne demande aucune référence mais veut uniquement connaître la taille et la masse moléculaire approximative du complexe. A partir d'un micrographe, il fait une proposition de sélection, puis on affine au fur et à mesure les paramètres de sensibilité jusqu'à obtenir une sélection en majorité correcte sur plusieurs micrographes. Une fois ceci fait, on lance le calcul de sélection automatique. Cette méthode est beaucoup plus rapide que celle de Relion, mais pour des petits complexes comme les miens, cela génère plus de faux positifs.

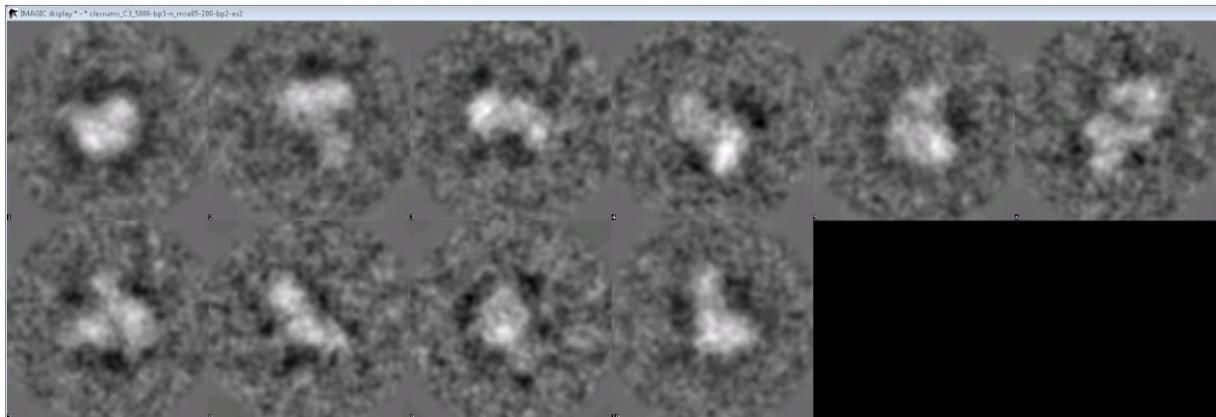


**Figure 61** : Exemple d'une interface de sélection de particules automatique (cisTEM).

#### 2.4.4 La classification 2D

La classification 2D permet de regrouper les particules similaires ayant environ la même orientation ensemble au sein d'une même classe. Pour ce faire il y a deux approches, la première est une méthode d'analyse statistique multivariée (MSA; Multivariate Statistical Analysis). Il y a plusieurs algorithmes utilisés comme par exemple, la K-moyennes (K-means). Le principe est qu'on souhaite répartir  $n$  images en  $k$  classes. Pour ce faire on sélectionne au hasard  $k$  images parmi l'ensemble  $n$ , ce sera les premières graines de classifications. Ensuite on compare toutes les autres images au  $k$  première images une à une. On attribue ainsi à chaque image une classe en fonction de sa ressemblance avec la graine qui a initiée la classe. Après une première itération, chacune des  $k$  graines d'origines se déplace vers le centre des valeurs  $k'$  de tous les éléments du total de  $n$  qui se sont avérés être les plus proches de la graine aléatoire d'origine  $k$ . Tous les éléments  $n$  originaux de l'ensemble de données sont ensuite classés par rapport aux  $k'$ , alors nouveaux centres de classification, puis le processus est répété itérativement.

La seconde est une approche de maximum de vraisemblance (Maximum Likelihood ; ML). Les logiciels répartissent les particules de manière aléatoire dans un ensemble de classes (le nombre de classes demandées par l'utilisateur). Chaque particule est soumise à une série de translations et de rotations puis comparée aux autres afin de déterminer si des particules se ressemblent. A la fin, une première moyenne de classe est alors générée. Cette moyenne servira de référence pour l'itération suivante où là les particules sont comparées à chaque moyenne de classe pour être triées correctement. La classification continue ainsi de manière itérative avec un nombre d'itérations prévu au départ par l'utilisateur. Cette méthode permet également de regrouper les particules issues des faux-positifs et ainsi de les éliminer. Les algorithmes sont souvent basés sur la méthode de vraisemblance maximale, c'est à dire qu'il détermine la probabilité pour une particule unique de son appartenance à chacune des classes de la classification. Cette méthode permet de ne pas faire de choix strict. En effet les probabilités désignent une classe, ce choix n'est pas figé pour l'itération suivante, d'où les problèmes d'instabilité de la procédure en comparaison avec une classification basée sur MSA. Pour cette opération, j'ai testé plusieurs logiciels qui donnent des résultats différents pour un même jeu de données de départ. Il y a notamment les logiciels Relion, cisTEM, EMAN2, IMAGIC, cryoSPARC et cryoSPARC 2 (en version beta). A la suite de cette opération, on sélectionne les classes qui nous intéressent pour la suite de la reconstruction 3D.

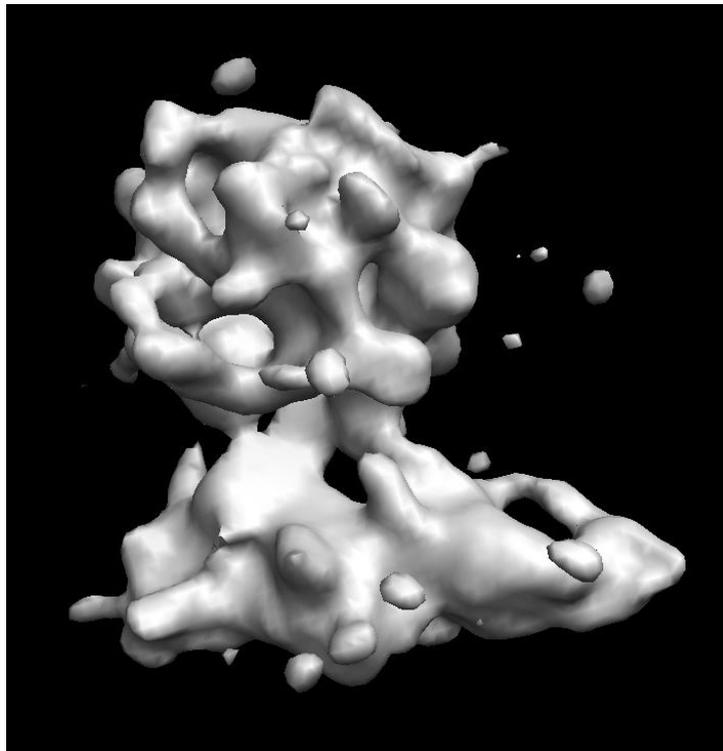


**Figure 62** : Exemple de classification 2D pour un complexe ERR $\alpha$ -ADN-PGC-1 $\alpha$  (IMAGIC).

### 2.4.5 Détermination d'une structure initiale

Cette étape permet d'obtenir une première reconstruction 3D à partir des images 2D des particules. Les principes fondamentaux de ce procédé sont expliqués dans la partie introduction. Ici l'enjeu de cette méthode est de générer une première structure à résolution modérée mais fiable, représentant bien le jeu de données qui servira de référence par la suite. Plusieurs logiciels permettent de générer

une structure initiale. J'ai notamment utilisé Relion, cisTEM, EMAN2, cryoSPARC et cryoSPARC 2 (en version bêta).

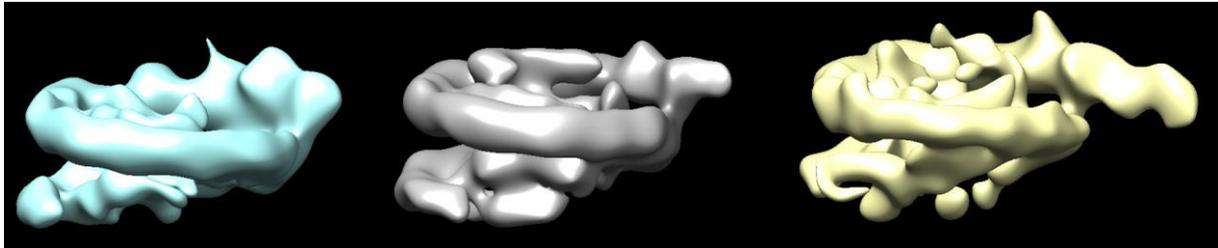


**Figure 63** : Exemple de structure initiale pour un complexe ERR-ADN (cryoSPARC). On reconnaît les domaines LBD en haut et l'ADN allongé en bas.

#### 2.4.6 Classification 3D

Cette étape prend en entrée l'ensemble des classes 2D retenues et la structure initiale préalablement générée. L'approche est la même que pour la classification 2D, les particules sont réparties aléatoirement dans les classes 3D de références, à savoir qu'ici la première référence est la structure initiale filtrée avec un filtre passe-bande. Ce filtre permet d'éliminer toute les hautes fréquences de la structure de référence, pour éviter de créer un biais dans la reconstruction. Ici, la méthode du maximum de vraisemblance va attribuer à chaque particule une pondération selon la probabilité pour une orientation donnée pour une classe donnée. Comme pour la classification 2D, cette classification se fait de manière itérative. Les jeux de données sont souvent hétérogènes. Il y a des sous-populations, soit en composition, parce qu'une protéine manque sur une partie des particules par exemple ou bien parce que l'échantillon étant figé directement en solution, plusieurs conformations différentes des zones flexibles peuvent coexister. Le choix des classes 3D dépend aussi

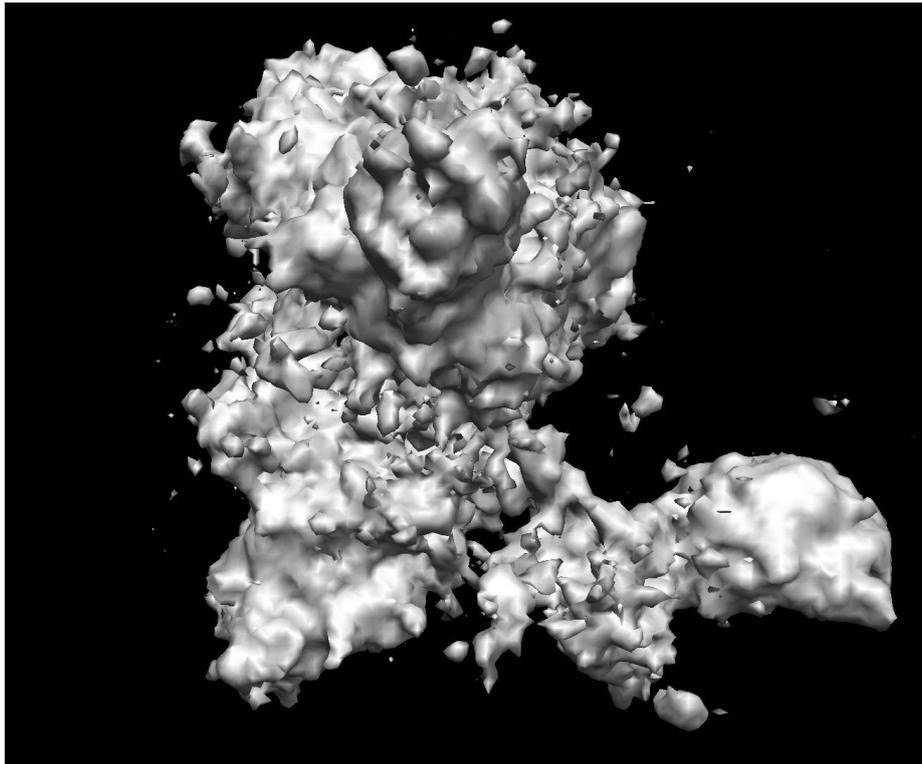
bien des différences entre les classes que du nombre de particules par classes. J'ai réalisé cette étape sur les logiciels suivants, Relion, cisTEM, cryoSPARC et cryoSPARC 2 (en version beta).



**Figure 64** : Exemple de classification 3D pour un complexe de nucléosome-ERR (Relion). On voit différentes étapes de l'ouverture de l'ADN.

### 2.4.7 Affinement de la structure 3D

Après la classification 3D, on peut définir la méthode à adopter pour l'affinement de la ou des structures à partir du jeu de données. Si la différence entre les classes est trop petite, ou si le nombre de particules totales n'est pas suffisant, il peut être plus adapté de sélectionner les meilleures classes 3D qui contiennent la majorité des particules et les fusionner ensemble. Par contre, si le nombre de particules le permet et si on constate une vraie variance entre les sous-populations, on peut également conserver ces classes et les affiner indépendamment les unes des autres. Cette étape consiste à affiner itérativement les paramètres d'orientation de chaque particule par rapport à la structure de référence issue de la classification 3D sur laquelle on applique un filtre passe-bande. L'affinement se fait en séparant les données en deux lots aléatoires et totalement indépendants. Les deux lots sont affinés en parallèle sans jamais se croiser, les reconstructions 3D générées itérativement sont appelées des demi-cartes. A chaque itération, la structure obtenue par l'itération devient la nouvelle référence qui se voit à nouveau appliquer un filtre passe-bande qui est progressivement diminué pour laisser passer les courtes fréquences spatiales. Ce procédé permet d'accéder aux données de haute résolution progressivement et évite l'accumulation du bruit au fur et à mesure des itérations. L'affinement se déroule de manière automatique et le programme arrête son affinement lorsque la résolution estimée ne s'améliore plus. Dans ce cas, l'étape finale est la fusion des deux demi-cartes afin de générer la structure finale de l'étape d'affinement. Pour cette étape j'ai utilisé les logiciels Relion, cisTEM, EMAN2, cryoSPARC et cryoSPARC 2 (en version bêta).



**Figure 65** : Exemple de carte affinée pour le complexe ERR $\alpha$ -ADN-PGC-1 $\alpha$  (cisTEM). Cette carte n'est pas filtrée pour enlever le bruit de haute résolution. Il s'agit de la carte brute.

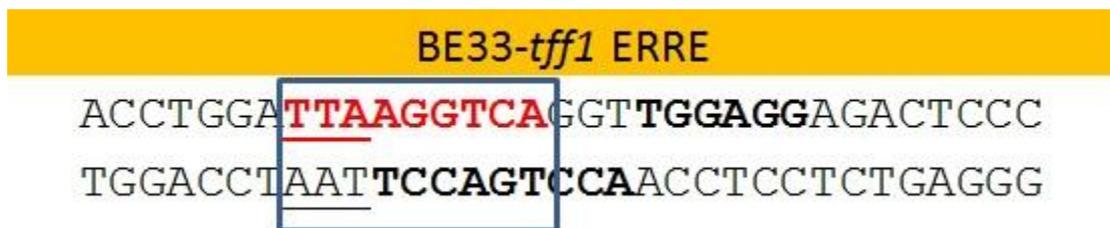
### 3. Résultats et discussion

Les complexes avec ERR $\alpha$  étant petits et asymétriques, la détermination de la structure a présenté des difficultés car les logiciels et méthodes actuellement développés pour la cryo-ME sont plus adaptés à l'étude de grands complexes tels que le ribosome par exemple. Après avoir réalisé plusieurs jeux de données à haute résolution de plusieurs complexes comprenant ERR $\alpha$ , j'ai utilisé et testé la plupart des logiciels de traitement de données actuellement disponibles, et ce dès leur sortie en version bêta pour les plus récents d'entre eux. Pour ces derniers, j'ai réalisé quelques retours et rapports de bugs aux développeurs. Pour progresser dans mon traitement de données, j'ai combiné ces différents logiciels, car utilisés seuls, ils ne permettent pas de mener à bien une reconstruction 3D dans mon cas d'étude. J'ai utilisé par exemple pour le prétraitement, MotionCor2, Gctf, CTFFind4, et pour le traitement, cisTEM, RELION, EMAN-2, IMAGIC, Sphire, cryoSPARC et cryoSPARC 2.

### 3.1 Complexe ERR $\alpha$ -ADN BE33-*tff1*ERRE

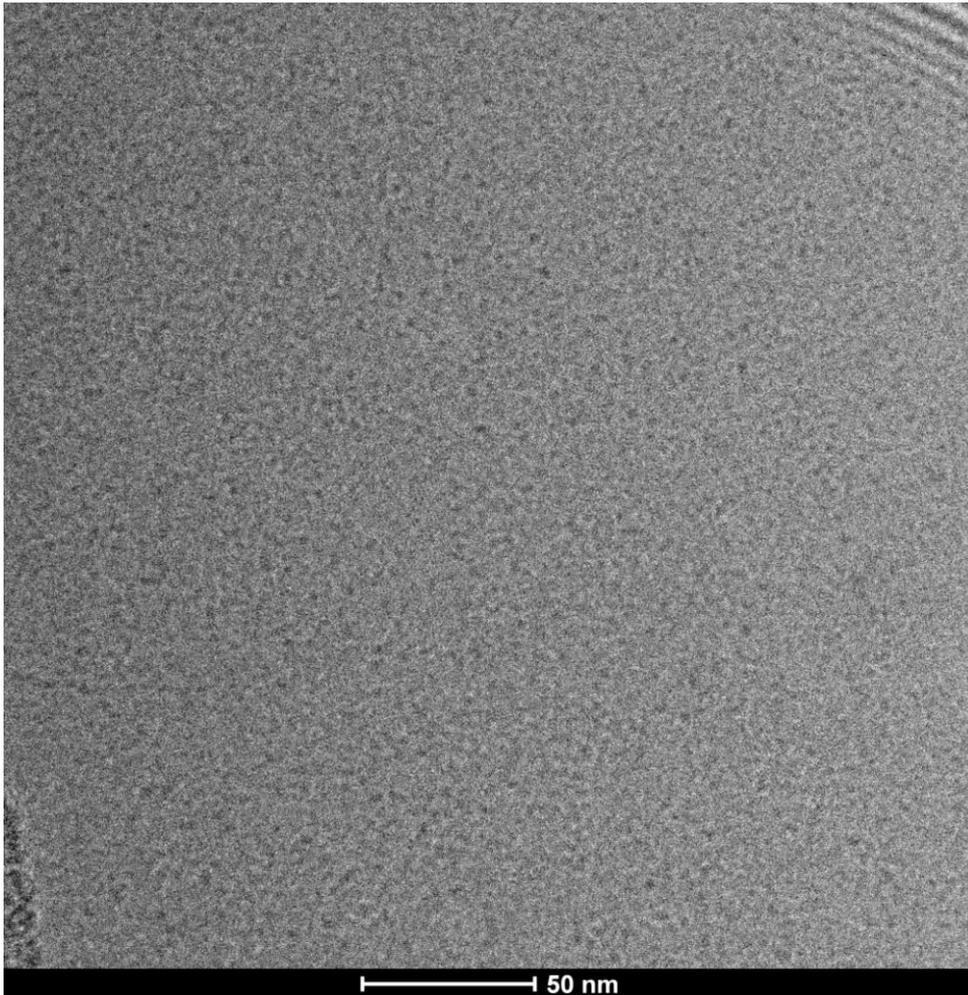
#### 3.1.1 Préparation d'échantillons

Ce complexe est le premier complexe sur lequel j'ai travaillé et a permis de faire les premières étapes d'optimisation de la préparation des échantillons et des grilles.



**Figure 66** : Séquence de l'ADN BE33-*tff1* avec l'élément de réponse ERRE encadré.

Les premières préparations sont faites dans un tampon TRIS pH7. On peut voir les particules sur un microscope électronique Polara à 100 kV mais j'ai également constaté en observant une grille de tampon seul qu'il y avait un fond très bruité qui est totalement indépendant de l'échantillon. Par la suite d'autres tests ont été réalisés avec un tampon HEPES pH 7,5, ce qui a permis de considérablement améliorer la qualité des micrographes avec moins de résidus liés au tampon.



**Figure 67** : Micrographe de tampon Tris pH7 seul. On note la présence de nombreux résidus d'environ 5nm pouvant être facilement confondus avec de petites protéines. Un complexe ERR-ADN faisant entre 8 et 11 nm de largeur en fonction de son orientation.

La concentration en sel a aussi été optimisée pour des variations entre 60 et 150 mM KCl. D'un point de vue de l'imagerie, l'augmentation du sel diminue le contraste, cependant la diminution du sel favorise les agrégations pour des raisons électrostatiques sur la protéine. Pour les différents complexes ERR $\alpha$ -ADN une concentration de 100 mM de KCl a été sélectionnée car elle représente un bon compromis entre contraste et absence d'agrégation. Cependant il restait encore à ce stade un problème majeur : la distribution aléatoire et non reproductible des particules. En effet, une série de grilles faites à quelques minutes d'intervalles dans les mêmes conditions et avec le même échantillon n'ont pas du tout le même aspect une fois dans le microscope. Ce problème est également constaté sur une même grille, où l'aspect de l'échantillon peut complètement changer entre différentes régions de la même grille.



**Figure 68** : Micrographe illustrant les problèmes de distributions des particules. On peut voir ici que la majorité des particules sont présente sur le bord inférieur du micrographe.

Ces problèmes rendent plus difficile l'optimisation des paramètres, car le côté aléatoire de la répartition de l'échantillon crée une difficulté majeure. On ne sait pas précisément si les changements observés sont dus aux changements de conditions ou juste au hasard. Pour ces raisons, le criblage des grilles a pris plus de temps, car l'ensemble de chaque grille est vérifié puisqu'une région n'est pas forcément représentative de l'ensemble de la grille. Dans un premier temps pour corriger ce problème, j'ai fait varier les conditions de glow discharge (Elmo cordouan) et les paramètres du plongeur (Vitrobot) afin de faire varier les propriétés hydrophiles du carbone et l'épaisseur de la glace. J'ai observé que dans une glace trop fine, il n'y a aucune particule, elles ont tendance à se placer sur les zones de carbone et non dans les trous où l'on procède aux acquisitions des images. Avec une glace épaisse et moyenne la distribution est meilleure car des particules sont présentes dans les trous, mais cela ne corrige pas les problèmes de distribution et ce, même en utilisant différents types de grilles Quantifoil. Il y a cependant eu une amélioration en utilisant des grilles faites maison avec du carbone continu. La présence de carbone et donc l'absence d'une

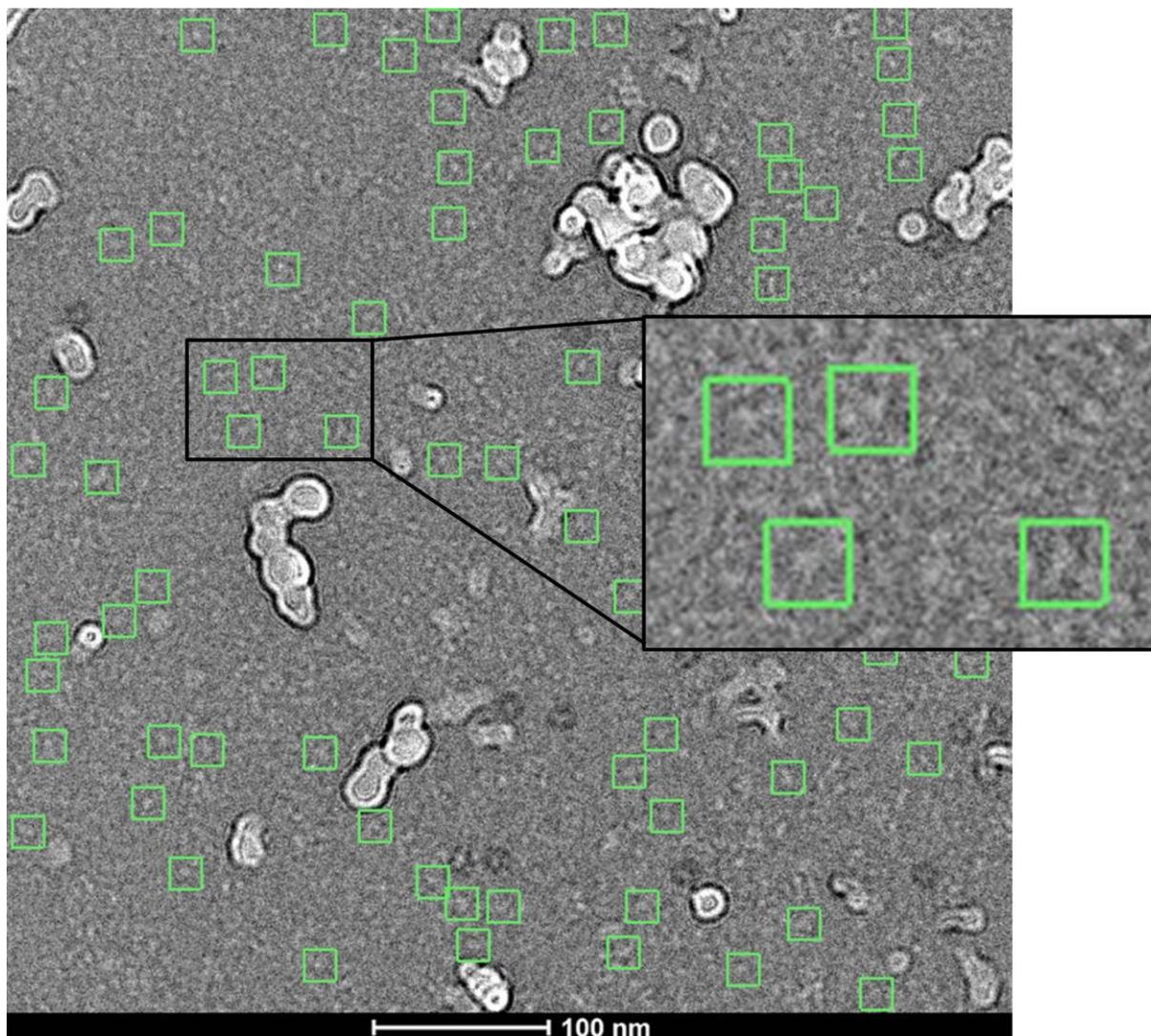
seconde interface échantillon/air dans les doubles ménisques des trous du carbone aide à la distribution des particules. Le principal problème de cette méthode est la diminution du contraste des particules, d'autant plus que les complexes d'intérêts sont petits et ont de fait pas beaucoup de contraste.

J'ai essayé une autre méthode permettant de conserver le contraste tout en modifiant les propriétés de distribution de l'échantillon. Au lieu de distribuer l'échantillon sur un carbone continu, on peut ajouter un détergent à la composition du tampon. Cependant cette méthode doit être utilisée avec précaution : il est nécessaire de travailler avec des concentrations en détergent très faibles, nettement inférieures à la valeur de la CMC (Concentration Micellaire Critique) pour ne pas ajouter des artefacts sur les micrographes (Gewering et al., 2018). Des premiers essais ont été réalisés avec du CHAPS qui a une CMC de 6 mM. Le CHAPS, utilisé avec une concentration de 1mM entraîne une forte diminution de la qualité des images, le détergeant, même à concentration très faible inférieur à sa CMC crée des artefacts. Cette méthode n'étant pas concluante, j'ai tout de même pu réaliser une acquisition sur le microscope Polara d'une grille correcte mais non reproductible. D'autres tests, visant à éliminer la présence de détergent tout en conservant une distribution homogène ont été réalisés avec de l'amphipols à 0.025%. Cette molécule est un polymère amphiphile qui est souvent utilisée pour la solubilité et la distribution des protéines membranaires lors d'expériences de cryo-ME. Cependant son utilisation n'a pas montré d'améliorations significatives du comportement du complexe ERR-ADN étudié.

---

### 3.1.2 Acquisitions

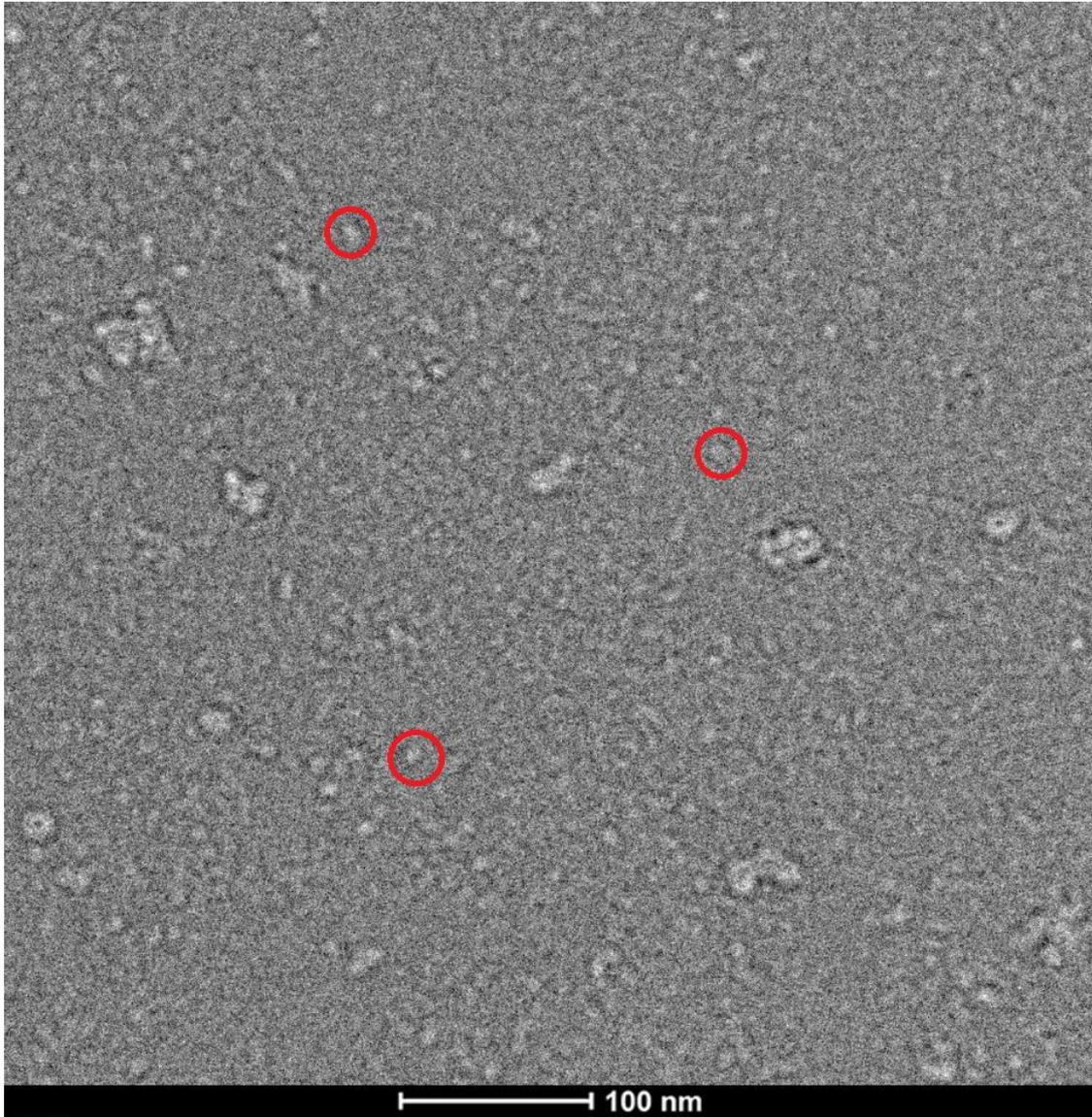
J'ai réalisé une première acquisition pour le complexe  $ERR\alpha$ -ADN BE33-*tff1* ERRE. Elle est réalisée sur le microscope Polara à 100 kV équipé d'une caméra Falcon I. Un grandissement de 78 000 fois, un temps d'exposition de 2 secondes pour une dose totale de 20  $e/\text{\AA}^2$ . Pour ce jeu de données de 700 micrographes, enregistré grâce au logiciel d'acquisition automatique EPU, un total de 17 000 particules a été sélectionné avec le module semi-automatique de EMAN2.



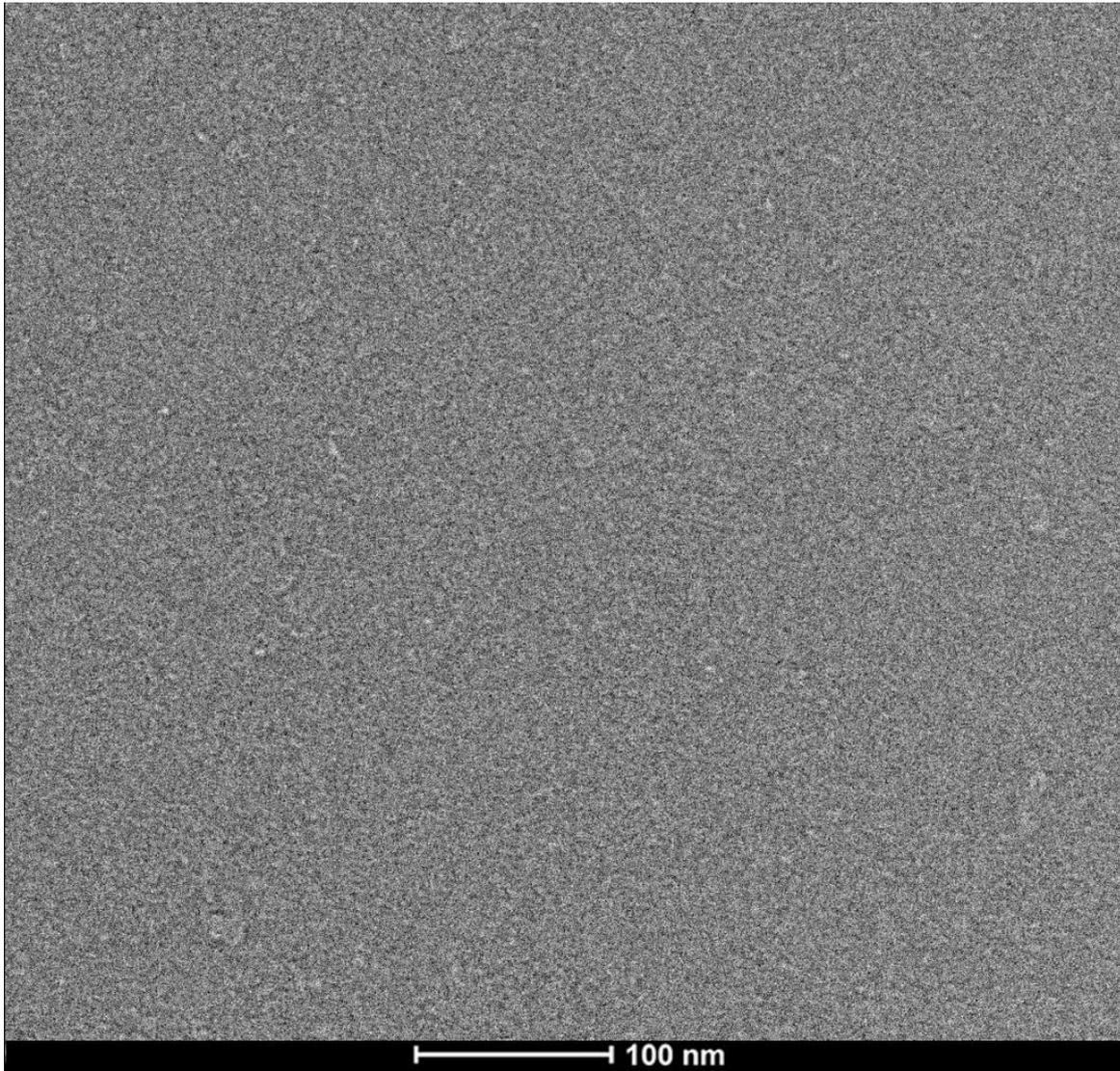
**Figure 69** : Micrographe de l'acquisition du complexe  $ERR\alpha$ -ADN BE33-*tff1* ERRE sur le microscope Polara avec une caméra Falcon I, défocalisation de  $-3,5\mu\text{m}$ , voltage de 100 kV, grandissement 78 000 fois. La sélection des particules est matérialisée par les carrés verts. Les objets de grande taille blancs sont des contaminants présents sur la grille par-dessus la glace. Pour reconnaître les particules on peut se fier à leur taille et à la présence d'une structure interne visible. Un contaminant, de la même taille que la particule sera plus lisse qu'une protéine, qui elle contient du détail. Sur les micrographes, les particules sont noires, ici elles apparaissent blanche car le contraste est inversé dans le logiciel EMAN2. Il s'agit d'un héritage du traitement par film photographique qui utilise donc des négatifs où les particules étaient blanches. Dans les autres logiciels tels que Relion ou cisTEM on peut spécifier si les particules sont blanches ou noires sur les micrographes.

Des essais sur le Titan Krios ont été fait en vue d'une acquisition à 300kV. Au moment de l'acquisition, le Titan Krios de l'Institut était équipé d'une caméra Falcon II et d'un correcteur Cs. Cette caméra n'était malheureusement pas suffisamment performante pour l'acquisition de complexes aussi petits que les miens. On pouvait voir les particules avec une très forte défocalisation, mais on perdait ainsi les fréquences de la haute résolution. Lorsqu'on appliquait une défocalisation comprise entre  $-1,5$  et  $-4,5\mu\text{m}$  on ne distinguait pas les particules, le contraste de

l'objet étant insuffisant. Par conséquent, j'ai rencontré une limitation technique qui a restreint le traitement d'images à des images produites sur le Polara de l'Institut, à une tension de 100 kV et une caméra Falcon I, ce qui risquait de limiter la résolution à une dizaine d'angströms de résolution comme pour les travaux précédents sur les complexes RXR/VDR et USP/EcR cités précédemment, qui ont été enregistrés sur ce même équipement.



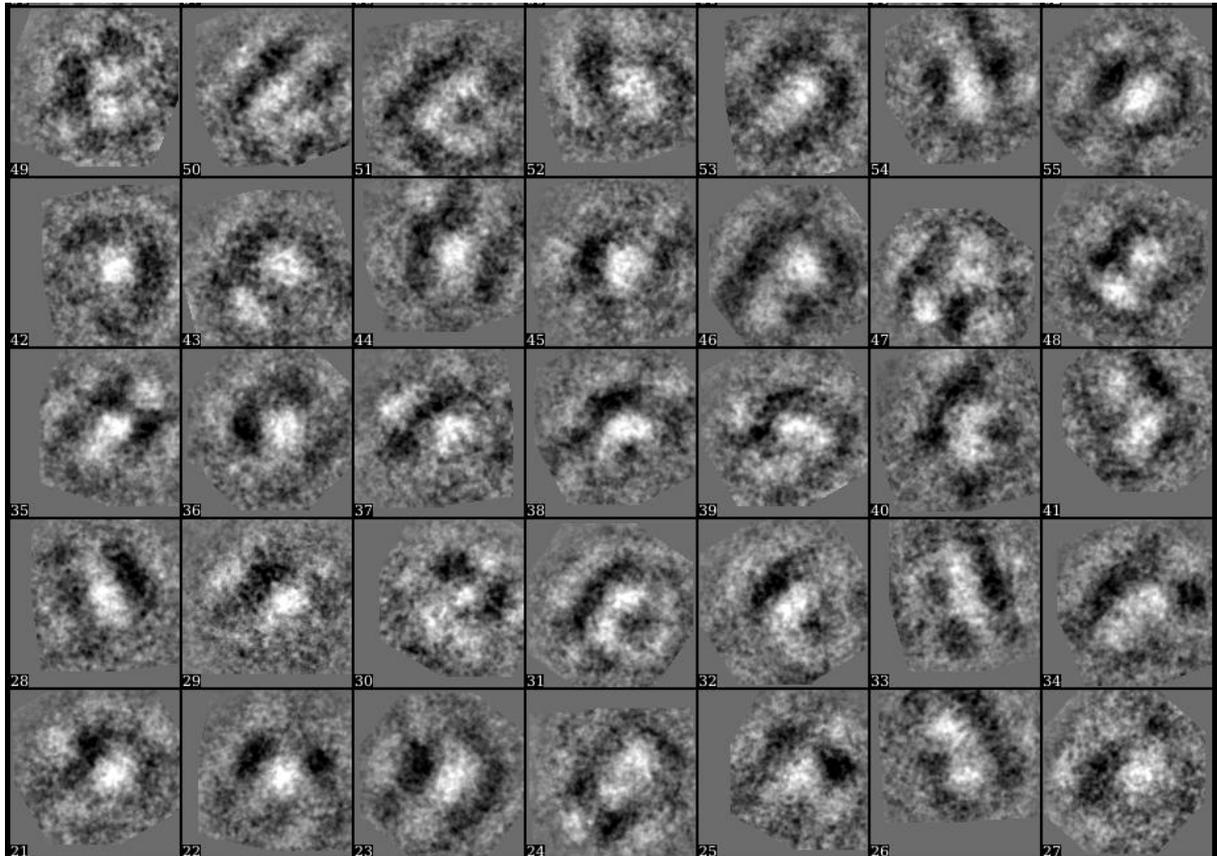
**Figure 70** : Micrographe réalisé sur le microscope Titan Krios du complexe  $ERR\alpha$ -ADN BE33-*tff1* ERRE avec une défocalisation de  $-12\ \mu\text{m}$ . A cette défocalisation, les détails des particules sont totalement perdus mais le contraste est augmenté. On voit quelques contaminants, et les particules (exemple en rouge) qui ont une taille d'environ 10 nm.



**Figure 71** : Micrographe réalisé sur le microscope Titan Krios du complexe  $ERR\alpha$ -ADN BE33-*tff1* ERRE avec une défocalisation de  $-3,5 \mu\text{m}$ . A cette défocalisation nous n'avons plus suffisamment de contraste pour distinguer les particules. Nous voyons seulement quelques contaminants.

### 3.1.3 Classification 2D

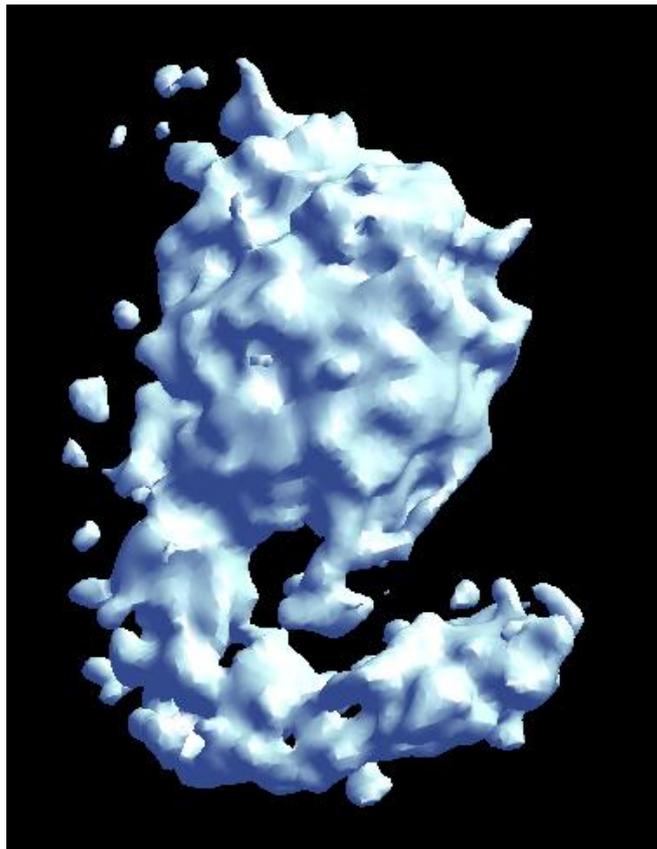
Une classification 2D a été faite à l'aide du logiciel EMAN2 à partir des 17000 particules sélectionnées avec 25 cycle itératifs et 100 classes demandées. Suite à cette classification 15000 particules ont été retenues.



**Figure 72** : Classes 2D du complexe ERR $\alpha$ -ADN BE33-*tff1* ERRE (EMAN2).

### 3.1.4 Reconstruction 3D

La détermination 3D a débuté par la génération d'une reconstruction 3D initiale sur EMAN2, puis une série de cycles d'affinement 3D. Une première structure à faible résolution et comprenant un peu de bruit a pu être obtenue.



**Figure 73** : Première structure du complexe ERR $\alpha$ -ADN BE33-*tff1* ERRE après quelques cycles d'affinement (EMAN2). On reconnaît les domaines LBD en haut et l'ADN allongé en bas.

Ce premier résultat à basse résolution permet déjà de décrire la topologie globale du complexe. La grande masse en haut correspond en taille au dimère des LBD, puis le domaine charnière (hinge) qui connecte les DBD sur l'ADN qui correspondrait à la densité allongée en bas.

---

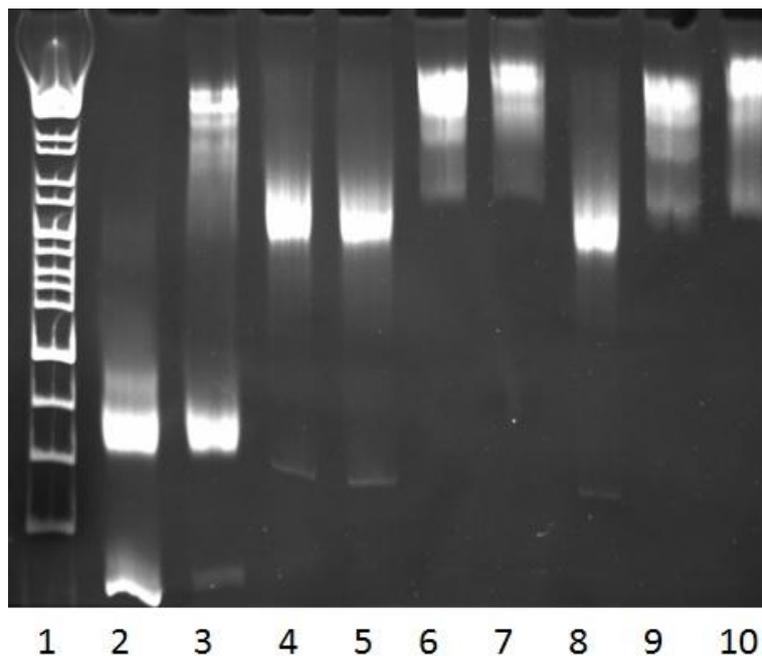
### 3.1.5 Limites-problèmes

Continuer les cycles d'affinement a pour conséquence de faire disparaître progressivement la densité du fragment d'ADN. Cela suggère qu'il y a une hétérogénéité dans le jeu de données avec une partie des complexes qui ne sont peut-être pas liés à leur ADN. Je n'ai donc pas poursuivi l'affinement de cette structure. Une classification 3D avec Relion par exemple est probablement nécessaire pour continuer avec ce jeu de données, mais un jeu de données plus grand serait nécessaire pour cela.

## 3.2 Complexe ERR-Di Nucléosome et ERR-Nucléosome

### 3.2.1 Préparation d'échantillons

Ces complexes sont les deuxièmes complexes sur lesquels j'ai travaillé, en parallèle avec les dernières optimisations apportées aux complexes  $ERR\alpha$ -BE33-*tff1* ERRE. Pour la préparation des échantillons, j'ai utilisé comme base la méthode de préparation, les dernières améliorations pour le complexe  $ERR\alpha$ -BE33-*tff1* ERRE. Au niveau du tampon les premiers tests ont donc été avec de l'HEPES, et on a également essayé le tampon EPPS pH 7,5. Il a apporté une légère amélioration au niveau de la distribution dans le cas de ce type de complexes avec du nucléosome, mais cette amélioration ne se vérifie pas pour le complexe  $ERR\alpha$ -BE33-*tff1* ERRE pour lequel deux tampons donnent des résultats équivalents. Le détergent CHAPS a été conservé également dans le tampon EPPS à 0.0025% pour des raisons de distribution, les artefacts créés par le détergeant sont moins gênants ici car le nucléosome a plus de contraste car il a plus d'ADN que les complexes ERR-ADN et il a une masse plus importante. Pour ces complexes la concentration en sels et surtout en magnésium représente un facteur limitant à la formation du complexe. En effet de légères variations changent du tout au tout la formation du complexe observé sur gel. L'intervalle de concentration du magnésium est très petit, nous utilisons 0,2 mM. Un manque de magnésium entraîne la dissociation du complexe, ERR ne se fixe pas sur l'ADN du nucléosome. Au contraire un léger excès de magnésium provoque l'agrégation des nucléosomes. Pour la distribution sur les grilles de ce complexe, j'ai rencontré les mêmes problèmes de distributions que pour les complexes  $ERR\alpha$ -ADN. Cependant, contrairement à ces derniers, le changement de grilles permet de faire d'importantes optimisations. En effet la distribution aléatoire observée avec des grilles Quantifoil R2/2 devient reproductible avec des grilles Quantifoil R1.2/1.3.



**Figure 74** : Gel natif 5% EMSA de plusieurs complexes NCP-ERR $\alpha$ . La colonne 1 est le marqueur de poids moléculaires. La colonne 2 est un nucléosome seul, la colonne 3 est le complexe NCP-ERR $\alpha$ , les colonnes 4 et 5 sont un Di-NCP avec un ADN 314pb. La colonne 6 est le complexe Di-NCP-ERR $\alpha$  avec un ADN 314pb. La colonne 7 est le complexe Di-NCP-ERR $\alpha$ -PGC-1 $\alpha$  avec un ADN 314pb. La colonnes 8 est un Di-NCP avec un ADN 304pb. La colonne 9 est le complexe Di-NCP-ERR $\alpha$  avec un ADN 304pb. La colonne 10 est le complexe Di-NCP-ERR $\alpha$ -PGC-1 $\alpha$  avec un ADN 304pb.

Ces observations sont valables pour l'ensemble des complexes produits avec des nucléosomes, à savoir un complexe avec ERR fixé sur l'ADN qui relie un di-nucléosome, et un complexe avec ERR directement fixé sur un ADN.

## Linker DiNCP21 ERRE/ERE (314)

```

TCAAGGTCACACTGACCTTGA
AGTTCCAGTGTGACTGGAACT

```

## Linker DiNCP11 ERRE (304)

```

TCAAGGTCACACA
AGTTCCAGTGT

```

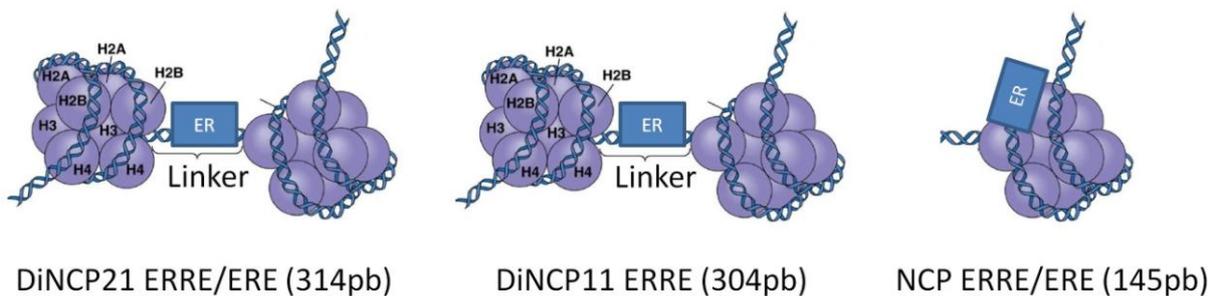
## NCP ERRE/ERE (145)

```

TCAAGGTCACACTGACCTTGA
AGTTCCAGTGTGACTGGAACT

```

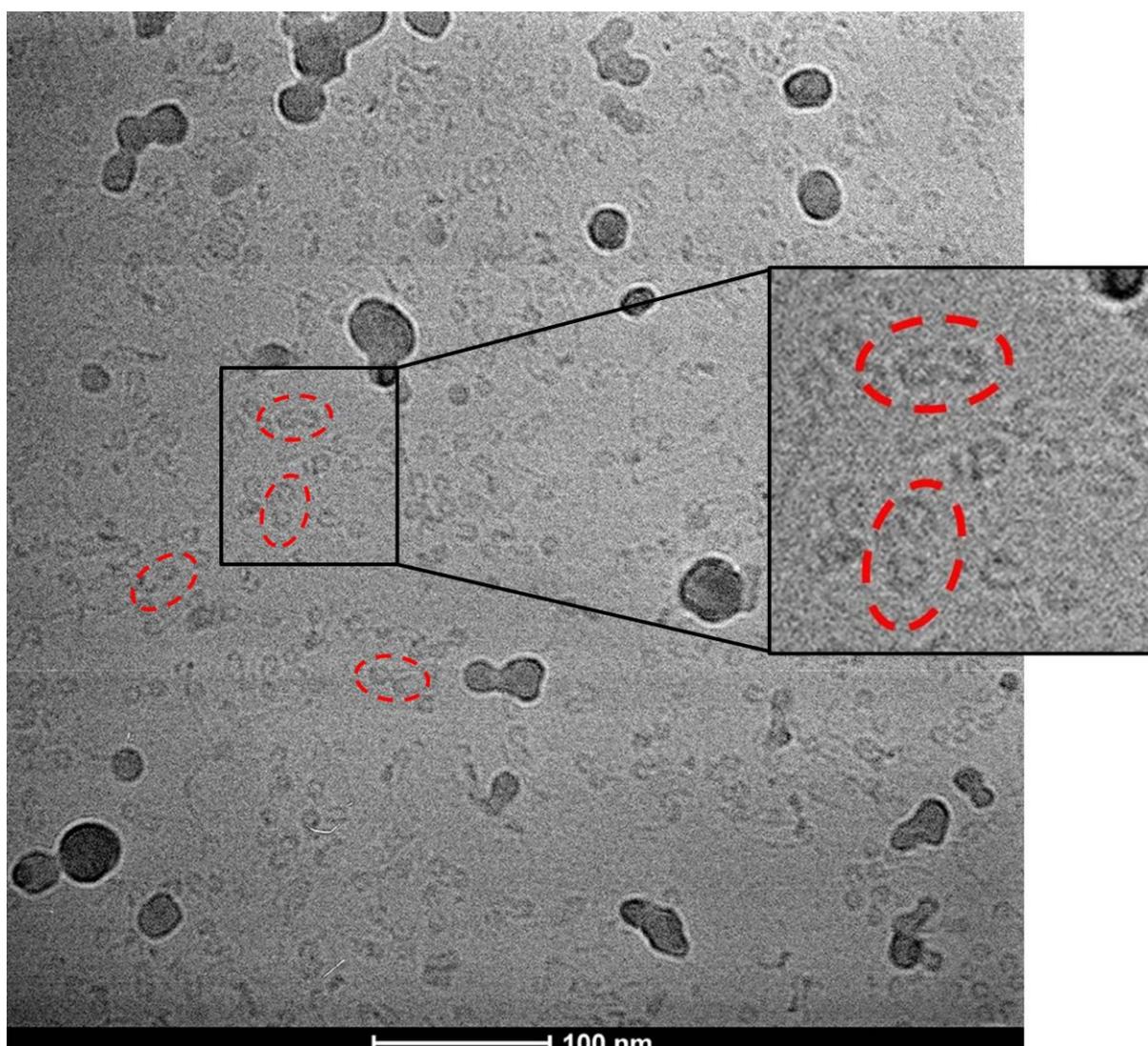
**Figure 75** : Séquence des éléments de réponses utilisés dans les complexes nucléosomes. La première séquence est la séquence du lien entre les deux nucléosomes où est placé l'élément de réponse ERRE/ERE, en 5' et 3'. La séquence continue avec un nucléosome complet de part et d'autre (147 pb pour chaque nucléosome, donc 314 paires de bases au totale). La seconde séquence est la séquence du lien entre les deux nucléosomes où est placé l'élément de réponse ERRE, en 5' et 3'. La séquence continue avec un nucléosome complet de part et d'autre (147 pb pour chaque nucléosome, donc 304 paires de bases au totale). La troisième séquence est la séquence au milieu d'un ADN de nucléosome où est placé l'élément de réponse ERRE/ERE, en 5' et 3' la séquence continue avec 62 pb (145 paire de bases au total).



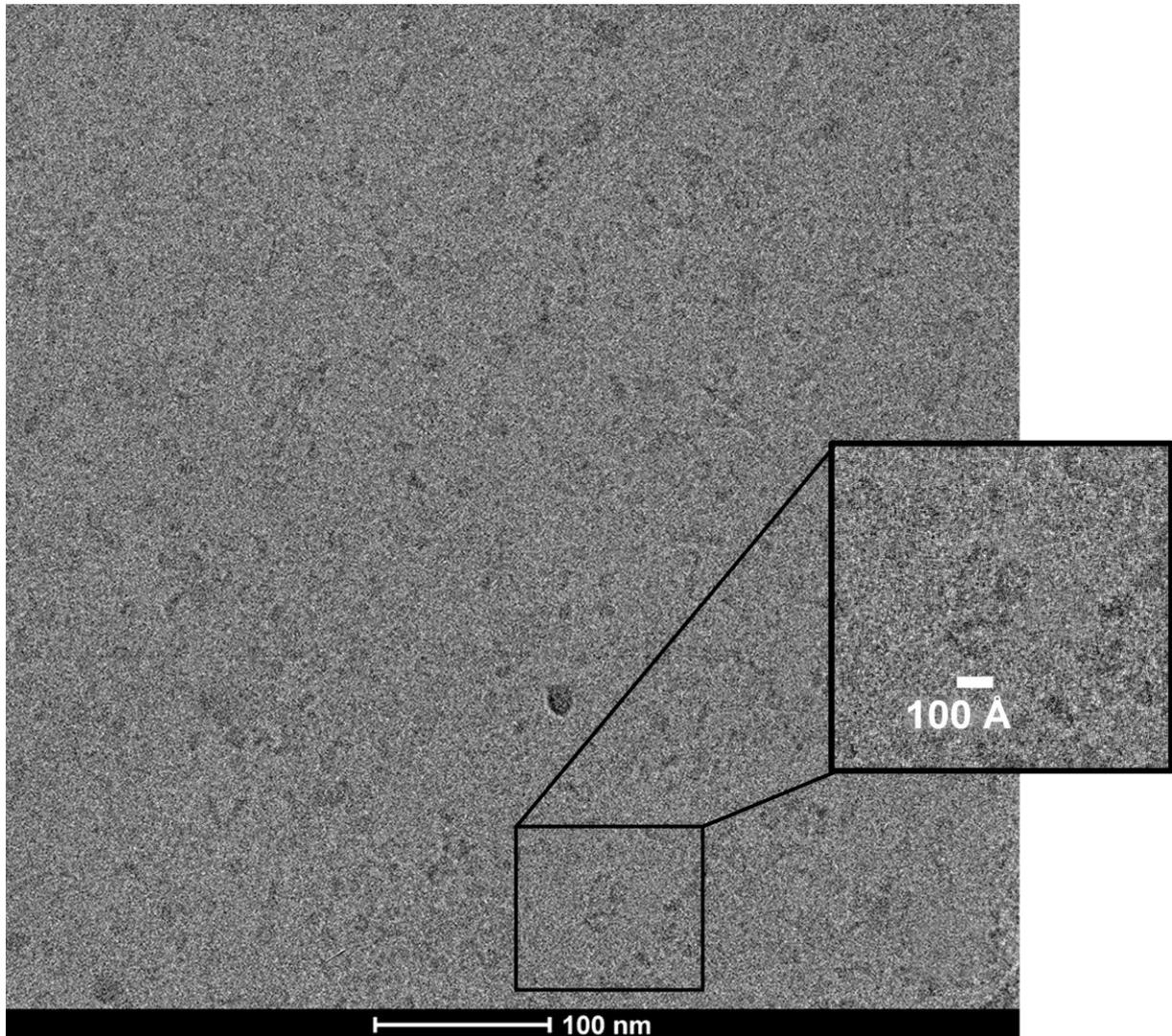
**Figure 76** : Schéma de la position de l'élément de réponse (ER) dans les complexes avec nucléosomes. (image adaptée de Caputi and Romualdi, 2017).

Le principal problème pour ces complexes n'a été résolu qu'en partie au cours de ma thèse. Il s'agit de la stabilité du complexe dans des conditions cryogéniques. En effet, les gels natifs montrent une

très bonne stabilité du complexe en solution, par contre une fois cryogénisé, les complexes ERR-nucléosomes se dissocient. Il y a des nucléosomes, des petites particules qui correspondent au domaine homodimère LBD d'ERR d'un point de vue taille, par contre aucun complexe entier n'est observé. On note également dans certains cas le début d'une ouverture de l'ADN aux extrémités. Pour remédier à ce problème on a essayé la méthode de réticulation de l'échantillon (cross-link) avec un agent de réticulation tel que le glutaraldéhyde qui permet de lier chimiquement les protéines entre elles pour augmenter la stabilité du complexe. C'est une opération irréversible. J'ai aussi utilisé du formaldéhyde qui lui est plus adapté pour lier chimiquement un ADN à une protéine. Dans ce cas on observe une amélioration, où environ 10% des complexes observés sont entiers. Il s'agit cependant encore d'un résultat fragile qui mériterait d'avantage d'optimisation.



**Figure 77** : Micrographe du complexe Di-NCP-ERR $\alpha$  qui a un lien de 11 paires de bases entre les deux nucléosomes. Image réalisée sur le microscope Polara avec une caméra Falcon I, défocalisation de  $-3,5\mu\text{m}$ , voltage de 100 kV, grandissement 78 000 fois. Exemples de Di-NCP entouré en rouge.

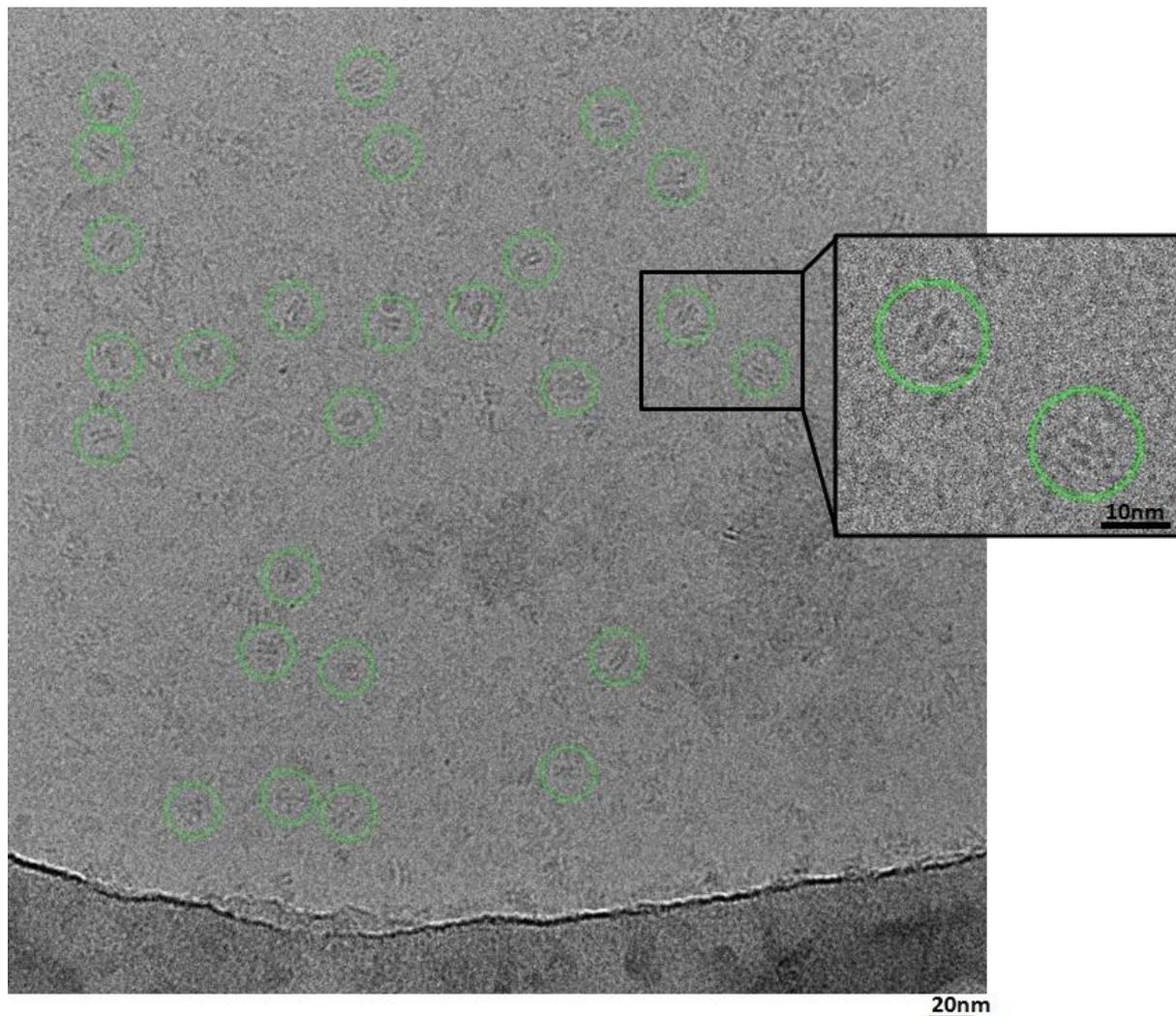


**Figure 78** : Micrographe du complexe Di-NCP-ERR $\alpha$  qui a un lien de 21 paires de bases entre les deux nucléosomes. Image réalisée sur le microscope Polara avec une caméra Falcon I, défocalisation de -2,5 $\mu$ m, voltage de 100 kV, grandissement 78 000 fois. L'exemple de particules encadrées montre un complexe complet avec ERR lié au linker entre les deux nucléosomes.

### 3.2.2 Acquisitions

J'ai effectué des premiers tests sur le Titan Krios de l'Institut avec la caméra Falcon II, ont montré que les nucléosomes sont visibles, contrairement à ERR, mais le contraste et donc le rapport signal/bruit est tout de même trop faible pour un traitement d'images. Plus tard, lors de deux mises à jour successives du Titan Krios j'ai pu réaliser des premiers essais d'acquisitions pour deux complexes de

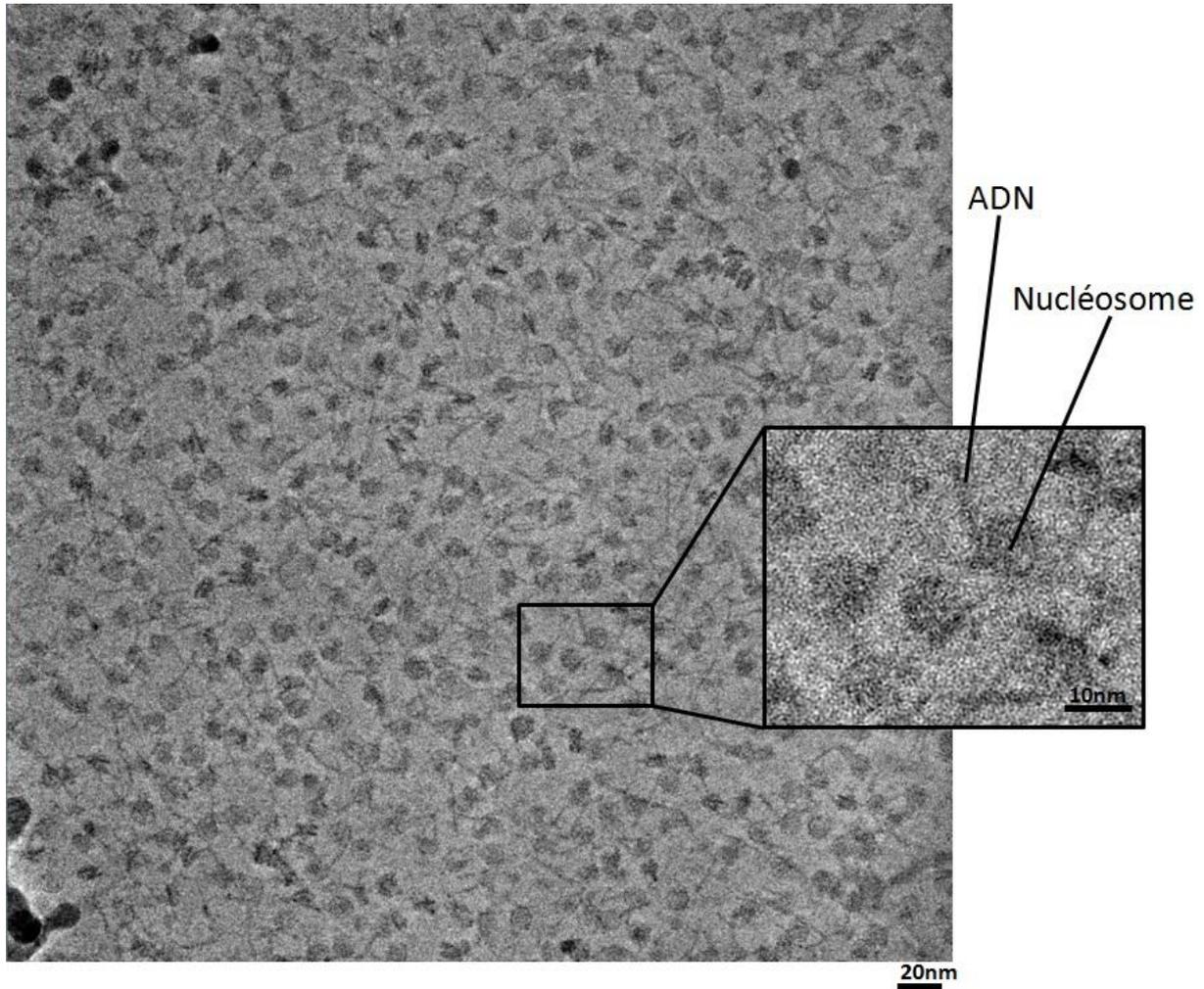
ERR fixé à un nucléosome. La première mise à jour importante est l'installation d'une caméra à détection directe d'électrons de nouvelle génération, la Gatan K2 Summit. La première acquisition a été réalisée avec cette caméra et pour la première fois je pouvais voir clairement les particules à un voltage de 300 kV sur un microscope très stable. Un grandissement de 105 000 fois, un temps d'exposition de 7,1 secondes pour une dose totale de  $50 \text{ é}/\text{Å}^2$  répartie entre 28 frames. Pour ce jeu de données de 1042 micrographes, enregistré grâce au logiciel d'acquisition automatique EPU, un total de 120 000 particules a été sélectionné avec le module automatique de Relion 1.4.



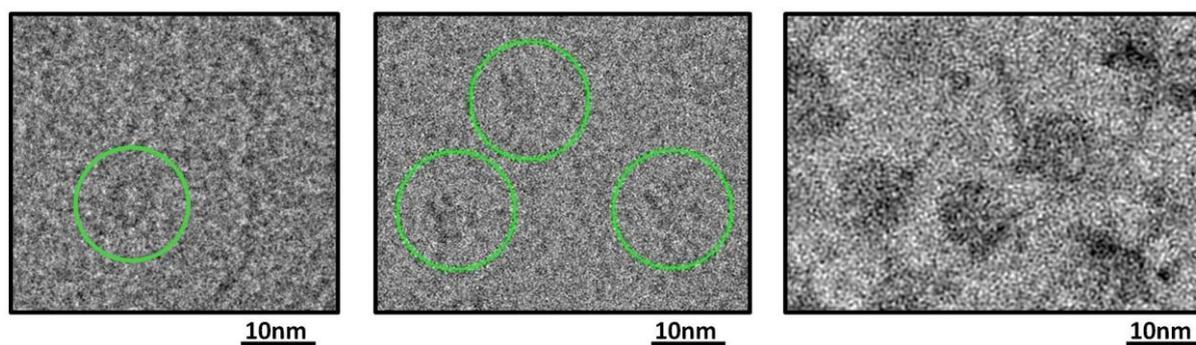
**Figure 79** : Micrographe du complexe NCP-ERR $\alpha$ . Image réalisée sur le microscope Titan Krios avec une caméra Gatan K2 summit, défocalisation de  $-3 \mu\text{m}$ , voltage de 300 kV, grandissement 105 000 fois. La sélection des particules est matérialisée par les cercles verts (Relion 1.4).

La seconde mise à jour importante est l'installation d'une phase plate. Deux mois après la première acquisition, une seconde acquisition a été réalisée avec la Volta phase plate et la caméra Gatan K2 summit et l'amélioration du contraste est encore plus impressionnant. Elle est réalisée sur le

microscope Titan Krios à 300 kV équipé d'une caméra Gatan K2 Summit. Un grandissement de 105 000 fois, un temps d'exposition de 7,1 secondes pour une dose totale de 50  $\text{é}/\text{Å}^2$  répartie entre 28 frames. Pour ce jeu de données de 1211 micrographes, enregistré grâce au logiciel d'acquisition automatique SerialEM, un total de 232 000 particules a été sélectionné avec le module automatique de Relion 1.4.



**Figure 80** : Micrographe du complexe NCP-ERR $\alpha$ . Image réalisée sur le microscope Titan Krios avec une caméra Gatan K2 summit et une Volta phase plate, défocalisation de  $-0,5 \mu\text{m}$ , voltage de 300 kV, grandissement 105 000 fois. On voit ici très clairement les particules et l'ADN lorsqu'il est débobiné.

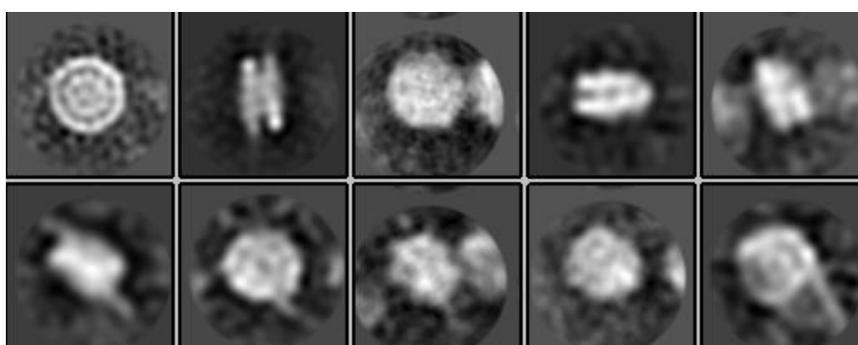


**Figure 81** : Comparatif de l'évolution de l'équipement d'acquisition durant la thèse. Les 3 images ont été faites sur le microscope Titan Krios. Il s'agit du même échantillon, à savoir le complexe NCP-ERR $\alpha$ . Les 3 images ont été faites avec un voltage de 300 kV. La première image est acquise avec une caméra Falcon II à un défocus de  $-4,5 \mu\text{m}$ . La seconde image est acquise avec une caméra Gatan K2 Summit et un filtre d'énergie (GIF) à un défocus de  $-2,5 \mu\text{m}$ . La troisième image est acquise avec une caméra Gatan K2 Summit, un filtre d'énergie (GIF) et une Volta phase plate à un défocus de  $-0,5 \mu\text{m}$ . On constate l'amélioration du signal, d'autant plus qu'on a plus de hautes fréquences à des défocus proches de zéro.

### 3.2.3 Classification 2D

Une faible proportion de complexes contenait ERR et la dominance du nucléosome d'un point de vue contraste de la particule a fait tendre les données vers un centrage sur le nucléosome lui-même.

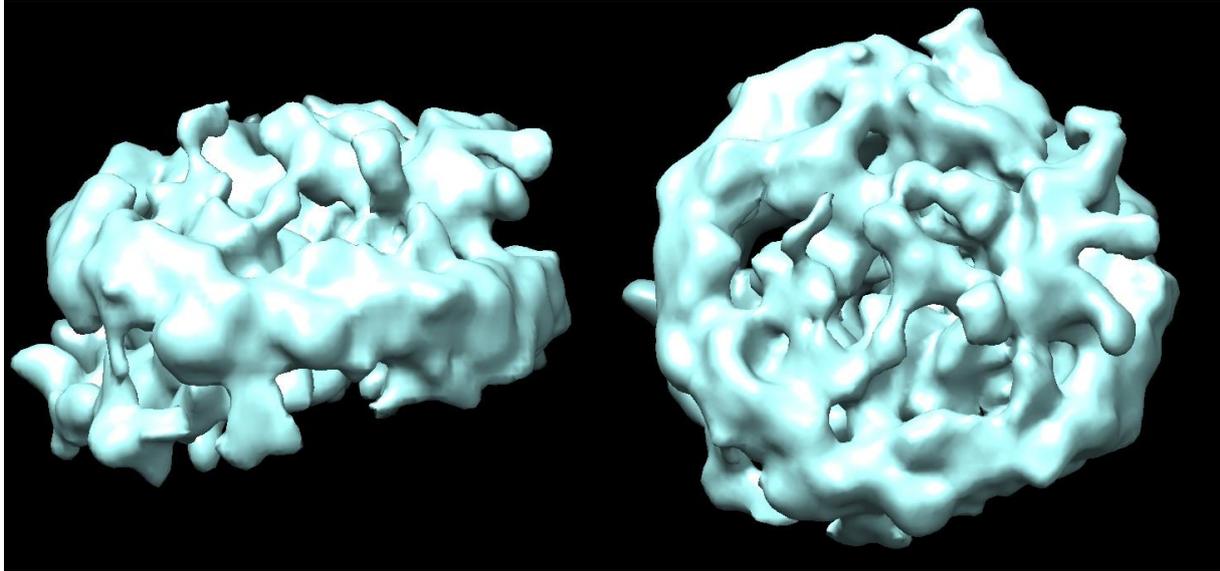
Une classification 2D a été faite à l'aide du logiciel Relion1.4 à partir des 232 000 particules sélectionnées du jeu de données acquis avec la caméra Gatan K2 Summit et la phase plate. Elle a été réalisée avec 25 cycles itératifs et 200 classes demandées. Suite à cette classification 142 000 particules ont été retenues.



**Figure 82** : Exemple de classes 2D du complexe NCP-ERR $\alpha$ . On observe certaine classe avec plusieurs particules présente comme la classe 3 et la classe 8. Ceci est dû à la densité importante de l'échantillon. Pour le traitement il est préférable d'éliminer ces particules, surtout au début du traitement.

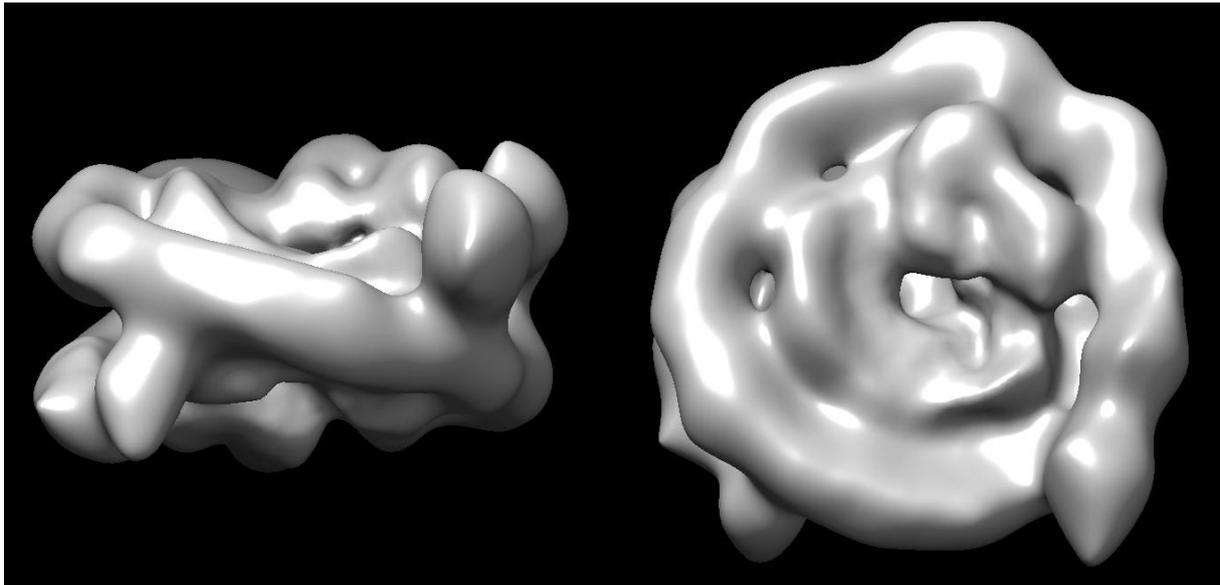
### 3.2.4 Reconstruction 3D

La version 1.4 de Relion n'est pas capable de produire une structure initiale. Pour pouvoir continuer l'affinement dans Relion, j'ai produit plusieurs structures initiales, avec les logiciels EMAN2 et CryoSPARC.



**Figure 83** : Structure initiale du complexe ERR-nucléosome généré avec cryoSPARC. On distingue déjà la partie ADN et la partie histone.

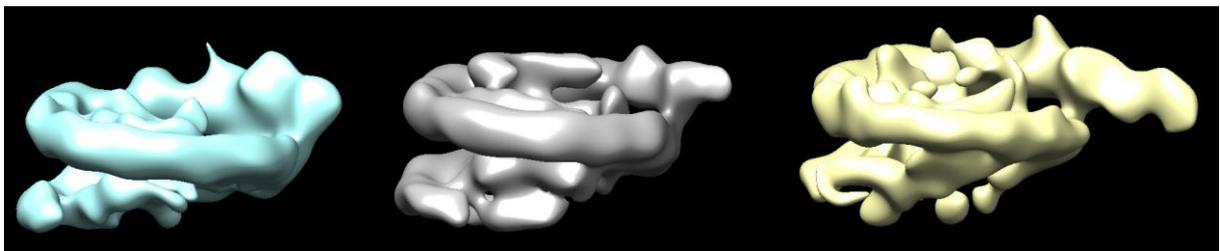
A partir de cette structure initiale, j'ai réalisé un premier affinement 3D avec Relion 1.4. On peut voir que le cœur du nucléosome est clairement identifiable mais il y a une mauvaise interprétation des données au niveau des deux extrémités d'ADN. De plus il n'y a aucune trace de la présence d'ERR sur ce premier affinement. On pouvait pourtant espérer récupérer une faible densité dans la région où il est fixé. L'absence totale de cette densité est probablement due à la trop faible proportion du complexe contenant ERR ainsi qu'à l'alignement des particules qui n'est pas encore suffisamment précis. Le nucléosome ayant plus de contraste que ERR, en plus d'être majoritaire en nombre, les classes ont tendance à se centrer sur le nucléosome. Par conséquent les ERR présents sont noyés dans le bruit.



**Figure 84** : Structure après le premier cycle d'affinement fait avec Relion 2.1. On distingue correctement l'ADN en bordure et le cœur protéique au centre.

### 3.2.5 Classification 3D

Une première classification 3D a été réalisée pour essayer de séparer les complexes complets des nucléosomes isolés. Les résultats obtenus n'ont pas permis de faire cette séparation, cependant la classification 3D a permis de séparer les nucléosomes en fonction de leurs degré d'ouverture de l'ADN.



**Figure 85** : Classification 3D pour un complexe de nucléosome-ERR (Relion). On voit différents intermédiaires de l'ouverture de l'ADN.

### 3.2.6 Limites-problèmes

La principale limite ici est la nature du complexe qui n'est pas stable avec les conditions utilisées pour la cryogénisation. Cependant il a été intéressant de poursuivre les premières étapes d'affinement pour appréhender l'utilisation et le traitement de données issues des nouvelles technologies disponibles, à savoir le mode super-résolution de la caméra Gatan K2 Summit et l'utilisation de données phase plate qui modifient les étapes de prétraitement à cause du décalage de phase, mais qui facilitent le traitement de données.

## 3.3 Complexe ERR $\alpha$ -ADN BE33-embedded ERRE/ERE

### 3.3.1 Préparation d'échantillons

La préparation de l'échantillon est faite en suivant les dernières avancées de l'optimisation du premier complexe ERR $\alpha$ -BE33-*tff1* ERRE. En m'appuyant sur les travaux menés sur les complexes avec le nucléosome, j'ai choisi ici de former un complexe avec l'agent de réticulation glutaraldéhyde dans un tampon HEPES pH7,9.



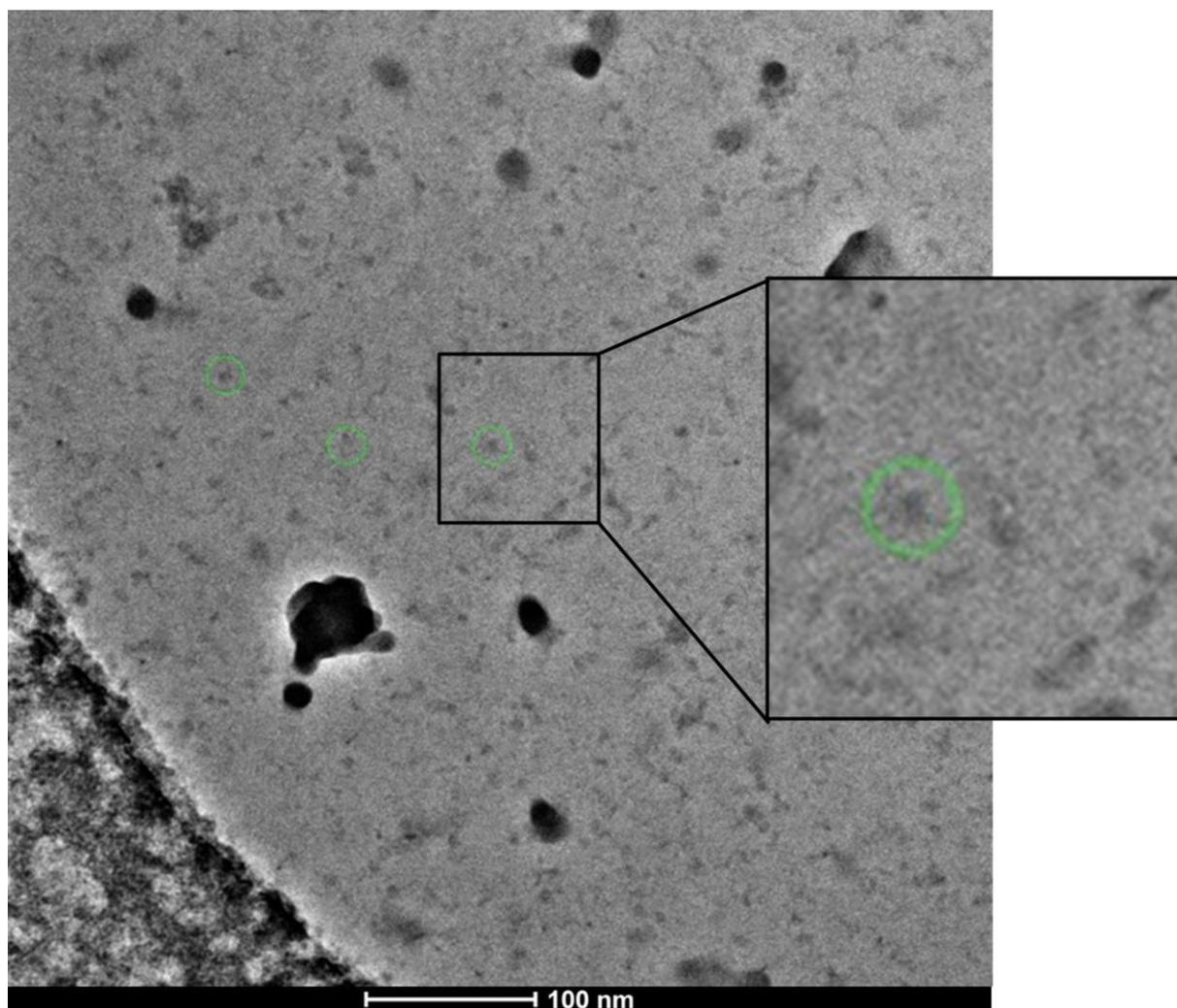
**Figure 86** : Séquence de l'ADN BE33-embedded IR3 avec l'élément de réponse ERRE/ERE encadré.

### 3.3.2 Acquisitions

Suite à l'évolution technologique du Titan Krios que j'ai utilisé durant ma thèse, j'ai procédé à l'acquisition d'un nouveau jeu de données, sur le Titan Krios cette fois d'un complexe ERR $\alpha$ -ADN BE33-embedded ERRE/ERE. Pour ce faire, j'ai utilisé la caméra Gatan K2 Summit, la volta phase plate avec un changement de position toutes les 100 minutes, un voltage de 300 kV, une défocalisation constante de -0,5 $\mu$ m, un grossissement de 105 000 fois ce qui donne une taille de pixel de 1,09 Å

(0.545 Å/px lors de l'acquisition en mode super-résolution). Les films ont été collectés avec 28 images par film, un temps total de 8,4 sec d'acquisition par film, 0,3 seconde par image de film et une dose de 7,1 é/Å<sup>2</sup>/s, soit une dose totale de 60 é/Å<sup>2</sup>.

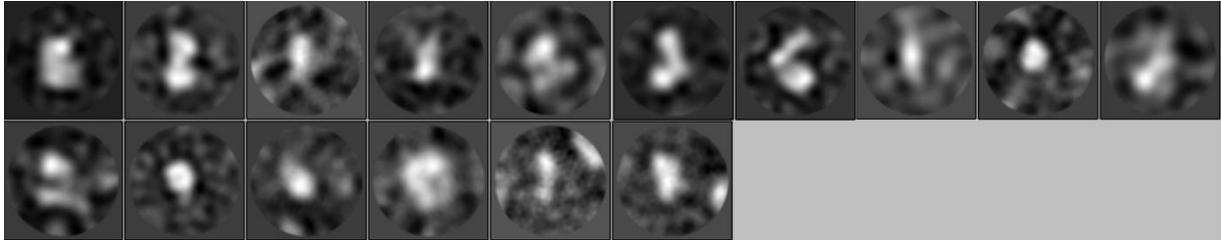
Pour ce jeu de données de 1022 micrographes, enregistré grâce au logiciel d'acquisition automatique SerialEM, un total de 38 000 particules a été sélectionné avec le module automatique de Relion2.0.



**Figure 87** : Micrographe du complexe ERR $\alpha$ -ADN. Image réalisée sur le microscope Titan Krios avec une caméra Gatan K2 summit et une Volta phase plate, défocalisation de -0,5  $\mu$ m, voltage de 300 kV, grandissement 105 000 fois. On voit ici une très bonne amélioration du contraste, en absence de Volta phase plate on ne peut pas voire des particules aussi petites correctement à 300 kV.

### 3.3.3 Classification 2D

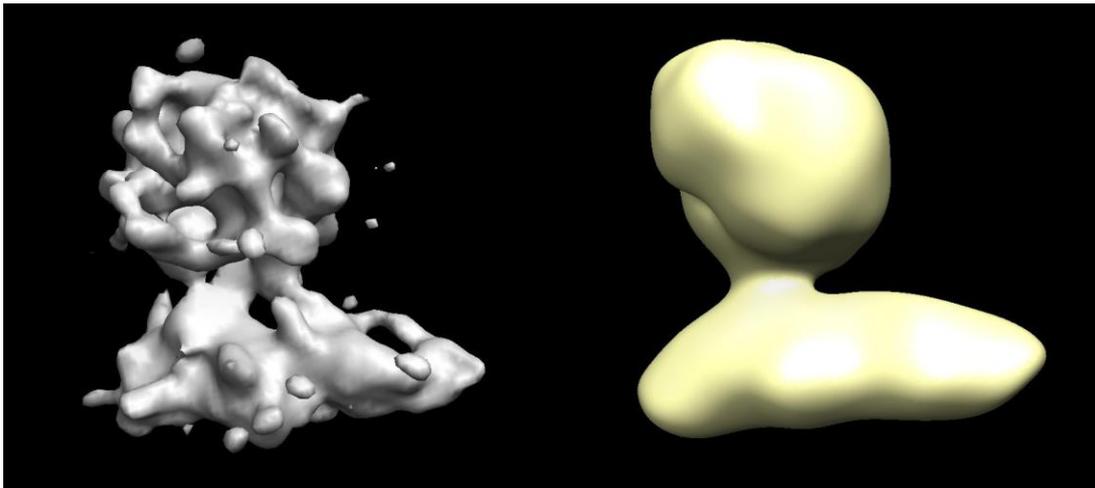
Une classification 2D a été faite à l'aide du logiciel Relion2.0.



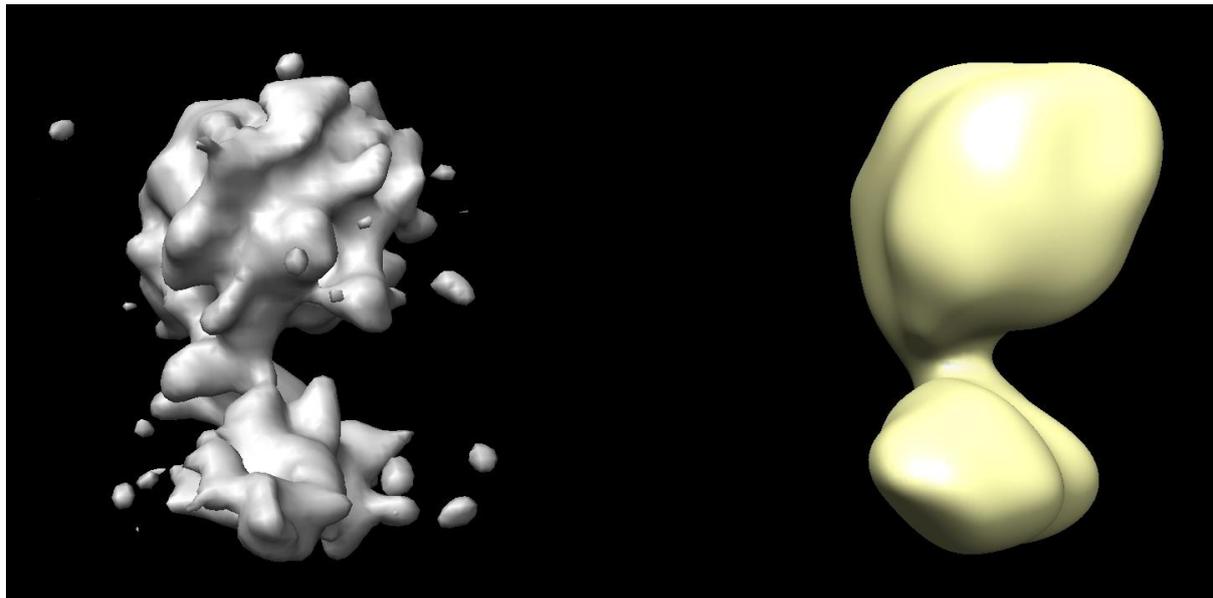
**Figure 88** : Classe 2D du complexe ERR $\alpha$ -ADN (Relion 2.0). On peut noter que les classes ne sont pas clairement définies.

### 3.3.4 Reconstruction 3D

J'ai produit une première structure initiale à partir du logiciel CryoSPARC, puis j'ai testé un affinement dans CryoSPARC et un affinement dans Relion1.4. J'ai obtenu un résultat d'affinement similaire entre les deux logiciels. Dans ce cas, la structure initiale contient des détails qui sont en accord avec la topologie d'un récepteur nucléaire et ressemble aux résultats obtenus préalablement. Cependant au cours de l'affinement, quelque soit les paramètres utilisés, il y a un lissage de la structure qui perd l'ensemble de ses détails.



**Figure 89** : A gauche, la structure initiale générée dans cryoSPARC, à droite la structure après 25 cycles d'affinement dans cryoSPARC. On note la topologie similaire entre le premier ADN utilisé (ADN BE33-tff1ERRE) et celui-ci. L'affinement dans cryoSPARC lisse la carte au lieu de faire ressortir les détails de moyenne résolution.



**Figure 90** : A gauche, la structure initiale générée dans cryoSPARC, à droite la structure après 25 cycles d'affinement dans cryoSPARC. (tournée de 90° dans l'axe Y)

### 3.3.5 Limites-problèmes

Pour cette étape j'ai eu des difficultés dans la phase d'affinement, il semble que les logiciels ne parviennent pas à attribuer correctement les angles d'Euler précis des particules. Dans ce cas précis cela peut provenir d'une hétérogénéité des particules sélectionnées ou alors de la présence de petits agrégats dû au glutaraldéhyde. De plus lors de la classification 2D, aucun logiciel de traitement à part IMAGIC actuellement testé n'est capable de faire une série de bonnes classes. Ces questions ont été approfondies pour le complexe suivant.

## 3.4 Complexe ERR $\alpha$ - BE29-embedded ERRE/ERE -PGC-1 $\alpha$

### 3.4.1 Préparation d'échantillons

Ce complexe est le complexe le plus récent sur lequel j'ai travaillé. Pour la préparation des échantillons, j'ai utilisé comme base la procédure de préparation, une première fois optimisée pour

le complexe  $ERR\alpha$ -BE33-embedded ERRE/ERE, le complexe étant basé sur ce dernier. Pour ce complexe, le fragment de PGC-1 $\alpha$  fait 516 acides aminés. J'ai décidé d'éliminer tout agent de réticulation, au vu des résultats obtenus avec le complexe précédemment acquis, d'autant plus que pour ce complexe nous n'avons pas observé de problèmes de stabilité. Pour une meilleure distribution, j'ai ajouté un détergent, le n-Dodecyl  $\beta$ -D-maltoside (DDM). Ce détergent améliore significativement la reproductibilité et la distribution des complexes sur les grilles. Ces nouvelles conditions améliorent différentes problématiques rencontrées précédemment. Cependant, ce procédé n'a pour le moment pas été appliqué aux complexes  $ERR\alpha$ -ADN seul ni au complexe comprenant les nucléosomes.

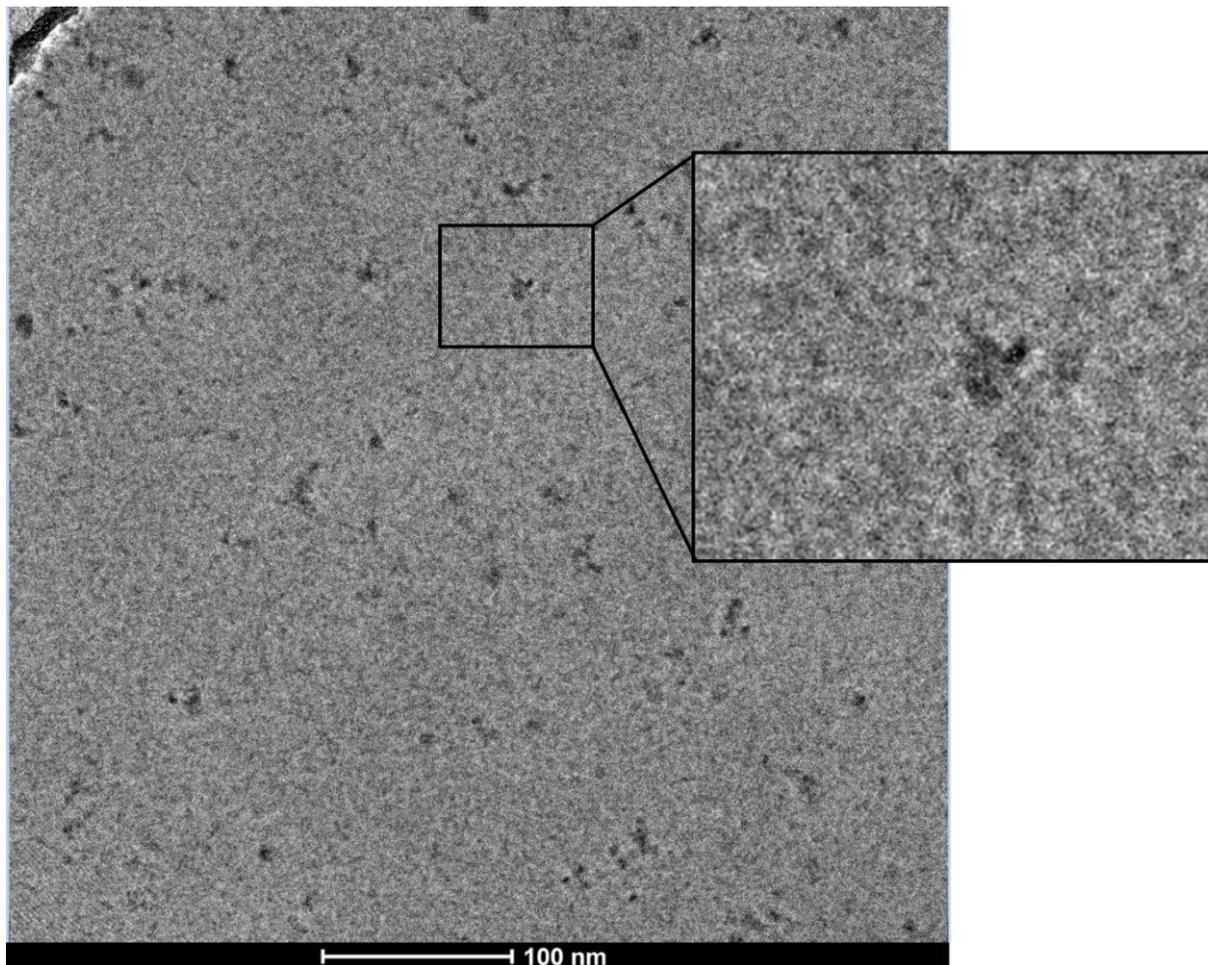


**Figure 91** : Séquence de l'ADN WC29-embedded-AA avec l'élément de réponse ERRE/ERE encadré.

### 3.4.2 Acquisitions

Cette acquisition a été faite dans les mêmes conditions que la précédente, à savoir, avec une caméra Gatan K2 Summit, la Volta phase plate avec un changement de positions toutes les 120 minutes, un voltage de 300 kV, une défocalisation constante de -0,5  $\mu$ m, un grossissement de 105 000 fois ce qui donne une taille par pixel de 1,09  $\text{\AA}$  (0.545  $\text{\AA}$ /px lors de l'acquisition en mode super-résolution). Les films ont été collectés différemment avec 40 images par film, un temps total de 8 sec d'acquisition par film et une dose de 5,6  $\text{é}/\text{\AA}^2/\text{s}$ , soit une dose totale de 45  $\text{é}/\text{\AA}^2$ .

Pour ce jeu de données de 2536 micrographes, enregistré grâce au logiciel d'acquisition automatique SerialEM, un total de 33 000 particules a été sélectionné entièrement manuellement avec le module de sélection manuel de Relion2.1. La sélection manuelle c'est avérée nécessaire pour avoir une meilleure qualité de sélections, en évitant ainsi de sélectionner des artefacts et contaminants sur les micrographes et également pour avoir un bon centrage des particules.



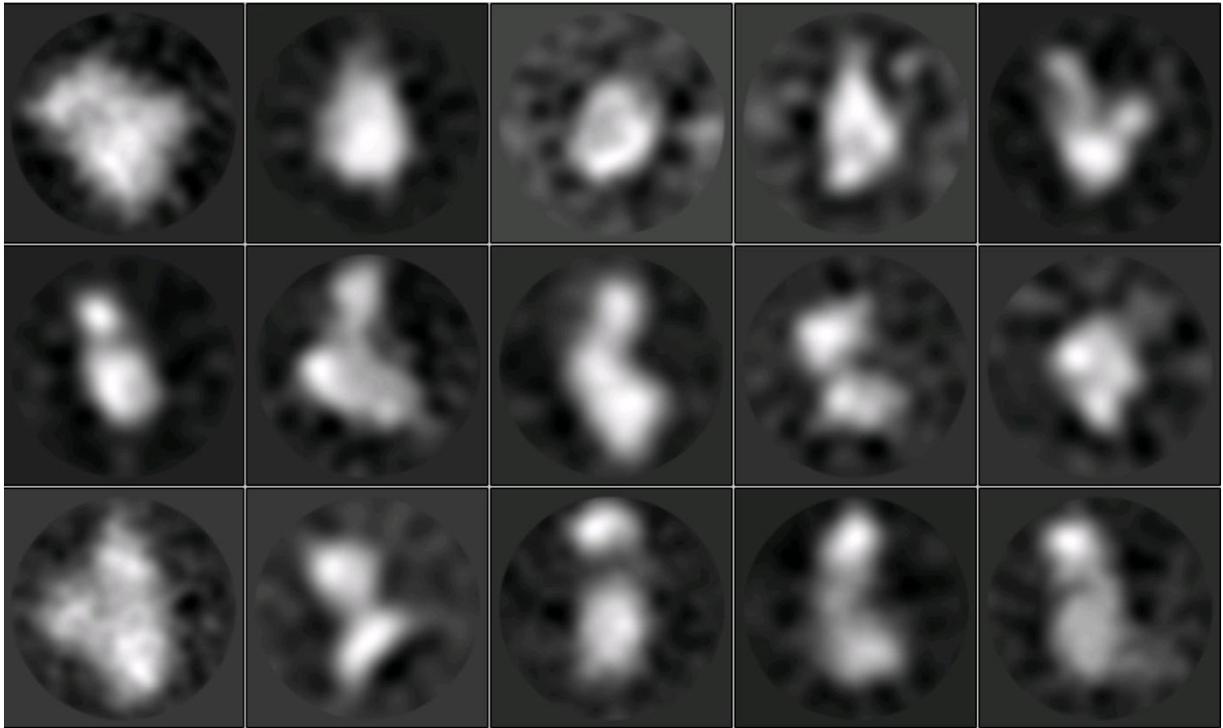
**Figure 92** : Micrographe du complexe ERR $\alpha$ -ADN-PGC-1 $\alpha$ . Image réalisée sur le microscope Titan Krios avec une caméra Gatan K2 summit et une Volta phase plate, défocalisation de -0,5  $\mu$ m, voltage de 300 kV, grandissement 105 000 fois.

### 3.4.3 Classification 2D

Pour ce jeu de données, j'ai procédé à de nombreux tests de classification 2D, à savoir avec les logiciels Relion 2.0 puis 2.1 qui ont un résultat similaire, avec cisTEM et avec IMAGIC.

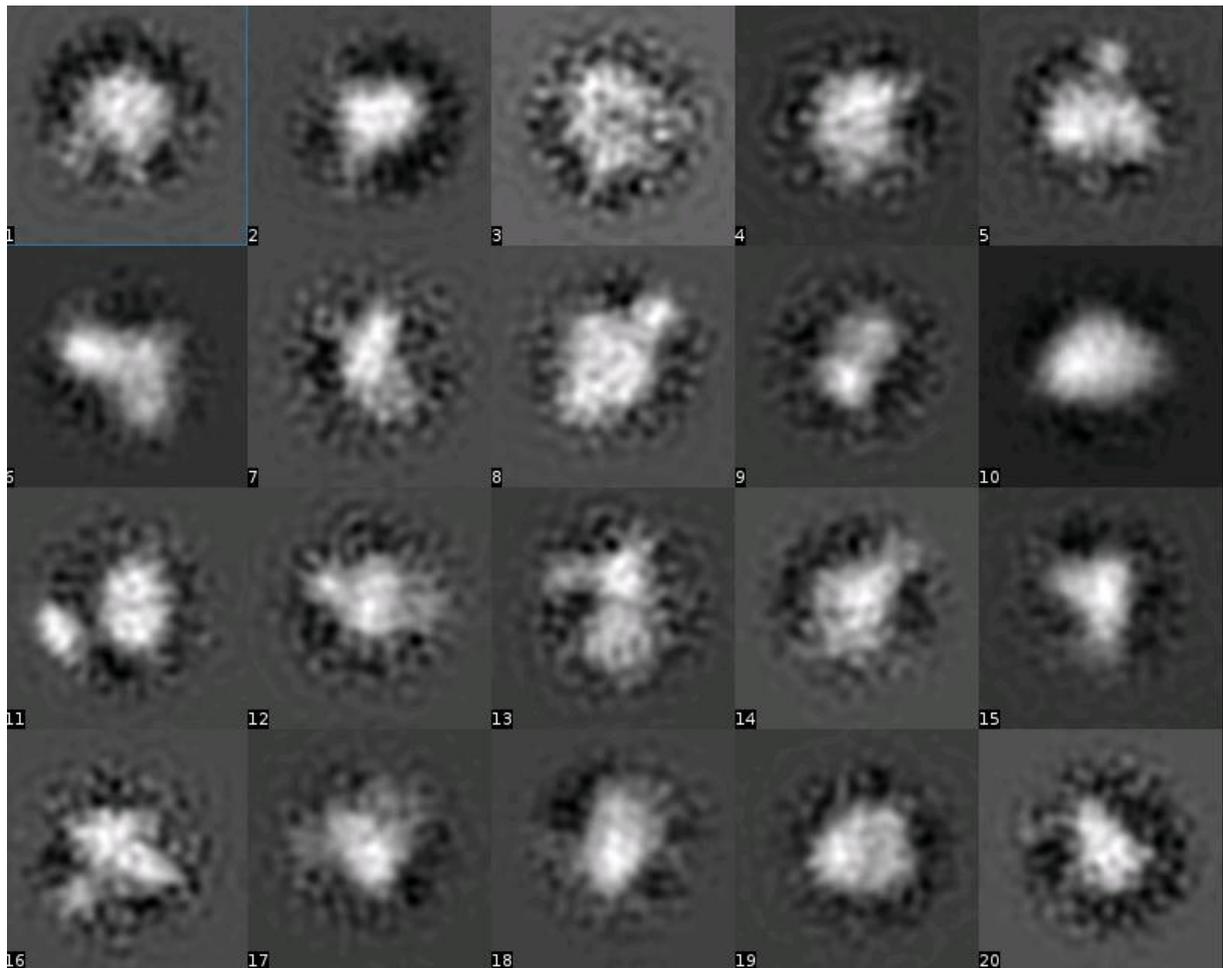
Avec Relion j'ai obtenu des classes de meilleur qualité que lors de la précédente acquisition. La question est de savoir si cela est dû aux données elles-mêmes ou à la nouvelle génération de logiciel. J'ai refait une classification 2D du jeu de données précédent avec la version 2.1 de Relion, et les résultats sont similaires à la classification faite avec Relion 1.4 et Relion 2.0. Ce n'est donc pas une question de version. La différence vient des données, de meilleures qualités et sélectionnées manuellement, ce qui assure un meilleur centrage que par des méthodes automatiques avec

références. En effet, les logiciels de sélection de manière générale ont du mal avec des petites particules comme dans ce cas.



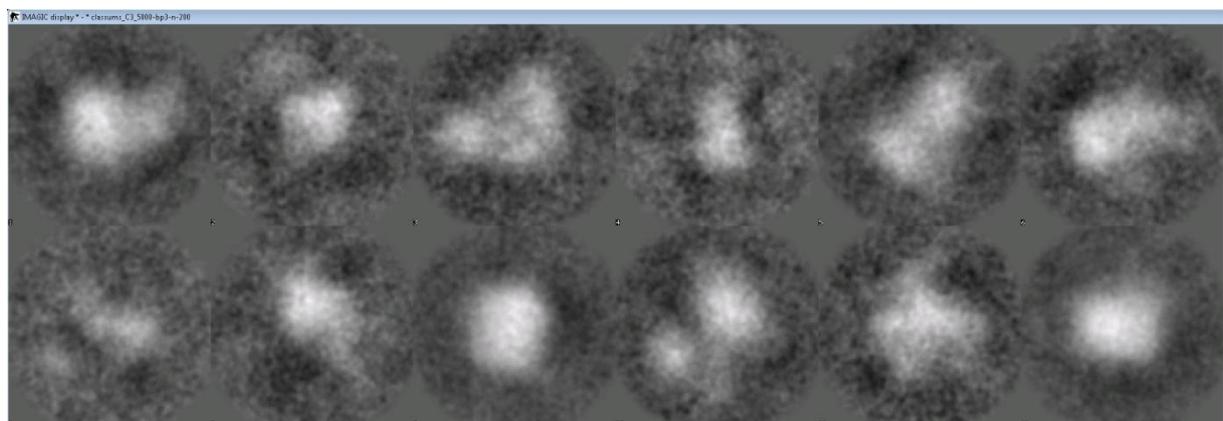
**Figure 93** : Classe 2D du complexe ERR $\alpha$ -ADN-PGC-1 $\alpha$  (Relion 2.1).

En utilisant le même jeu de données, cisTEM fournit un résultat similaire. La principale différence est l'apparition de bruit de haute fréquence dans cisTEM, et ce malgré des paramètres lui imposant de ne pas utiliser les hautes fréquences pour la classification 2D.

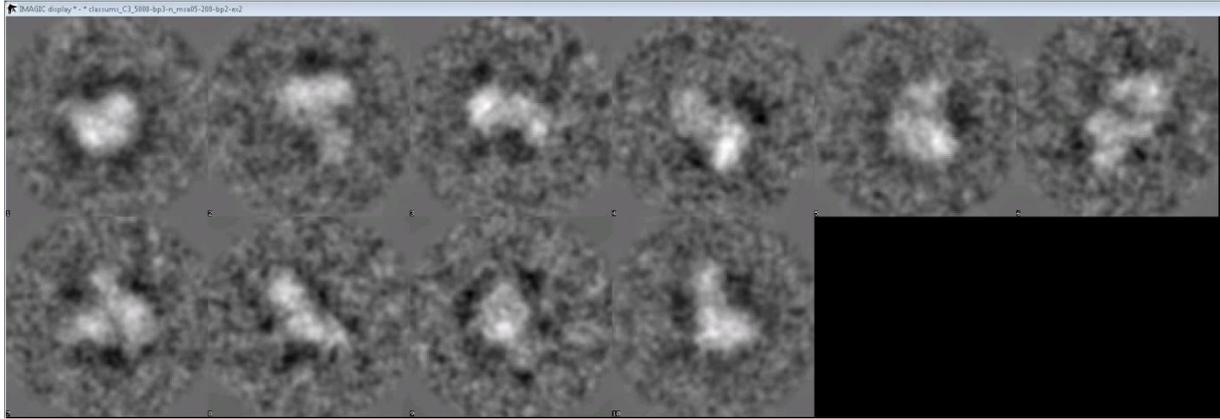


**Figure 94** : Classe 2D du complexe ERR $\alpha$ -ADN-PGC-1 $\alpha$  (cisTEM).

Avec IMAGIC qui utilise une autre approche, à savoir une méthode d'analyse statistique multivariée (MSA; Multivariate Statistical Analysis), on obtient de meilleurs résultats en utilisant des images filtrées directement dans IMAGIC.



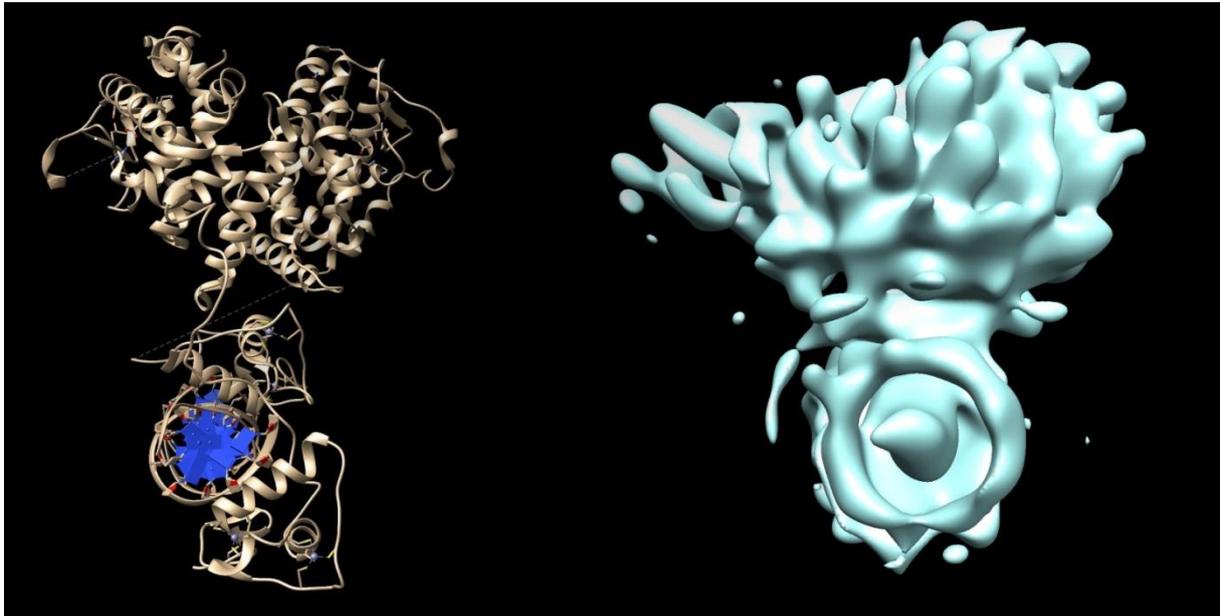
**Figure 95** : Classe 2D du complexe ERR $\alpha$ -ADN-PGC-1 $\alpha$  (IMAGIC sans filtre).



**Figure 96** : Classe 2D du complexe ERR $\alpha$ -ADN-PGC-1 $\alpha$  (IMAGIC avec filtre passe-bande).

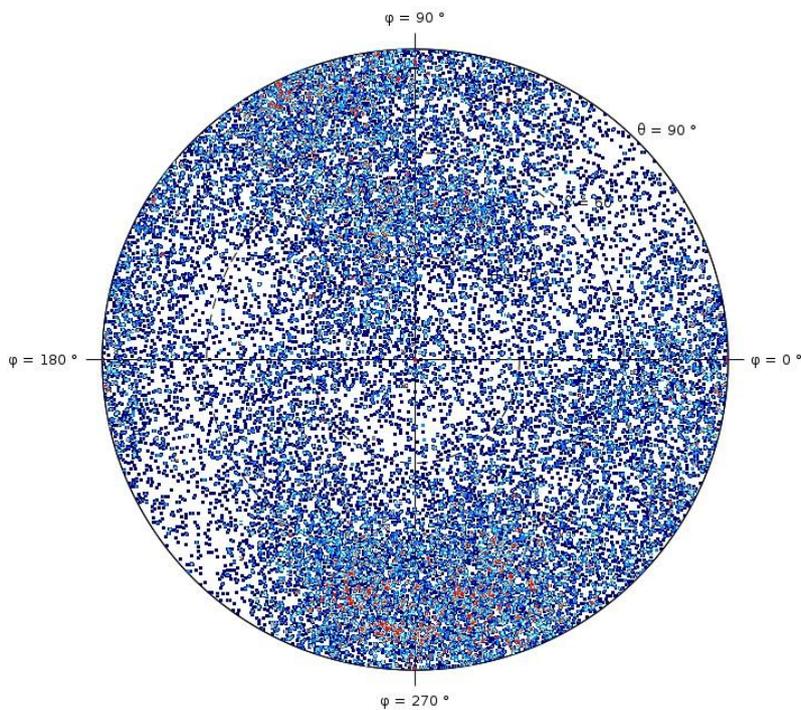
### 3.4.4 Reconstruction 3D

La structure initiale a été faite dans le logiciel cisTEM.



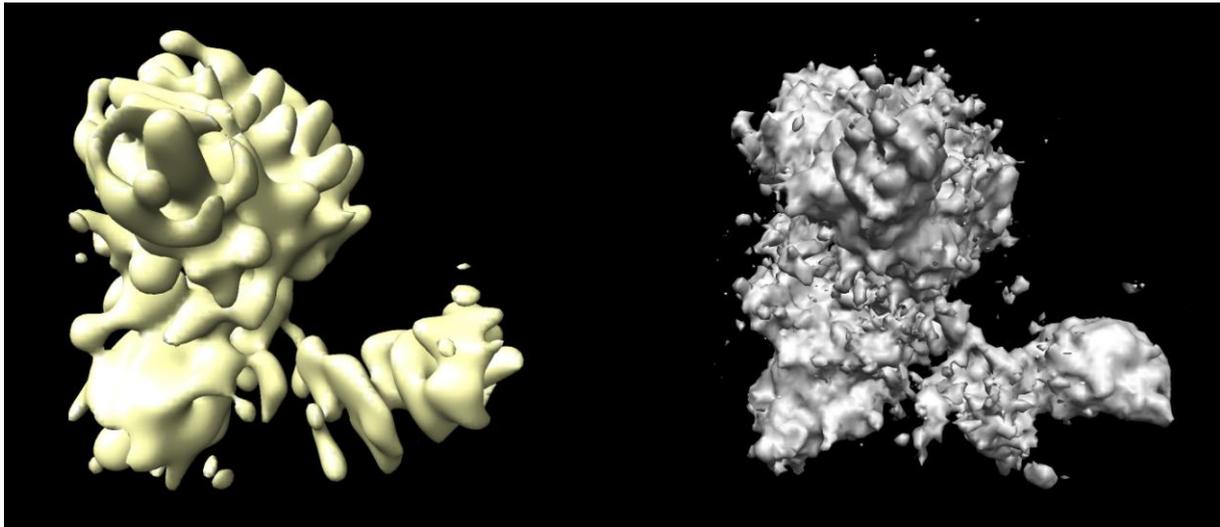
**Figure 97** : Comparaison entre une structure cristallographique et la carte de cryo-ME. A gauche, une structure cristallographique (PDB :4IQR) de HNF-4 servant ici de référence visuelle, mais n'est pas utilisée comme référence pour le traitement d'images. A droite la structure initiale du complexe qui laisse apparaître du côté gauche une excroissance qui peut correspondre à PGC-1 $\alpha$ .

Les premières étapes d'affinement ont été réalisées sur le logiciel cisTEM et sur Relion. On voit grâce à la distribution des angles d'Euler du complexe qu'il y a des orientations préférentielles, cependant il y a une bonne représentation des orientations minoritaires du complexe.

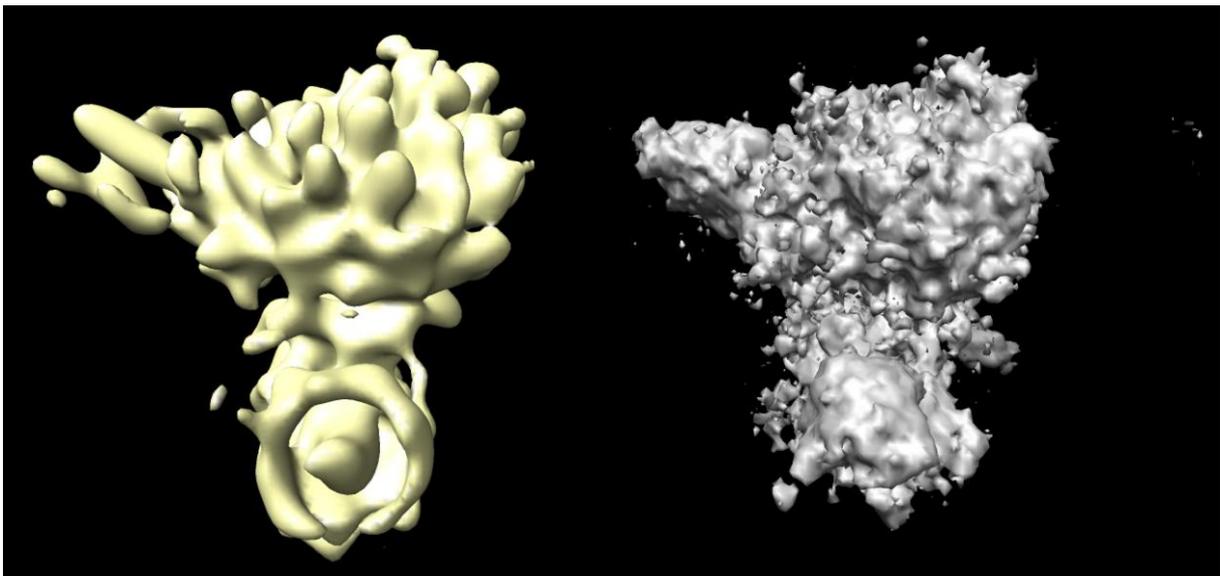


**Figure 98** : Distribution des angles d'Euler des particules du jeu de données (cisTEM).

Les résultats des premiers cycles d'affinement sur cisTEM font ressortir un bruit de haute fréquence plutôt important de la carte, et ce même en utilisant des paramètres limitant la résolution finale à atteindre.

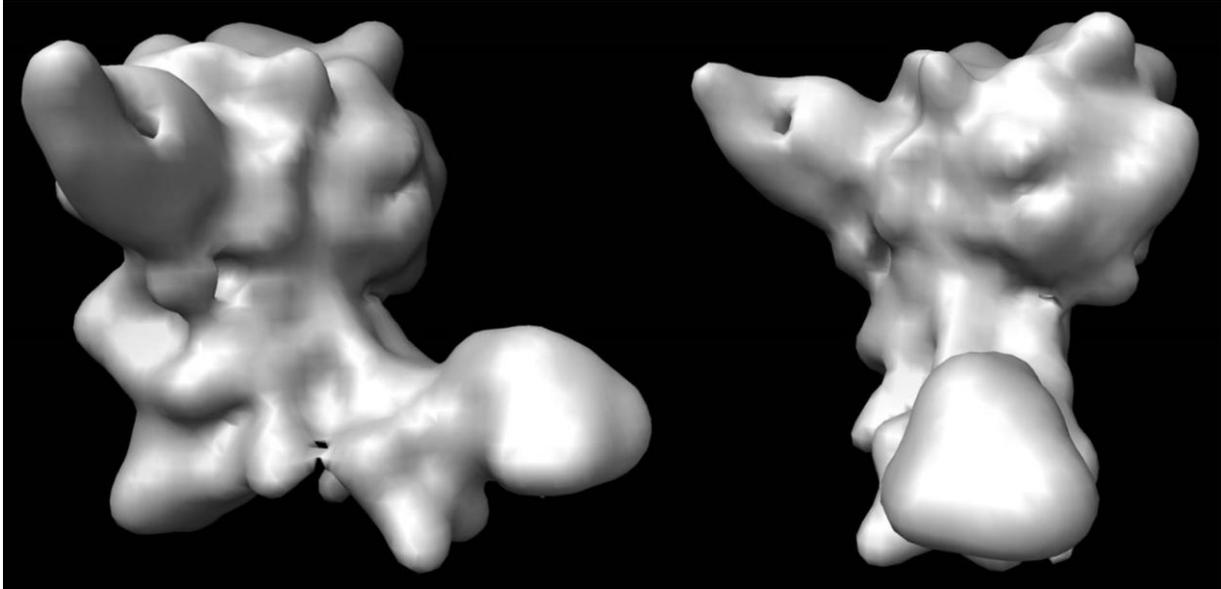


**Figure 99** : A gauche, la structure initiale, à droite le résultat après plusieurs cycles d'affinement avec cisTEM. La carte affinée n'est pas filtrée pour enlever le bruit de haute résolution. Il s'agit de la carte brute.



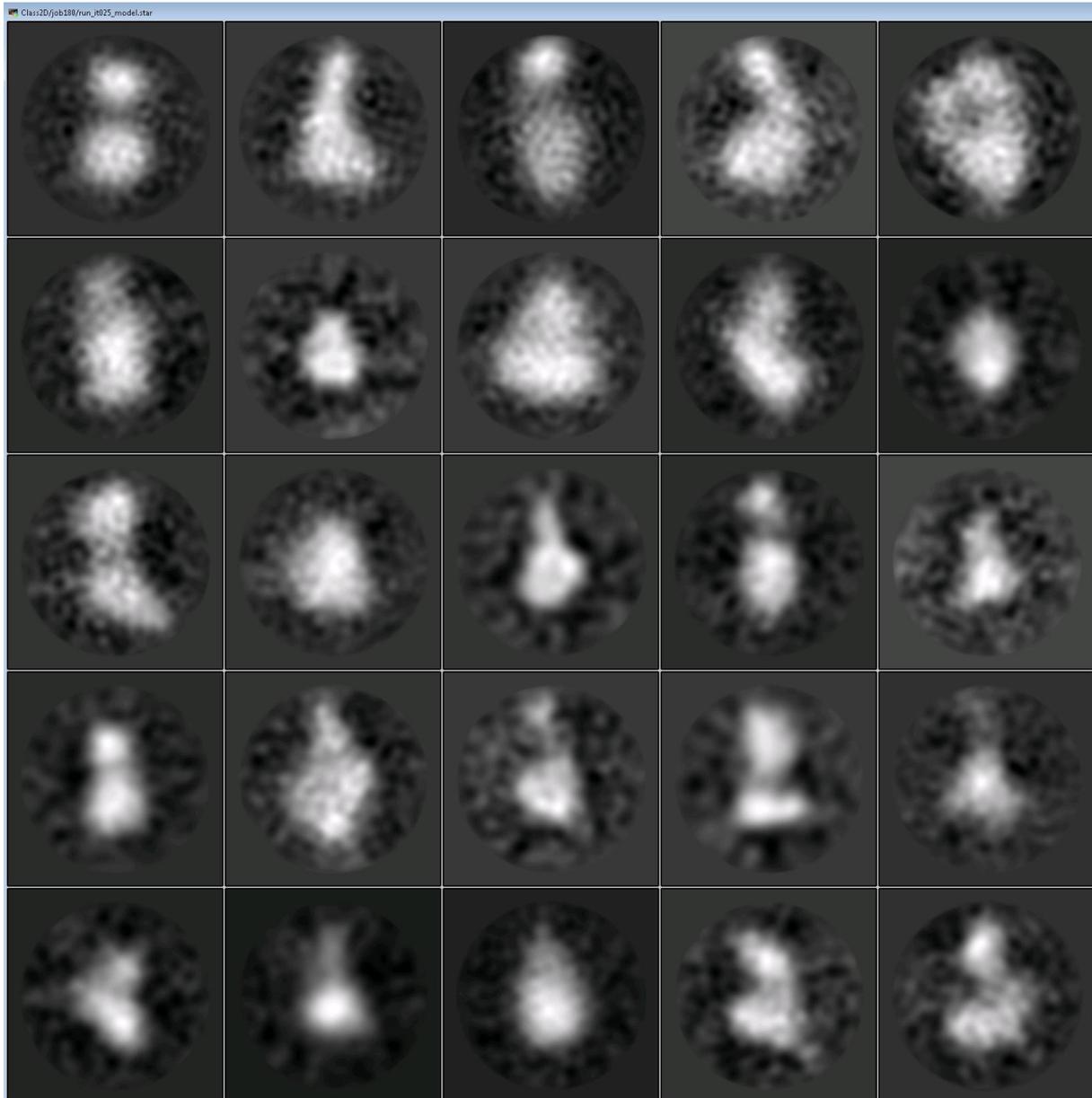
**Figure 100** : A gauche, la structure initiale, à droite le résultat après plusieurs cycles d'affinement avec CisTEM.

Par contre avec Relion2.1 j'obtiens de meilleurs résultats, il y a une meilleure gestion du bruit des hautes fréquences, car il s'arrête plus tôt dans les cycles d'affinement que CisTEM, la gamme de résolution utilisée est ainsi plus basse, mais le résultat est plus fiable. Il est néanmoins nécessaire d'utiliser la carte affinée et filtrée de cisTEM comme référence. A partir de la structure initiale de cisTEM, Relion ne parvient pas à attribuer correctement les angles.



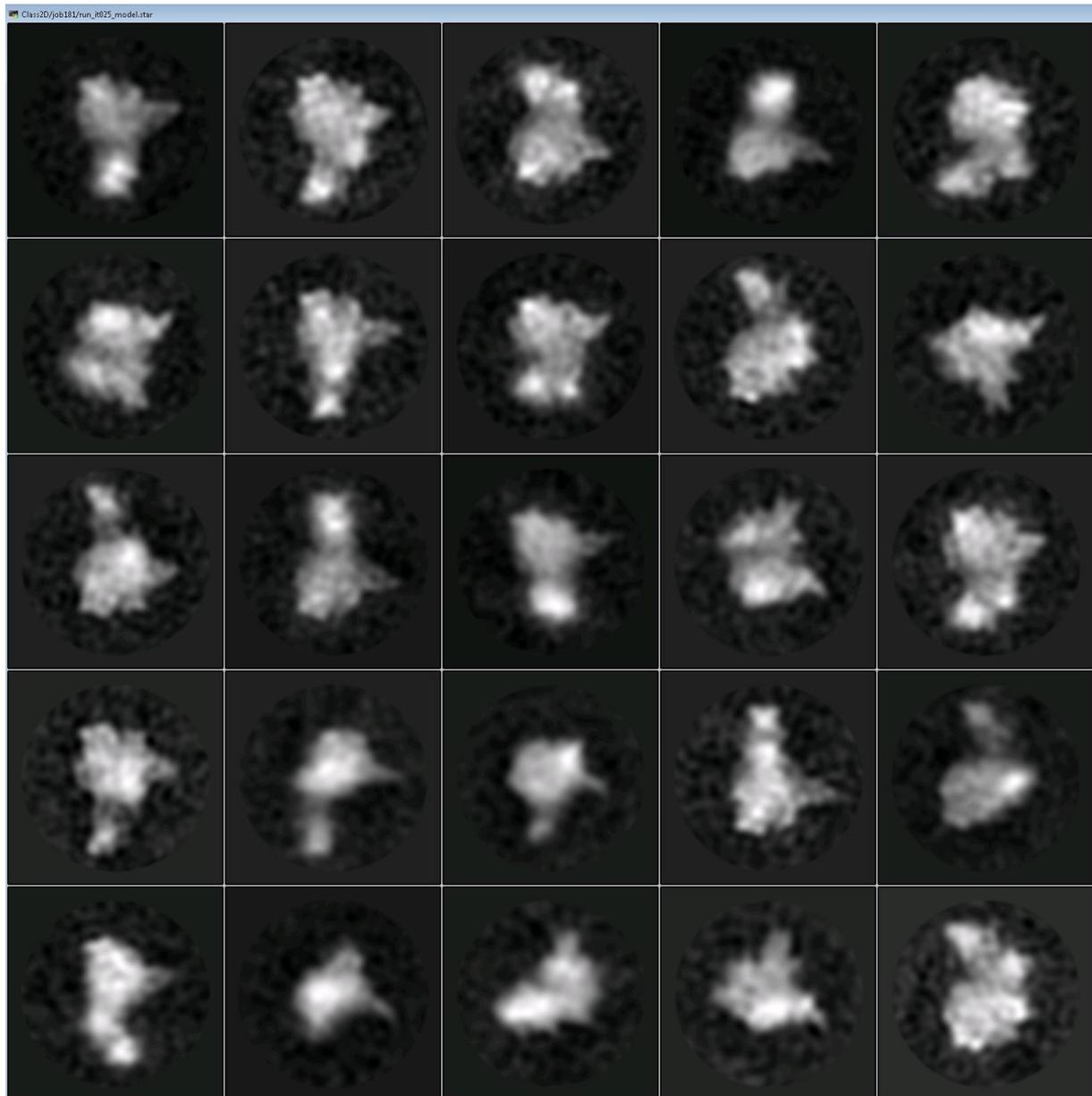
**Figure 101** : Structure après plusieurs cycles d'affinement avec Relion2.1.

Détails intéressants dans le cas de Relion, si on utilise les particules après alignement 3D lors des cycles d'affinement, avec des angles d'Euler corrects d'attribution. Si on les réinjecte dans une classification 2D en lui demandant d'affiner les angles, on obtient la classification 2D suivante.



**Figure 102** : Classification 2D de Relion2.1 à partir de particules avec des angles d'Euler correctement attribués par un cycle d'affinement 3D. Ici on demande à Relion d'optimiser les angles d'Euler.

On voit clairement ici qu'il y a une dérive et que les classes ne sont pas meilleures que la première classification 2D. Par contre en effectuant une classification 2D avec Relion2.1 et en mettant comme paramètre de n'effectuer aucune optimisation d'angle d'Euler, mais uniquement une classification, on obtient la classification 2D suivante.

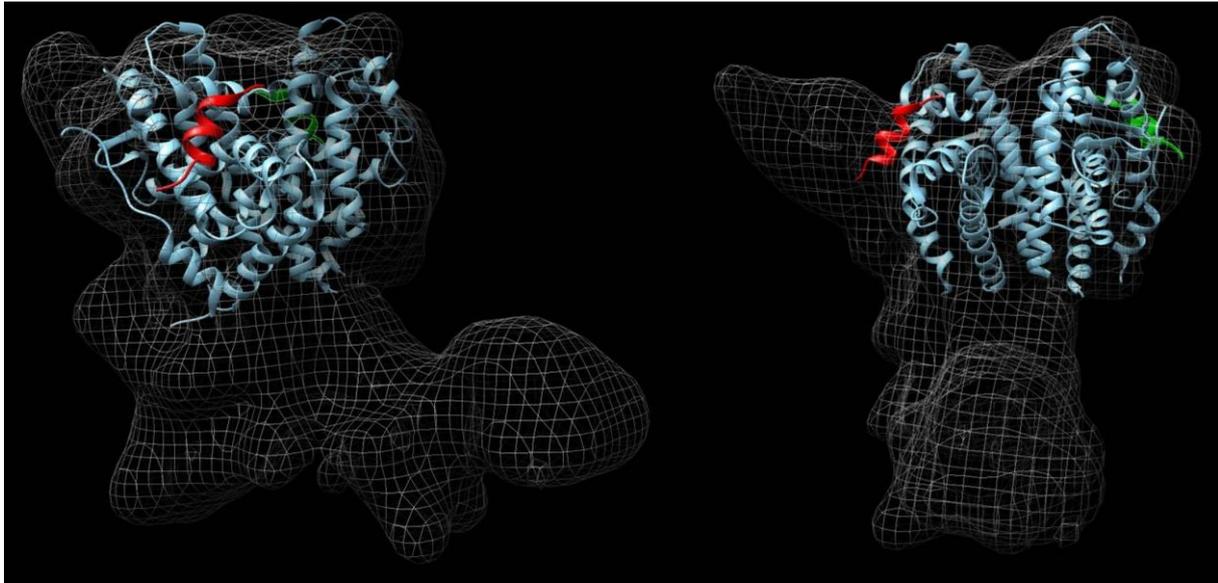


**Figure 103** : Classification 2D de Relion2.1 à partir de particules avec des angles d'Euler correctement attribués par un cycle d'affinement 3D. Ici on ne demande pas à Relion d'optimiser les angles d'Euler.

Ici les classes 2D sont de bien meilleures qualités, on voit clairement des éléments de structures secondaires apparaitre, alors qu'avec l'optimisation des angles d'Euler, ces détails sont perdus. On peut constater par conséquent que les différents logiciels actuellement disponibles comme cisTEM et Relion, n'arrivent pas à faire une bonne classification 2D de petites particules comme ce complexe d'étude, et ce même lorsqu'elles sont bien centrées et avec des angles d'Euler de départ correctes. Ce résultat permet d'expliquer en partie une série de difficultés rencontrées lors des traitements d'images précédents. Il est possible que ceci vienne d'une instabilité des paramètres en cours

d'affinement, en utilisant des méthodes de maximum de vraisemblance (Relion, cisTEM, cryoSPARC) alors qu'avec des classifications basées sur du MSA (IMAGIC) l'analyse des données semble meilleure.

Une première interprétation de la carte de Cryo-ME a été faite en positionnant et en recalant la structure cristallographique des LBD de ERR $\alpha$  avec un petit fragment de PGC-1 $\alpha$  (PDB 3D24 (Greschik et al., 2008)).



**Figure 104** : Positionnement d'une structure cristallographique des LBD de ERR $\alpha$  avec un petit fragment de PGC-1 $\alpha$  dans la carte de cryo-ME.

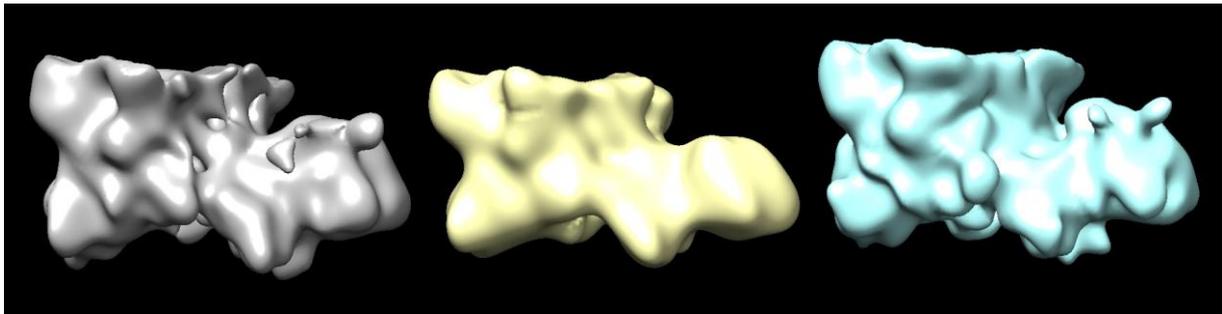
Les LBD sont représentés en bleu, et les fragments du coactivateur PGC-1 $\alpha$  sont représentés en rouge et en vert. On remarque que la densité additionnelle présente à gauche est probablement le coactivateur PGC-1 $\alpha$ , de plus l'absence de symétrie claire dans la carte de densité indique que PGC-1 $\alpha$  se fixe toujours du même côté sur le même LBD, et qu'il n'est pas de l'autre côté.

#### 3.4.5 Classification 3D

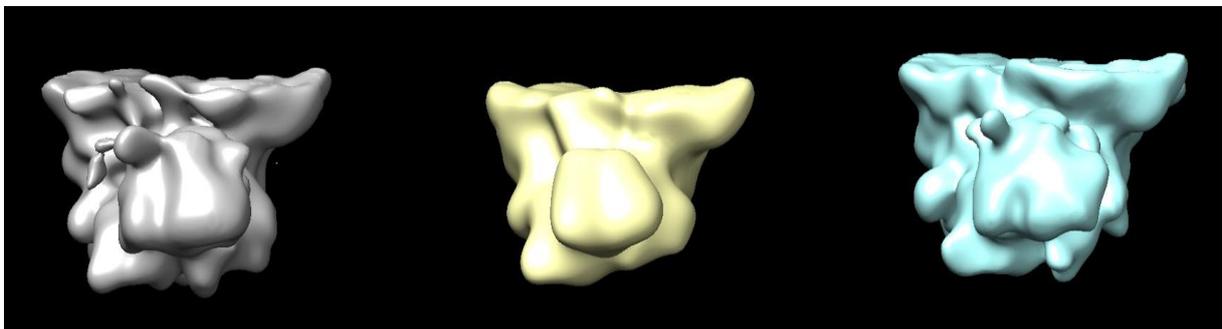
Un résultat obtenu récemment dans l'équipe, indique que les DBD de ERR $\alpha$  peuvent se fixer sur l'élément de réponse ERRE/ERE utilisé ici de deux manières différentes. En effet des résultats cristallographiques ont montré que 3 DBD seuls peuvent se fixer en simultanément sur cet élément de réponse. Dans ce cas il n'y a aucune contrainte stérique liée au reste du récepteur nucléaire,

néanmoins il est à envisager que les 2 DBD du complexe occupent 2 des 3 sites de liaisons différents observés, il pourrait donc y avoir une variabilité de fixation.

Pour vérifier si c'est aussi le cas en solution avec le complexe complet, j'ai fait une classification 3D de sous-région de la structure. Cependant je n'ai pour le moment pas constaté de différences indiquant ce phénomène en solution, tout au moins au niveau de résolution actuelle de la structure.



**Figure 105** : Classification 3D ciblé sur la partie ADN/DBD de la structure (focus refinement). On ne constate pas de grande variabilité entre les classes 3D. La différence de détails est principalement dû à la différence du nombre de particules de chaque classe. Ici la classe du milieu à trois fois moins de particules que les deux autres.



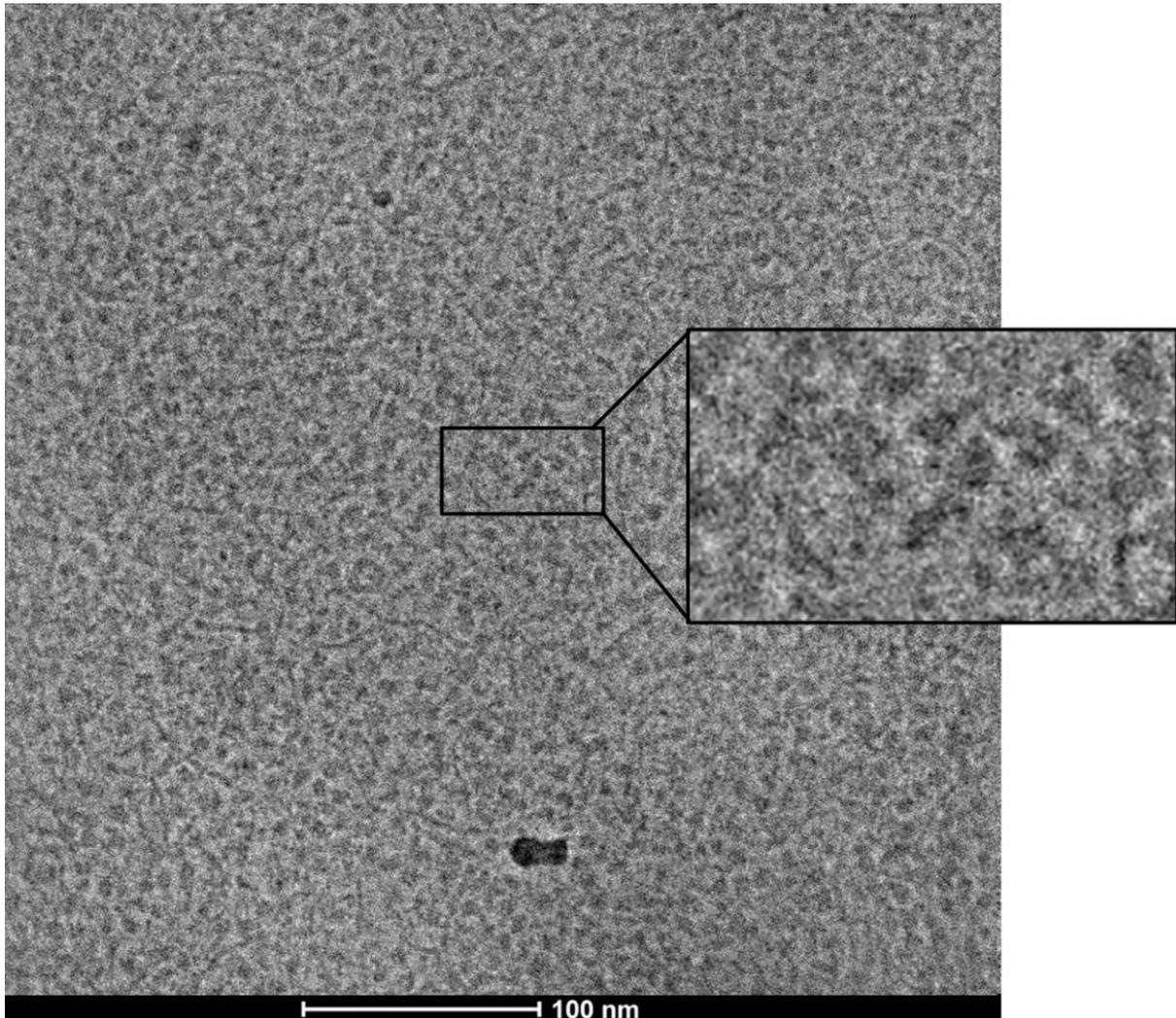
**Figure 106** : Classification 3D ciblé sur la partie ADN/DBD de la structure. Rotation de 90° sur l'axe Y.

On ne voit pas de différences notables entre les classes à cette résolution. Pour le moment les particularités de fixations observées en cristallographie avec les domaines DBD seul ne semblent pas être reproductibles avec le récepteur entier.

#### 3.4.6 Limites-problèmes

Pour ce jeu de données, je suis limité par le nombre de particules. En effet 32 000 particules avec un jeu de données présentant de l'hétérogénéité n'est pas suffisant. Pour remédier à ce problème, j'ai

acquis récemment un second jeu de données avec une concentration plus importante. L'acquisition a été faite sur le microscope Titan Krios avec à 300 kV équipé d'une caméra Gatan K2 Summit. Un grandissement de 130 000 fois (soit 0.89 Å/px contre 1.09 Å/px pour les acquisitions précédentes), un temps d'exposition de 8,4 secondes pour une dose totale de 50 é/Å<sup>2</sup> répartie entre 40 frames. Pour ce jeu de données de 3002 micrographes, enregistré grâce au logiciel d'acquisition automatique SerialEM. Le début des analyses montre qu'il sera probablement possible d'obtenir environ 400 000 particules après trie, sélectionnées avec le module automatique de cisTEM.



**Figure 107** : Micrographe du complexe ERR $\alpha$ -ADN-PGC-1 $\alpha$ . Image réalisée sur le microscope Titan Krios avec une caméra Gatan K2 summit et une Volta phase plate, défocalisation de -0,5  $\mu$ m, voltage de 300 kV, grandissement 130 000 fois.

Au vue de la forte densité de particules, il a été possible d'augmenter le grandissement. La forte densité représente aussi la principale difficulté de ce jeux de données. En effet lors des étapes de sélections automatiques, la sélection se fait fréquemment avec un mauvais centrage, la conséquence

pour la classification 2D est que le logiciel considère et centre 2 ou 3 particules comme s'il ne s'agissait que d'un seul objet.

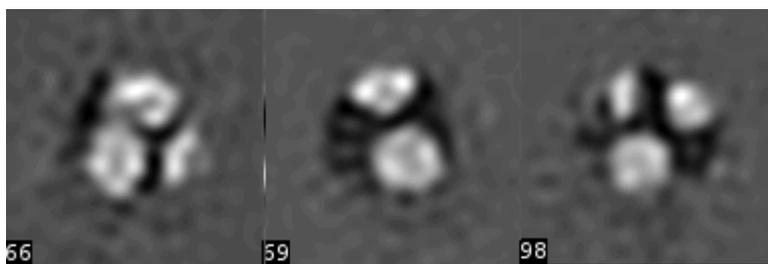


Figure 108 : Exemple de classes 2D de particules mal centré

Cette difficulté peut être corrigée en paramétrant le logiciel (ici cisTEM) pour qu'il recherche des particules plus petites que leur vraie taille, ainsi le centrage est de bien meilleure qualité et la majorité des classes où il y a plusieurs particules disparaissent.

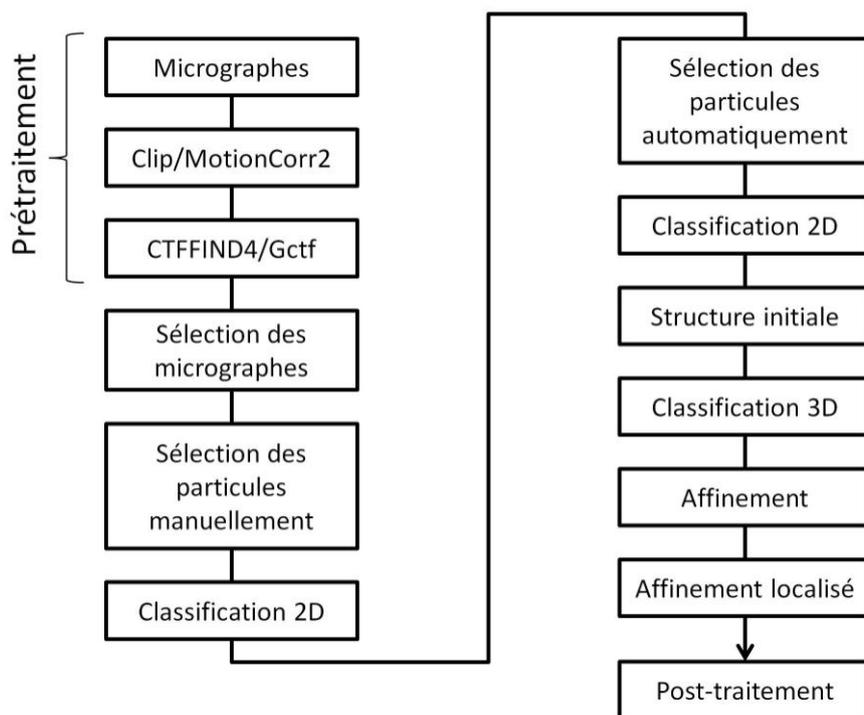


Figure 109 : Schéma des différentes étapes de traitement d'images.

Complexe	ERR $\alpha$ -ADN BE33- <i>tf</i> IERRE	ERR-Nucléosome (sans Volta phase plate)	ERR-Nucléosome (avec Volta phase plate)	ERR $\alpha$ -ADN BE33-embedded ERRE	Complexe ERR $\alpha$ - BE29-embedded ERRE -PGC-1 $\alpha$	Complexe ERR $\alpha$ - BE29-embedded ERRE -PGC-1 $\alpha$
<b>Microscope</b>	Polara	Titan	Titan	Titan	Titan	Titan
<b>Caméra</b>	Falcon I	K2	K2	K2	K2	K2
<b>Grandissement</b>	78 000	105 000	105 000	105 000	105 000	130 000
<b>Voltage</b>	100	300	300	300	300	300
<b>Frame</b>	1	28	28	28	40	40
<b>Temps d'exposition</b>	2 sec	7,1 sec	7,1 sec	8,4 sec	8 sec	8,4 sec
<b>Dose totale</b>	20é	50é	50é	60é	45é	50é
<b>Micrographes</b>	700	1042	1211	1022	2536	3002
<b>Particules</b>	17 000	120 000	232 000	38 000	33 00	400 00

**Tableau 3** : Récapitulatif des jeux de données principaux acquis au cours de la thèse.

#### 4. Conclusions et perspectives

L'optimisation de ces échantillons a demandé beaucoup de tests et d'efforts d'optimisations. Après de nombreux tests, nous avons isolé des conditions favorables pour le complexe ERR $\alpha$ -ADN-PGC-1 $\alpha$ , je peux essayer de les appliquer afin d'améliorer les complexes ERR $\alpha$ -ADN et les complexes avec les nucléosomes dans une optique de résolution structurale à haute résolution. Par exemple, l'utilisation du détergent tel que le DDM à faible concentration et la présence de nouveaux équipements pour la préparation des grilles, notamment pour le glow discharge permettent de faire des grilles plus reproductibles dans le cas de nos complexes d'intérêts. De plus, l'utilisation de DDM pourrait améliorer la stabilité des complexes avec le nucléosome en conditions cryogéniques. Pour ces complexes une autre solution a été envisagée mais non réalisée pour le moment : l'utilisation d'autres conditions de réticulation, avec par exemple un mixte de glutaraldéhyde avec du formaldéhyde, ou l'utilisation d'autres agents de réticulation tel que le BS3 (bis(sulfosuccinimidyl)suberate) (Shi et al., 2017). En attendant ces études m'ont permis d'avoir une première idée de la topologie générale du complexe. Pour les complexes avec le nucléosome, il n'y a pas de structure avec ERR, faute d'un jeu de données où le complexe est formé. Cependant les

micrographes permettent d'avoir une première idée de l'orientation du récepteur nucléaire par rapport au nucléosome, à savoir qu'il serait orthogonal par rapport au plan du nucléosome dans le cas du complexe avec un seul nucléosome. Pour les complexes de  $ERR\alpha$ -ADN-PGC-1 $\alpha$  on note l'asymétrie du complexe, il semble que PGC-1 $\alpha$  soit fixé qu'en un exemplaire sur les complexes. En théorie il y a deux sites de liaisons pour PGC-1 $\alpha$ , un sur chaque LBD. L'autre point est que PGC-1 $\alpha$  se fixe toujours du même côté dans le cas de ce complexe, pour cette ADN précis (BE29-*tff1* ERRE). De plus la position de PGC-1 $\alpha$  correspond bien à ce qui est observé dans les structures cristallographique qui contiennent un peptide de PGC-1 $\alpha$ .

De plus les nouvelles technologies qui sont apparues au cours de ma thèse permettent à présent de faire des acquisitions de micrographes à haute résolution de petits objets à 300 kV sur un microscope Titan Krios grâce aux caméras à détection directe d'électrons et à la phase plate. L'enjeu principal est d'améliorer la qualité du traitement d'images afin d'atteindre une résolution quasi atomique qu'il est encore difficile d'atteindre pour des petits complexes asymétriques.

## Chapitre 3

# Développements informatiques

---

### 1. IBiSS, un développement bioinformatique d'un outil interactif pour l'analyse structure-séquence de grands complexes macromoléculaires

#### 1.1 Introduction et contexte

Les complexes macromoléculaires, tels que les ribosomes, les complexes de régulation de la transcriptions etc., nécessitent des outils adaptés pour l'analyse des données, de types et d'origines différentes. Ces outils permettent de mener une étude dans une logique de biologie intégrative, alliant les aspects structure-séquence-fonction. Or encore aujourd'hui ces trois aspects sont le plus souvent traités de manière séparée mais méritent néanmoins d'être regroupés dans leur analyse. En effet, les séquences 2D d'acides aminés et leur repliement 3D sont étroitement liés avec des implications fonctionnelles directes.

Il y a de nombreux logiciels à utiliser en local ou sur internet qui sont disponibles pour différentes approches d'analyses biologiques. Pour une visualisation de données structurales, parmi les logiciels les plus complets et les plus puissants, nous trouvons PyMOL (DeLano, 2002) et Chimera (Pettersen et al., 2004) qui sont à utiliser en local. Ils permettent une manipulation avancée de la structure 3D des protéines, de superposer plusieurs structures entre elles ou encore de visualiser la séquence 2D correspondante à la structure grâce à un lien interactif entre les deux. Cependant, les deux outils ne sont pas disponibles en application web. D'autres applications qui elles fournissent un service web, proposent un nombre de fonctions légèrement inférieures, tel que Jmol (Cammer, 2007; Herráez, 2006) qui permet une visualisation de macromolécules avec toutes les principales fonctionnalités, mais où deux protéines issues de deux fichiers PDB différents ne peuvent pas être superposées.

### Chapitre 3 - IBiSS, un développement bioinformatique d'un outil interactif pour l'analyse structure-séquence de grands complexes macromoléculaires

D'autres logiciels sont spécifiques à l'analyse de séquences 2D de macromolécules constituées d'acides nucléiques et d'acides aminés.

Pour des alignements multiples et globaux de séquences, le principal programme utilisé est ClustalW (Thompson et al., 1994) qui peut être utilisé en local mais également sur internet. D'autres types de programmes d'alignement tel que BLAST (Altschul et al., 1990) existent. Ils ne permettent pas un alignement global multiple mais un alignement local, deux à deux, entre une séquence biologique spécifiée et les autres séquences de même nature, comprises dans une base de données. Ainsi ce programme peut directement être lié à une base de données et permet de retrouver les séquences qui ont le plus de similarité avec notre séquence d'intérêt. Jalview (Clamp et al., 2004; Waterhouse et al., 2009) est un programme pour visualiser et éditer les alignements multiples de séquences. Les versions les plus récentes permettent de le combiner directement avec Jmol afin de relier interactivement une séquence d'un alignement avec une structure 3D. Il suffit de lui spécifier la séquence et la structure. L'utilisation de ce programme est comme pour Jmol possible en version web JApplet.

Les bases de données biologiques sont devenues incontournables pour la recherche scientifique. Elles permettent de rassembler des quantités importantes d'informations venant de divers champs de recherches, notamment de la biologie structurale, la génomique, la protéomique ou encore la phylogénie. Grâce aux nouvelles technologies expérimentales appliquées en biologie, qui tendent vers des flux de données à haut débit, les besoins en bioinformatique et en traitement des données deviennent de plus en plus importants pour la recherche actuelle. La première difficulté consiste à organiser cette énorme masse d'informations pour la rendre disponible à l'ensemble de la communauté scientifique. Pour répondre à cette difficulté, la mise en place de différentes bases de données a permis de rendre l'information disponible via internet. Il y a trois grandes institutions qui sont en charge de la gestion de ces données, le NCBI (<http://www.ncbi.nlm.nih.gov/>; Ostell and Kans, 1998) aux USA, l'EBI (<http://www.ebi.ac.uk/>; Emmert et al., 1994) en Europe et le DDBJ (<http://www.ddbj.nig.ac.jp/>; Tatenno and Gojobori, 1997) au Japon. De ces bases découlent des sous-parties notamment pour les séquences protéiques avec la base de données UniProt (<http://www.uniprot.org/>; Apweiler et al., 2004) ou encore TrEMBL (Bairoch and Apweiler, 1996). D'autres bases de données sont spécialisées dans les structures 3D de protéines comme la base de données PDB (<http://www.rcsb.org/pdb/home/home.do>; Bernstein et al., 1977; Meyer, 1997; Peitsch et al., 1995) et EMDatabank (<http://www.emdatabank.org/>; Lawson et al., 2011). Devant la masse d'informations générée et tous les traitements informatiques automatiques réalisés, il est parfois difficile de vérifier la fiabilité d'une information qui peut être le fruit d'une erreur

d'annotation automatique. De plus, ces erreurs se reportent de base en base en raison de leurs implications pour les annotations automatiques des nouvelles données. C'est pour cette raison que certaines bases de données ne mettent à disposition que des données qui ont subi une vérification manuelle telle que Swiss-Prot (Bairoch and Boeckmann, 1991) qui fournit des séquences protéiques avec une plus grande fiabilité.

L'ensemble des données biologiques disponibles sur internet est facilement accessible. Cependant, leur exploitation ainsi que la manière de les corréler entre elles quelles que soient leurs origines n'est pas encore chose aisée. Des bases de données commencent à mettre à disposition des outils permettant une plus grande flexibilité pour l'utilisateur mais elles restent spécialisées sur un type de données bien précis comme par exemple StringDB (<http://string-db.org/> Snel et al., 2000) qui propose des données d'interactomique. Dans la plupart des cas le modèle courant est la mise à disposition des données biologiques dans une base de données publique. L'exploitation des données est ensuite placée à la charge de l'utilisateur. Celui-ci télécharge ce dont il a besoin puis exploite les données avec des logiciels installés sur son poste individuel.

La bioinformatique intégrative a pour objectif de permettre un accès unifié des données en permettant de les relier entre elles, pour former un tout cohérent, représentant l'ensemble des connaissances biologiques actuelles. Ainsi, avoir la possibilité de corréler des données structurales avec des alignements de séquences est un atout non négligeable, car l'alignement de séquences permet de mettre en évidence des motifs conservés dans une famille de protéines, pour un groupe d'espèces donné ou l'ensemble des espèces présentes. Par exemple, seulement deux acides aminés très éloignés dans la séquence peuvent se retrouver côte à côte une fois la protéine repliée. On peut ainsi retrouver une zone hydrophobe ou la formation d'une liaison hydrogène qui peuvent être importantes pour la formation d'un complexe ou encore pour un site catalytique. Or en se limitant à la séquence 2D il est difficile de prédire une telle interaction. Ainsi pour une étude complète de ces complexes, une approche combinant structure 3D, alignement de séquences 2D et évolution des complexes permet de mieux comprendre les relations structure-séquence-fonction de ceux-ci.

Le côté évolutif peut être ajouté par un arbre taxonomique qui montre alors les relations de parentés entre des groupes d'êtres vivants et permet de classer le vivant dans des groupes en fonction de nombreux critères. C'est un outil précieux pour pouvoir comparer des séquences d'un alignement de séquences et déterminer les spécificités de chaque groupe d'espèces.

### Chapitre 3 - IBiSS, un développement bioinformatique d'un outil interactif pour l'analyse structure-séquence de grands complexes macromoléculaires

Certains complexes macromoléculaires sont formés d'un ensemble important et variable de macromolécules de divers types comme des protéines, des ARN ou encore des ADN. Pour de tels complexes à sous-unités multiples et/ou à composition variable il devient nécessaire de développer des outils bioinformatiques performants et multi-approches.

Grace au développement important des réseaux informatiques, particulièrement dans les domaines scientifiques, le web permet la mise en place d'un service simple et rapide pour l'utilisateur sans aucune installation complexe, ce qui favorise son utilisation immédiate.

Ce travail a donné lieu à une publication qui constitue la partie suivante des résultats (Beinsteiner et al., 2015).

#### 1.2 Résultats - Article scientifique

Databases and ontologies

## IBiSS, a versatile and interactive tool for integrated sequence and 3D structure analysis of large macromolecular complexes

Brice Beinsteiner<sup>1,2,3,4</sup>, Jonathan Michalon<sup>1,2,3,4</sup> and Bruno P. Klaholz<sup>1,2,3,4,\*</sup>

<sup>1</sup>Centre for Integrative Biology (CBI), Department of Integrated Structural Biology, IGBMC (Institute of Genetics and of Molecular and Cellular Biology), Illkirch, France, <sup>2</sup>Centre National de la Recherche Scientifique (CNRS) UMR 7104, Illkirch, France, <sup>3</sup>Institut National de la Santé et de la Recherche Médicale (INSERM) U964, Illkirch, France and <sup>4</sup>Université de Strasbourg, Strasbourg, France

\* To whom correspondence should be addressed

Associate Editor: Anna Tramontano

Received on October 1, 2014; revised on May 22, 2015; accepted on May 30, 2015

### Abstract

**Motivation:** In the past few years, an increasing number of crystal and cryo electron microscopy (cryo-EM) structures of large macromolecular complexes, such as the ribosome or the RNA polymerase, have become available from various species. These multi-subunit complexes can be difficult to analyze at the level of amino acid sequence in combination with the 3D structural organization of the complex. Therefore, novel tools for simultaneous analysis of structure and sequence information of complex assemblies are required to better understand the basis of molecular mechanisms and their functional implications.

**Results:** Here, we present a web-based tool, Integrative Biology of Sequences and Structures (IBiSS), which is designed for interactively displaying 3D structures and selected sequences of subunits from large macromolecular complexes thus allowing simultaneous structure-sequence analysis such as conserved residues involved in catalysis or protein-protein interfaces. This tool comprises a Graphic User Interface and uses a rapid-access internal database, containing the relevant pre-aligned multiple sequences across all species available and 3D structural information. These annotations are automatically retrieved and updated from UniProt and crystallographic and cryo-EM data available in the Protein Data Bank (PDB) and Electron Microscopy Data Bank (EMDB).

**Availability and implementation:** The database contains all currently available structures of ribosomes, RNA polymerases, nucleosomes, proteasome, photosystem I and II complexes. IBiSS is available at <http://ibiss.igbmc.fr>

**Contact:** [klaholz@igbmc.fr](mailto:klaholz@igbmc.fr)

### Introduction

To obtain a better understanding of large macromolecular complexes that contain multiple and variable subunits, such as the ribosome, the RNA polymerase or the proteasome, appropriate tools are required to allow the integrated analysis of several subunit proteins down to the

level of individual residues. For example, eukaryotic ribosomes contain ~80 proteins of which some are conserved across species, some only share conserved domains, while others are specific to either prokaryotic or eukaryotic species. In this context, protein sequence analysis greatly benefits from the available 3D structures and vice versa.

From an Integrated Structural Biology perspective, new tools are needed that allow researchers to conduct broader studies in which various techniques, such as bioinformatics, biochemistry, crystallography and cryo electron microscopy (cryo-EM), may be combined (Klostermeier and Hammann, 2013 Ménétret *et al.* chapter). Within such an interdisciplinary approach, the integration of diverse sequence–structure–function aspects of a macromolecular complex becomes critical, and it is challenging with regards to the growing amount of biological information. Moreover, because of frequently changing nomenclatures or different naming conventions of related proteins in different organisms, it can become quite tedious to perform a comprehensive study. For well-documented families, such as ribosomal proteins, Uniprot allows gathering through protein families to facilitate the search. In many cases, the nomenclatures of these proteins lack unifying names; an example of addressing this problem is the recent renaming of ribosomal proteins into a species-unifying nomenclature (Ban *et al.*, 2014; Ben-Shem *et al.*, 2011; Yusupov *et al.*, 2001). Moreover, errors can occur and propagate due to automatic annotations which in turn are used for other annotations. The above-mentioned constraints make the use of these data complicated and require additional searching and tedious sorting and validation by the user.

While convenient stand-alone tools exist for sequence alignments (e.g. ClustalW (Thompson *et al.*, 1994), PipeAlign (Plewniak *et al.*, 2003)), taxonomic analysis (e.g. Phylogeny.fr (Dereeper *et al.*, 2008), iTOL (Letunic and Bork, 2007)), and for structure analysis (e.g. Pymol (Delano, 2002) and Chimera (Pettersen *et al.*, 2004)), there is no easy way to combine all this information (Procter *et al.*, 2010) which would provide a better understanding of the basis of molecular mechanisms and their functional implications. Software such as Friend (Abyzov *et al.*, 2005), STRAP (Gille and Robinson, 2006), Chimera's MultAlign viewer (Meng *et al.*, 2006), ConSurf (Ashkenazy *et al.*, 2010) and PyMod (Bramucci *et al.*, 2012) are not connected to a database and require a lot of manual intervention by the user. In general, existing software are better adapted for manual analysis of individual proteins with relatively small data sets, while not being designed for large-scale sequence analysis required for multi-subunit complexes. Thus, interactivity between sequence alignments and 3D structures directly connected to a database often remains unsatisfactory.

Larger databases, such as the NCBI (Ostell and Kans, 1998) and EBI (Emmert *et al.*, 1994), are focusing their efforts on data storage and allow data mining to some extent. Other website tools such as PipeAlign (Plewniak *et al.*, 2003) focus on the processing of sequences, providing efficient and easy to use sequence alignment tools and classification into protein sub-families. Specialized databases such as the Protein Data Bank (PDB) (Bernstein *et al.*, 1977; Meyer, 1997; Peitsch *et al.*, 1995), allow a preview of the structural data online via a Jmol applet version, but simultaneous sequence analysis is not possible. The NCBI server allows finding structures with a sequence similar to that of the query by using CBLAST (Wang *et al.*, 2007), and structures can be viewed with the sequence alignment using the software Cn3D (Hogue, 1997) ('see in 3D'). However, there are some limitations in that it requires downloading the software and retrieve all the structure and sequence data to perform the analysis locally, and the sequence alignment is limited to two sequences only, while protein subunits may contain hundreds; also, there is no particular amino acid color code there which could facilitate the identification of residues. Taken together, there is a critical need for tools, which could directly combine multiple sequence alignments and structures specifically for large and complex molecular assemblies which are particularly challenging to analyze.

Here, we propose an interactive tool that directly combines structure and sequence analysis using an integrated database of the relevant protein sequence alignments and available 3D structures. Using a web database, it provides access to data relevant for all the subunits of a given macromolecular complex. Such a multi-subunit complex is handled as a set of proteins, each of which contains the relevant data (sequences and atomic coordinates) and which are validated beforehand (e.g. protein sequence fragments are removed). The database contains the pre-calculated sequence alignments which are viewable in a Jalview applet window, and structures can be visualized in a Jmol applet window. It thus becomes easy to correlate structure and sequence alignments directly online for all available structures. Other tools, such as phylogenetic trees, tools for coevolutionary analysis, comparison between species with different composition of complexes are also available.

## Methods

The development tools used for the database are web technologies, which are the most appropriate with respect to speed and ease of use. The database is based on the NoSQL language (Lith and Mattsson, 2010) and uses MongoDB (Chodorow and Dirolf, 2010), which allows heterogeneous data to be handled, irrespective of their nature and quantity and which ensures a fast access for intensive search requests. The online services are written in Java and JavaScript.

Proteins included in the database were retrieved from Uniprot (Apweiler *et al.*, 2004), provided they have the status 'reviewed' as annotated in Swiss-Prot (Bairoch and Boeckmann, 1991; Bairoch and Apweiler, 1996). For each protein family, selection criteria such as protein name, gene name, sequence length were manually applied to check for errors such as annotation errors and partial sequences, before adding a given protein into a list for automatic updating. All PDB files associated with the sequences have been integrated in the database, together with extensive links to other databases, e.g. PubMed (Liu and Altman, 1998) and EMBL (Hamm and Cameron, 1986), and references have been linked to each protein. For each protein, IDs to all other databases referenced by UniProt are used to create links to each of the corresponding databases. For each protein subunit of a macromolecular complex, a global, multiple sequence alignment has been pre-calculated with the software ClustalW2 and Mafft (Katoh *et al.*, 2002; Thompson *et al.*, 1994), integrated into the database, and the user can choose between the two sequence alignments. To enable the simultaneous visualization of sequence alignments and 3D structures, an interactive link between sequence and structure was created using the software packages Jalview (Clamp *et al.*, 2004; Waterhouse *et al.*, 2009) and Jmol (Herráez, 2006; Cammer, 2007) (Java applet version). This allows processing data directly online in a fully interactive manner with an interface that comprises two windows for 3D structure and sequence, respectively. Interfacing between sequence alignments and 3D structure visualization is facilitated by the Jalview version that is part of Jmol. Amino acids are thus connected between the 2D sequence alignment and 3D structure, and a simple mouse movement over the sequence or structure easily displays the link by changing the appearance of the residue of interest. All Jalview and Jmol applet options remain available, such as editing of sequence alignments. The color code of the sequence alignment is transferred onto the structure using the amino acid color code implemented in PipeAlign according to distinct physicochemical properties (Plewniak *et al.*, 2003). An additional user interface implemented in JavaScript allows quick visualization of structures without displaying the sequences.

## Results

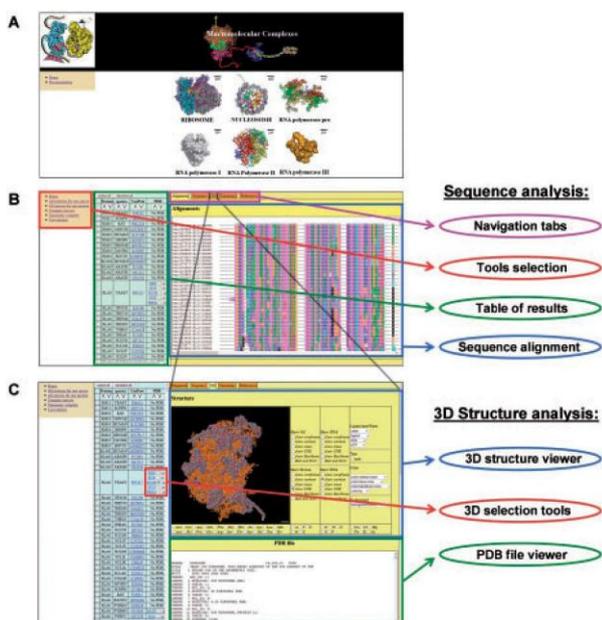
From the perspective of integrated structural biology, it is essential to be able to visualize and handle protein sequences and 3D structures in a correlated manner. The present tool, called Integrative Biology of Sequences and Structures (IBiSS), provides a user-friendly interface to interactively link and analyze sequences and 3D structures. Several types of information are needed to facilitate the connection between the analysis of 3D structures and sequence alignments from the level of the full complex down to individual protein subunits and residues. One of the key characteristics of IBiSS is to allow the identification of a conserved amino acid in a given sequence and have it simultaneously highlighted and localized in the 3D structure. This allows the user to analyze its role in molecular recognition in protein-protein or protein-nucleic acid interfaces and interactions (e.g. for translation and transcription complexes (Anger *et al.*, 2013; Ban *et al.*, 2000; Eiler *et al.*, 2013; Harms *et al.*, 2001; Klaholz *et al.*, 2004; Klaholz, 2011; Maletta *et al.*, 2014; Marzi *et al.*, 2007; Orlov *et al.*, 2012; Simonetti *et al.*, 2008; Simonetti *et al.*, 2013a, b; Wimberly *et al.*, 2000), in catalysis or in other functions.

To enlarge the potential applications of IBiSS, several important, widely studied macromolecular complexes involved in transcription or translation or other key cellular processes have been incorporated in the IBiSS database. These include the ribosome, the nucleosome, the RNA polymerase, the proteasome and photosystem complexes I and II (Fig. 1A) (Beck *et al.*, 2012; da Fonseca *et al.*, 2012; Golbeck, 1987; Guskov *et al.*, 2009; Low *et al.*, 2014; Loll *et al.*, 2005; Saenger *et al.*, 2002). For each complex, a set of protein subunits is defined. Additional, more specific functionalities allow the

comparison of protein composition in different species and taxonomic tree grouping based on species.

The development tools used for the database are web-based technologies, which have proven to be the most appropriate in terms of speed and ease of use. Indeed, the database NoSQL allows handling very heterogeneous data, as is often the case in biology. Web-based technologies allow the tool to be used on any operating system (OS), without special installation requirements or compatibility problems, facilitating software access and distribution. The choice of the MongoDB database that contains all the data provides high-speed interactivity. External databases, such as Uniprot and PDB are used only in the context of updating, but are not called directly through the interactive user interfaces. Because precalculated sequence alignments and PDB coordinates are stored in MongoDB, a quick and interactive access to the web site is achieved, which greatly enhances the data access speed through the web site. The update is done automatically on a monthly basis. To take into account modifications of sequence annotations and corrections, not only novel sequences are added but the entire sequences are updated with the associated structures. The aligned sequences are obtained from a global alignment to minimize alignment errors by using a large number of sequences. When required, it is possible to extract and display just two sequences out of many, while still benefiting from the better global alignment as compared to a simple pair-wise alignment (Fig. 1B). It is also possible to select a portion of the alignment for a quick taxonomic analysis using the checkbox, the results in the Sequence tab or directly on the co-evolution page.

In the following, we describe practical aspects of the IBiSS tool that are available to the user. On the main web page, the user selects the complex of interest, e.g. ribosome, nucleosome, etc. Next, a specific protein can be chosen. IBiSS then produces a page that contains all sequences related to this protein, and sequence alignment and 3D structure tools available for each sequence. The residue color code of the alignment is reported on the 3D structure and facilitates residue identification. A Jmol window allows visualizing the 3D structure and a Jalview window shows the sequence alignment. The two windows are connected interactively (Jmol is integrated in Jalview and serves for the 3D visualization). If the mouse is positioned over an amino acid of the 3D structure, the same amino acid in the corresponding sequence in the sequence alignment is automatically recentered and changes color in the Jalview window. Inversely, if the mouse is over an amino acid in the 2D sequence alignment window, the corresponding amino acid becomes highlighted in the structure (Fig. 2A). On the left panel, a table is included with the description of each sequence and an annotation of the protein describing name, species, UniProt identifier, tools for 3D visualization and correlation between sequences and structures that are displayed for each structure. Full information for each protein is also visible in a checkbox and by referring to a 'References tab' that allows a complete description of the protein with active links to other databases, in a manner as implemented in UniProt (Apweiler *et al.*, 2004). Displaying the results for the selected sequence is achieved without leaving the web page to avoid toggling between pages. A panel on the right of the interface provides access to sequence alignments, 3D structures, taxonomy and bibliographic references. Multiple sequence alignments can be retrieved from the IBiSS webpage. A portion of the alignment can be isolated to compare two or more sequences. The '3D' tab allows handling all the information relevant to the available 3D structures. A Jmol window is embedded in the web page with an interface to access the main features. The '3D' tab provides a quick and interactive view, for any of the available structures for this protein (structure of the protein alone, and structure of the protein complex). Displaying the PDB



**Fig. 1.** Integrated structure-function analysis with the IBiSS interface. (A) Overall home webpage from which a macromolecular complex can be selected. (B) Sequence analysis display. The purple box includes navigation tabs to access different functions to analyze the results of the table section. The red box allows changing analysis tools for a given complex. The blue box contains a scrollable sequence alignment tab. The green box contains a table of results for the tool 'All species for one protein'. (C) 3D structure analysis display. The blue box is part of the 3D structure content tab, which comprises a 3D structure viewer Jmol with a control panel for quick structure visualization. The red box contains the 3D selection tools. The green box is a PDB file viewer for displaying the PDB file content

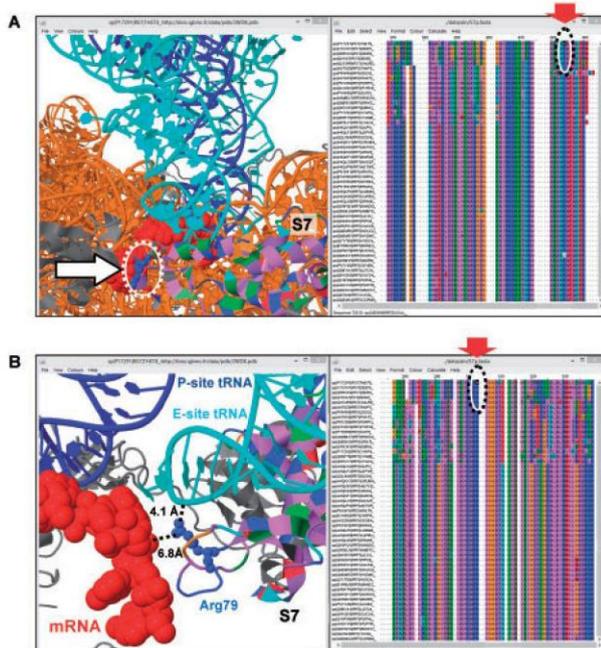
text file provides a view of the comments about the structures, such as the COMPND record that describes the macromolecular contents of an entry. The functionalities of Jmol are available by right clicking, in the same way as on the Jmol application of the PDB website, and some options are facilitated through the JavaScript interface (e.g. display options of macromolecules; Fig. 1C). All 3D functionalities, including those for structure-sequence correlation, will automatically load the Java applets. The 'taxonomy' tab allows displaying an interactive taxonomic tree summary for each species for which at least one sequence is present.

To analyze correlated amino acid changes, a coevolution analysis tool has been integrated in IBiSS which includes the entire protein database. It allows choosing a taxon for the protein of interest in a taxonomic tree, and visualizes the corresponding sequence alignments. Coevolutionary sequence variability between taxa can also be observed, e.g. if a taxon reveals a mutation in a binding site, the interacting protein partner of the same taxon may carry a complementary or compensatory mutation.

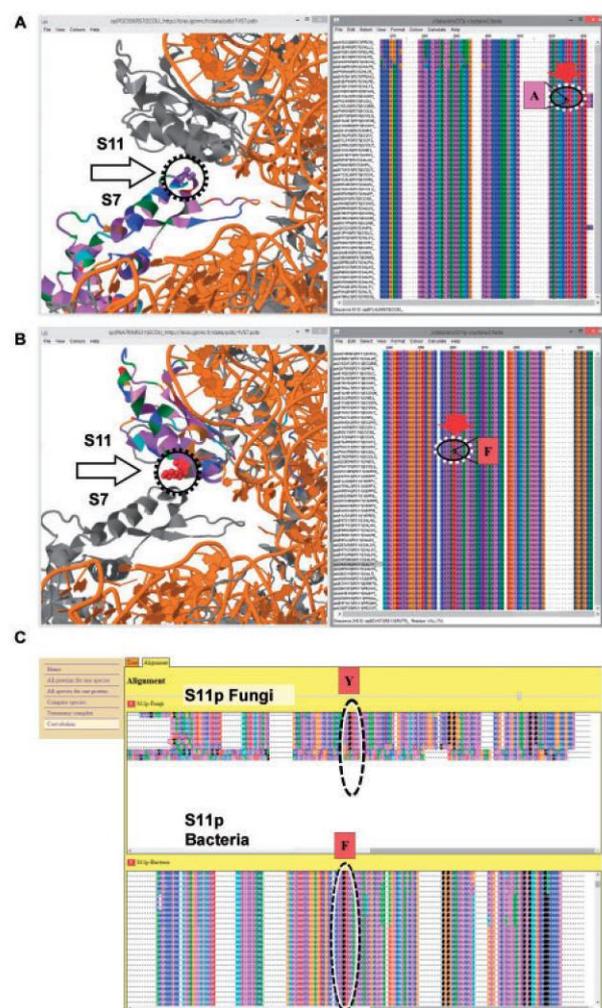
To illustrate the main features of the IBiSS tool, we present a typical example of a ribosomal protein, S7, which is part of the bacterial ribosome (~54 proteins, 3 rRNAs; Anger *et al.*, 2013; Ban *et al.*, 2000; Harms *et al.*, 2001; Wimberly *et al.*, 2000). S7 is located on the platform region of the 30S ribosomal subunit, close to the interface with the 50S subunit and the exit-site tRNA. It is involved in the binding of mRNA to the small ribosomal subunit. Using IBiSS, the number of sequences available for the S7p protein can be visualized at a glance. The sequence alignment reveals high sequence

conservation across species, with one variable domain between prokaryotes and eukaryotes.

This example illustrates the powerful potential of a convenient interactive analysis linking 3D structure and sequence alignment. By analyzing the amino acid conservation in the sequence alignment, a series of amino acids can be identified that are positively charged and highly conserved in all species of ribosomal protein S7. These residues are mainly located in the C-terminal region of protein S7. Because they do not interact with other proteins or with other ribosomal components, but instead are solvent-exposed, it can be concluded that they are important for the binding of mRNA considering that many of the conserved residues carry positively charged side-chains which could interact with the negatively charged phosphate moieties of the mRNA. Thus, it is very easy to localize surface-exposed residues that



**Fig. 2.** A concrete example of linking sequence alignments and structure information of a ribosomal protein, as part of the multi-proteinsubunit 70S ribosome. (A) Global view of the structure with protein S7 (small ribosomal subunit, 30S) interactively linked with the sequence alignment. The mRNA and tRNAs are indicated in red and blue, respectively. (B) In the example given, ribosomal protein S7 is localized on the structure window with the backbone chain residues labeled according to the residue color code in the sequence window. The solvent-exposed and conserved residue Arg 79 is highlighted in red as a good candidate for interactions with mRNA bound in the 30S platform region. mRNA and tRNAs are colored in red, blue and cyan respectively. For simplicity, ribosomal rRNAs and the 50S ribosomal subunit are not shown



**Fig. 3.** Example of coevolution study between ribosomal S7 and S11 proteins using IBiSS. (A) Global view of the ribosome structure with the S7 protein interactively linked with the corresponding sequence alignment. (B) Global view of the ribosome structure with the S11 protein interactively linked with the corresponding sequence alignment. (C) The selection of sequences of the multi-sequence alignment based on taxa reveals a coevolution at the interface between proteins S7 and S11. In protein S11, a phenylalanine (B) residue conserved in bacteria (label 'F') is replaced by a tyrosine (label 'Y') conserved in fungi; within a 3 Å distance, in protein S7, an alanine (A) residue is replaced by a serine residue, suggesting an additional hydrogen bond between protein S7 and S11 in fungi. With the IBiSS tool, this analysis is quick and easy thanks to its integrated database

could be involved in mRNA recognition (Fig. 2B). In a previous study, all of this comprised a tedious manual analysis, including the retrieval of structures, sequences and residue identification (Marzi *et al.*, 2007).

An additional taxonomy analysis is also possible, which in this example leads to a large number of proteins because of the broad representation of protein S7 in prokaryotes, eukaryotes or archaea. The taxonomy can be used to study coevolutionary aspects from the 'co-evolution' page. From the taxonomic tree, one can select the protein of interest for many taxa of interest and display the corresponding alignments. Because several amino acids of interest have been already identified, each taxon can be displayed to more easily identify at what point of evolution a given mutation occurred. It is also possible to use multiple Jalview and Jmol windows simultaneously (e.g. with multiple screens). The 3D structure shows that protein-protein interactions occur between protein S7 and proteins S9 and S11 (Fig. 3), for which the identification of amino acids is straightforward in IBiSS. Moreover, co-evolution analysis in IBiSS conveniently identifies a conserved residue pair at the interface between proteins S7 and S11 (Fig. 3). Inversely, strictly conserved amino acids in proteins S9 and S11 which are involved in protein-protein interactions with S7 can be identified easily within the ribosome complex.

Taken together, IBiSS provides a convenient, interactive and robust tool that facilitates combined structure and sequence analysis of complicated, multi-subunit macromolecular complexes. As compared to the combined usage of independent tools such as STRAP, Friend, Chimera and PyMOL, IBiSS represents a significant added value in that it provides an immediate, interactive and structured way to access the data through an integrated database which interactively links structures, pre-computed multi-sequence alignments, co-evolutionary and phylogeny tools. The data are standardized and grouped into protein families thus simplifying work when the same protein has several different names in the databanks. All features are integrated and can be used directly without the need of plugins. This is very useful for the identification of conserved amino acids and their 3D localization, eventually leading to design of new experiments for testing the functional implications of a particular residue. Using a web-based database, IBiSS gives immediate access to the relevant protein sequence alignments and available 3D structures for all the subunits of a given macromolecular complex. An additional typical feature of IBiSS is that phylogenetic analysis and coevolutionary tools are linked to multi-sequence alignments.

Taken together, the present tool provides an easy and unified access to structure and sequence information of large complexes and allows data mining in an interactive and correlated manner, thus facilitating the identification of conserved amino acids in a sequence and their localization in the 3D structure. In the example provided above, residues exposed on the surface of the small ribosomal subunit platform can be identified much more easily than in the previously published manual analysis (Marzi *et al.*, 2007). In general, IBiSS contributes to the development of integrative tools which are increasingly needed for integrated biology approaches. The preprocessed data which are accessible online allow performing bioinformatics analysis without requiring much data manipulation nor advanced knowledge in bioinformatics data processing. The database within IBiSS currently contains a series of typical large complexes, the ribosome (including the latest human ribosome structure; (Khatter *et al.*, 2015), translation initiation factors, the RNA polymerases, nucleosome, proteasome and photosystem I and II complexes. Currently, the database comprises 62 000 sequence entries and 2400 structure entries which are pre-processed to enhance interactivity. In the future, it could become interesting to extend IBiSS to

include user-defined complexes to create their own data base within IBiSS.

## Acknowledgements

We thank Julie Thompson, Jean-François Ménétret, Kareem Abdul Mohideen and Isabelle Billas for comments and the other group members for discussions, Remy Fritz for making the software available online, and the referees for nice suggestions.

## Funding

This work was supported by the European Research Council (ERC Starting Grant 243296), the BioScape Project, the Centre National pour la Recherche Scientifique (CNRS), the French Infrastructure for Integrated Structural Biology (FRISBI) ANR-10-INSB-05-01, and Instruct as part of the European Strategy Forum on Research Infrastructures (ESFRI) and the IGBMC facilities. The electron microscope facility is supported by the Alsace Region, the Fondation pour la Recherche Médicale (FRM), INSERM, CNRS and the Association pour la Recherche sur le Cancer (ARC). Université de Strasbourg (IDEX, Investissement d'Avenir).

*Conflict of Interest:* none declared.

## References

- Abyzov, A. *et al.* (2005) Friend, an integrated analytical front-end application for. *Bioinformatics*, **21**, 3677–3678.
- Anger, A.M. *et al.* (2013) Structures of the human and Drosophila 80S ribosome. *Nature*, **497**, 80–85.
- Apweiler, R. *et al.* (2004) UniProt: the Universal Protein knowledgebase. *Nucleic Acids Res.*, **32**, D115–D119.
- Ashkenazy, H. *et al.* (2010) ConSurf 2010: calculating evolutionary conservation in sequence and structure of proteins and nucleic acids. *Nucleic Acids Res.*, **38**, W529–W533.
- Bairoch, A. and Apweiler, R. (1996) The SWISS-PROT protein sequence data bank and its new supplement TREMBL. *Nucleic Acids Res.*, **24**, 21–25.
- Bairoch, A. and Boeckmann, B. (1991) The SWISS-PROT protein sequence data bank. *Nucleic Acids Res.*, **19**, 2247–2249.
- Ban, N. *et al.* (2014) A new system for naming ribosomal proteins. *Curr. Opin. Struct. Biol.*, **24**, 165–169.
- Ban, N. *et al.* (2000) The Complete Atomic Structure of the Large Ribosomal Subunit at 2.4 Å Resolution. *Science*, **289**, 905–920.
- Beck, F. *et al.* (2012) Near-atomic resolution structural model of the yeast 26S proteasome. *Proc. Natl. Acad. Sci. USA*, **109**, 14870–14875.
- Bernstein, F.C. *et al.* (1977) The protein data bank: A computer-based archival file for macromolecular structures. *J. Mol. Biol.*, **112**, 535–542.
- Bramucci, E. *et al.* (2012) PyMod: sequence similarity searches, multiple sequence-structure alignments, and homology modeling within PyMOL. *BMC Bioinformatics*, **13**, S2.
- Cammer, S. (2007) SchiSM2: creating interactive web page annotations of molecular structure models using Jmol. *Bioinformatics*, **23**, 383–384.
- Chodorow, K. and Dirolf, M. (2010) MongoDB: The Definitive Guide O'Reilly Media, Inc.
- Clamp, M. *et al.* (2004) The Jalview Java alignment editor. *Bioinformatics*, **20**, 426–427.
- Delano, W. (2002) The PyMOL Molecular Graphics System.
- Dereeper, A. *et al.* (2008) Phylogeny.fr: robust phylogenetic analysis for the non-specialist. *Nucleic Acids Res.*, **36**, W465–469.
- Eiler, D. *et al.* (2013) Initiation factor 2 crystal structure reveals a different domain organization from eukaryotic initiation factor 5B and mechanism among translational GTPases. *Proc. Natl. Acad. Sci. USA*, **110**, 15662–15667.
- Emmert, D.B. *et al.* (1994) The European Bioinformatics Institute (EBI) databases. *Nucleic Acids Res.*, **22**, 3445–3449.

- da Fonseca, P.C.A. et al. (2012) Molecular model of the human 26S proteasome. *Mol. Cell*, **46**, 54–66.
- Gille, C. and Robinson, P.N. (2006) HotSwap for bioinformatics: A STRAP tutorial. *BMC Bioinformatics*, **7**, 64.
- Golbeck, J.H. (1987) Structure, function and organization of the photosystem I reaction center complex. *Biochim. Biophys. Acta BBA - Rev. Bioenerg.*, **895**, 167–204.
- Guskov, A. et al. (2009) Cyanobacterial photosystem II at 2.9-Å resolution and the role of quinones, lipids, channels and chloride. *Nat. Struct. Mol. Biol.*, **16**, 334–342.
- Hamm, G.H. and Cameron, G.N. (1986) The EMBL data library. *Nucleic Acids Res.*, **14**, 5–9.
- Harms, J. et al. (2001) High Resolution Structure of the Large Ribosomal Subunit from a Mesophilic Eubacterium. *Cell*, **107**, 679–688.
- Herráez, A. (2006) Biomolecules in the computer: Jmol to the rescue. *Biochem. Mol. Biol. Educ. Bimon. Publ. Int. Union Biochem. Mol. Biol.*, **34**, 255–261.
- Hogue, C.W.V. (1997) Cn3D: a new generation of three-dimensional molecular structure viewer. *Trends Biochem. Sci.*, **22**, 314–316.
- Katoh, K. et al. (2002) MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res.*, **30**, 3059–3066.
- Khatter, H. et al. (2015) Structure of the human ribosome. *Nature*, **520**, 640–645.
- Klaholz, B.P. (2011) Molecular recognition and catalysis in translation termination complexes. *Trends Biochem. Sci.*, **36**, 282–292.
- Klaholz, B.P. et al. (2004) Visualization of release factor 3 on the ribosome during termination of protein synthesis. *Nature*, **427**, 862–865.
- Klostermeier, D. and Hammann, C. (2013) RNA Structure and Folding: Biophysical Techniques and Prediction Methods Walter de Gruyter (Ménéret et al. chapter : Integrative structure-function analysis of large nucleoprotein complexes. RNA structure and folding).
- Letunic, I. and Bork, P. (2007) Interactive Tree Of Life (iTOL): an online tool for phylogenetic tree display and annotation. *Bioinforma. Oxf. Engl.*, **23**, 127–128.
- Lith, A. and Mattsson, J. (2010) Investigating storage solutions for large data.
- Liu, X. and Altman, R.B. (1998) Updating a bibliography using the related articles function within PubMed. *Proc. AMIA Symp.*, 750–754.
- Loll, B. et al. (2005) Towards complete cofactor arrangement in the 3.0 Å resolution structure of photosystem II. *Nature*, **438**, 1040–1044.
- Low, H.H. et al. (2014) Structure of a type IV secretion system. *Nature*, **508**, 550–553.
- Maletta, M. et al. (2014) The palindromic DNA-bound USP/EcR nuclear receptor adopts an asymmetric organization with allosteric domain positioning. *Nat. Commun.*, **5**, 4139.
- Marzi, S. et al. (2007) Structured mRNAs Regulate Translation Initiation by Binding to the Platform of the Ribosome. *Cell*, **130**, 1019–1031.
- Meng, E.C. et al. (2006) Tools for integrated sequence-structure analysis with UCSF Chimera. *BMC Bioinformatics*, **7**, 339.
- Meyer, E.F. (1997) The first years of the Protein Data Bank. *Protein Sci. Publ. Protein Soc.*, **6**, 1591–1597.
- Orlov, I. et al. (2012) Structure of the full human RXR/VDR nuclear receptor heterodimer complex with its DR3 target DNA. *EMBO J.*, **31**, 291–300.
- Ostell, J.M. and Kans, J.A. (1998) The NCBI data model. *Methods Biochem. Anal.*, **39**, 121–144.
- Peitsch, M.C. et al. (1995) The Swiss-3DImage collection and PDB-Browser on the World-Wide Web. *Trends Biochem. Sci.*, **20**, 82–84.
- Pettersen, E.F. et al. (2004) UCSF Chimera—A visualization system for exploratory research and analysis. *J. Comput. Chem.*, **25**, 1605–1612.
- Plewniak, F. et al. (2003) PipeAlign: a new toolkit for protein family analysis. *Nucleic Acids Res.*, **31**, 3829–3832.
- Procter, J.B. et al. (2010) Visualization of multiple alignments, phylogenies and gene family evolution. *Nat. Methods*, **7**, S16–S25.
- Saenger, W. et al. (2002) The assembly of protein subunits and cofactors in photosystem I. *Curr. Opin. Struct. Biol.*, **12**, 244–254.
- Ben-Shem, A. et al. (2011) The structure of the eukaryotic ribosome at 3.0 Å resolution. *Science*, **334**, 1524–1529.
- Simonetti, A., Marzi, S., Billas, I.M.L., et al. (2013a) Involvement of protein IF2 N domain in ribosomal subunit joining revealed from architecture and function of the full-length initiation factor. *Proc. Natl. Acad. Sci. USA*, **110**, 15656–15661.
- Simonetti, A. et al. (2008) Structure of the 30S translation initiation complex. *Nature*, **455**, 416–420.
- Simonetti, A., Marzi, S., Fabbretti, A., et al. (2013b) Structure of the protein core of translation initiation factor 2 in apo, GTP-bound and GDP-bound forms. *Acta Crystallogr. D Biol. Crystallogr.*, **69**, 925–933.
- Thompson, J.D. et al. (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.*, **22**, 4673–4680.
- Wang, Y. et al. (2007) MMDB: annotating protein sequences with Entrez's 3D-structure database. *Nucleic Acids Res.*, **35**, D298–300.
- Waterhouse, A.M. et al. (2009) Jalview Version 2—a multiple sequence alignment editor and analysis workbench. *Bioinformatics*, **25**, 1189–1191.
- Wimberly, B.T. et al. (2000) Structure of the 30S ribosomal subunit. *Nature*, **407**, 327–339.
- Yusupov, M.M. et al. (2001) Crystal Structure of the Ribosome at 5.5 Å Resolution. *Science*, **292**, 883–896.

### 1.3 Discussion et perspective

Dans une optique de biologie intégrative il est nécessaire de pouvoir visualiser et de manipuler de manière corrélée de nombreux types d'informations entre elles, notamment pour faciliter le lien entre l'analyse de structure 3D et les alignements de séquences. Afin de faciliter l'accès à ces informations, le web est un outil de prédilection permettant de rendre un service sans autre besoin qu'un navigateur internet. C'est dans cette optique que les langages de programmations choisis sont exclusivement des langages orientés web comme le PHP coté serveur et le JavaScript coté client. L'autre avantage est l'accès direct à la base de données et par conséquent aux mises à jours constantes des données biologiques aussi bien en séquences qu'en structures procaryotes, eucaryotes et archées. Suite à ces mises à jour, il y a également l'évolution d'autres fichiers de ressources présent sur le serveur. Grâce à l'interface utilisateur du site, alliant les logiciels Jmol pour le 3D et Jalview pour le 2D qui fonctionnent de manière combinée, ainsi que la vue d'ensemble des protéines d'une famille plus l'arbre taxonomique, l'accès aux données est unifié. Cela permet de visualiser rapidement dans un même environnement de travail l'ensemble des données présentes pour une famille de protéines. L'ensemble des alignements et des fichiers PDB sont présents sur le serveur et leur utilisation se fait de manière transparente et rapide pour l'utilisateur. A aucun moment il n'a besoin de rentrer de paramètres, par exemple pour corréler une séquence d'un alignement de séquences avec un fichier structural PDB. Aucune manœuvre de réglage particulière n'est nécessaire, il suffit de sélectionner la structure qu'il souhaite observer et tout le reste est pris en charge par le site. L'intégralité des fonctionnalités est rendue la plus simple et la plus intuitive possible pour l'utilisateur. Le but étant que l'utilisateur n'ait besoin d'aucune connaissance en informatique au préalable et que la prise en main de ce service web soit le plus naturel possible. Ainsi toutes les tâches fastidieuses et sources d'erreurs de frappe ou autres erreurs, lors du traitement manuel des données, sont alors minimisées. L'utilisateur n'a par conséquent pas besoin de se soucier de la récupération des données ou de leur mise en forme, seules les observations et les interprétations sont à sa charge pour une étude séquence-structure-fonction d'une protéine seule ou dans le cadre d'un complexe.

La rapidité du site est rendu possible grâce à la centralisation des données dans une base de données MongoDB propre en NoSQL. L'utilisation d'autres bases de données, telle que celle d'UniProt n'est faite que pour les mises à jours de la base de donnée. Dans les autres cas, aucune donnée n'est requêtée dans une autre base directement par l'utilisateur. L'ensemble des fichiers nécessaires à l'utilisation du site sont également présent sur le serveur à la disposition des utilisateurs. La préparation des fichiers en avance comme ceux des alignements et la possibilité de les manipuler en

ne sélectionnant qu'un taxon en particulier permet également un gain en temps considérable, surtout si l'utilisateur veut voir l'alignement de séquences de beaucoup de familles de protéines comprises dans son complexe d'intérêt. De plus, le gain se fait aussi en quantité de calculs par le serveur qui n'a pas besoin de refaire plusieurs fois le même calcul. Il est réalisé une seule fois à l'aide du programme ClustalW, en amont de l'utilisateur et de façon automatisée. L'architecture de la base de données est également pensée pour permettre de garantir les performances qu'elle a actuellement au cours du temps, en prenant en compte l'ajout de plus en plus fréquent de protéines. En effet chaque famille de protéines a ses propres collections dans la base de données. La structure en complexe se retrouve grâce à des collections servant d'index afin de lister les protéines de celui-ci. Certaines collections, principalement pour les fonctionnalités réservées au complexe, sont également divisées en fonction des complexes présents. Ce système permet de répartir les données de manière efficace, évitant ainsi de parcourir des données inutiles lors de requêtes vers la base de données et également d'anticiper l'ajout de nombreuses familles de protéines et de nombreuses protéines à l'avenir.

Il y a plusieurs perspectives envisagées. Dans un premier temps la base de données doit être élargie pour de nouveaux complexes et ajouter une interface permettant de faire toutes les études proposées avec des protéines isolées. Ceci permettra de rendre disponible une majorité des protéines actuellement caractérisées et annotées correctement. Il est également envisagé de créer une partie utilisateur avec connexion sécurisée afin de donner la possibilité à tout utilisateur de créer ses propres complexes d'études. Une fois l'ensemble des protéines voulues sélectionnées, le serveur se chargera de créer les ensembles de données nécessaire à l'utilisation de toutes les fonctionnalités du site internet. Il peut être très intéressant d'ajouter des données et des outils d'interactomiques, qui seront notamment utiles dans le cadre de coévolution et pour les complexes transitoires. Ainsi, l'étude de protéines seules avec les différents éléments du complexe grâce à l'interactomique sera simplifiée.

## 2. Grid Files Manager un logiciel utilitaire de suivis d'échantillons en Cryo-ME

### 2.1 Introduction et contexte

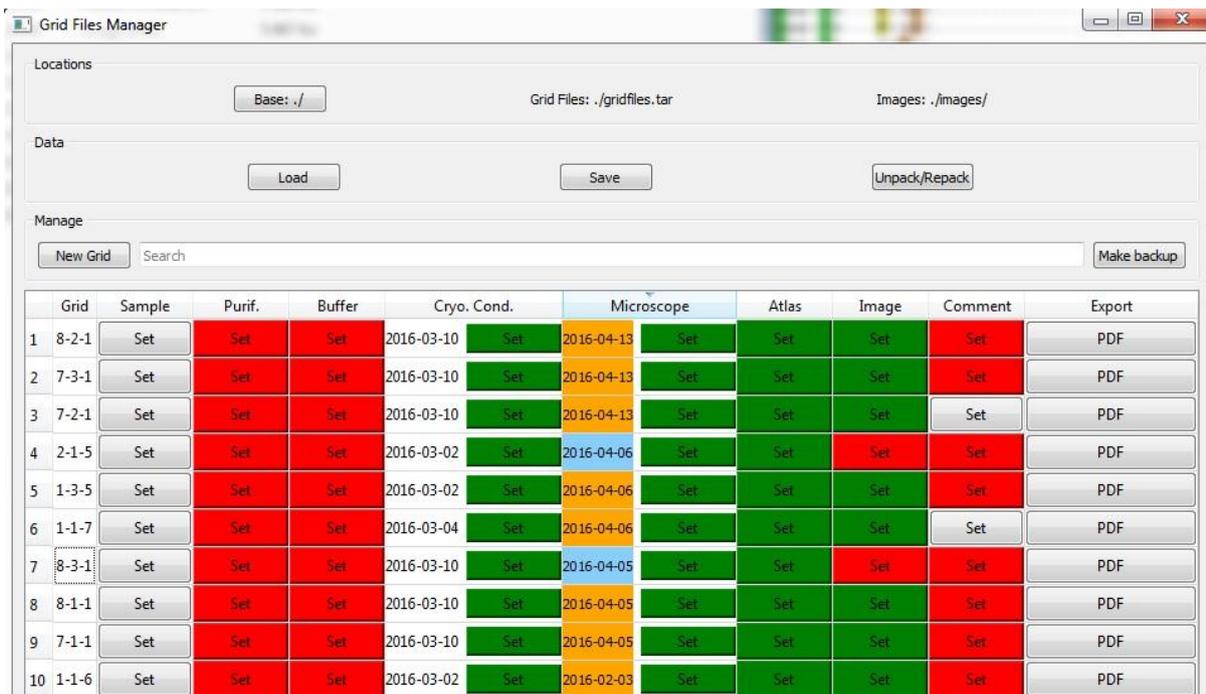
Démarrer un nouveau projet de détermination structurale en cryo-microscopie électronique demande en général une longue phase d'optimisation. Ce qui complexifie la gestion c'est la "boucle d'optimisation" entre biochimie et cryo-ME car la cryo-ME sert à visualiser et caractériser l'échantillon en plus de la caractérisation biophysique. Il y a tout d'abord l'optimisation de la production et de la purification de la protéine ou du complexe d'étude. Une fois ces étapes accomplies vient l'optimisation des conditions de cryogénéisation permettant de produire des grilles, reproductibles et de bonne qualité. Cette phase peut s'avérer très longue et fastidieuse dans certains cas. Ce fut le cas pour plusieurs complexes étudiés par cryo-microscopie électronique au cours de ma thèse. Le choix de la composition du tampon contenant l'échantillon est très important. En effet il doit être à la fois optimisé pour la phase de purification pour assurer la stabilité du complexe, et être adapté pour des études en cryo-microscopie électronique. Il est bon de savoir que l'utilisation de certaines molécules comme le glycérol et le sucrose, utilisées durant la purification des complexes pour assurer leur solubilité et stabilité, est fortement déconseillée pour la visualisation des particules notamment pour des complexes relativement petits. La concentration en sel est aussi un paramètre important : plus la concentration en sel est élevée, moins il y aura de contraste sur le micrographe. Le respect de ces règles ne doit pas interférer avec la stabilité de la protéine et sa solubilité afin de permettre une bonne distribution des particules sur la grille et éviter la formation d'agrégats.

Afin de faciliter ce travail et d'avoir une vue d'ensemble sur les corrélations entre paramètres expérimentaux et les résultats obtenus. Un collègue informaticien, Jonathan Michalon et moi-même avons mis au point un outil de suivi d'échantillons.

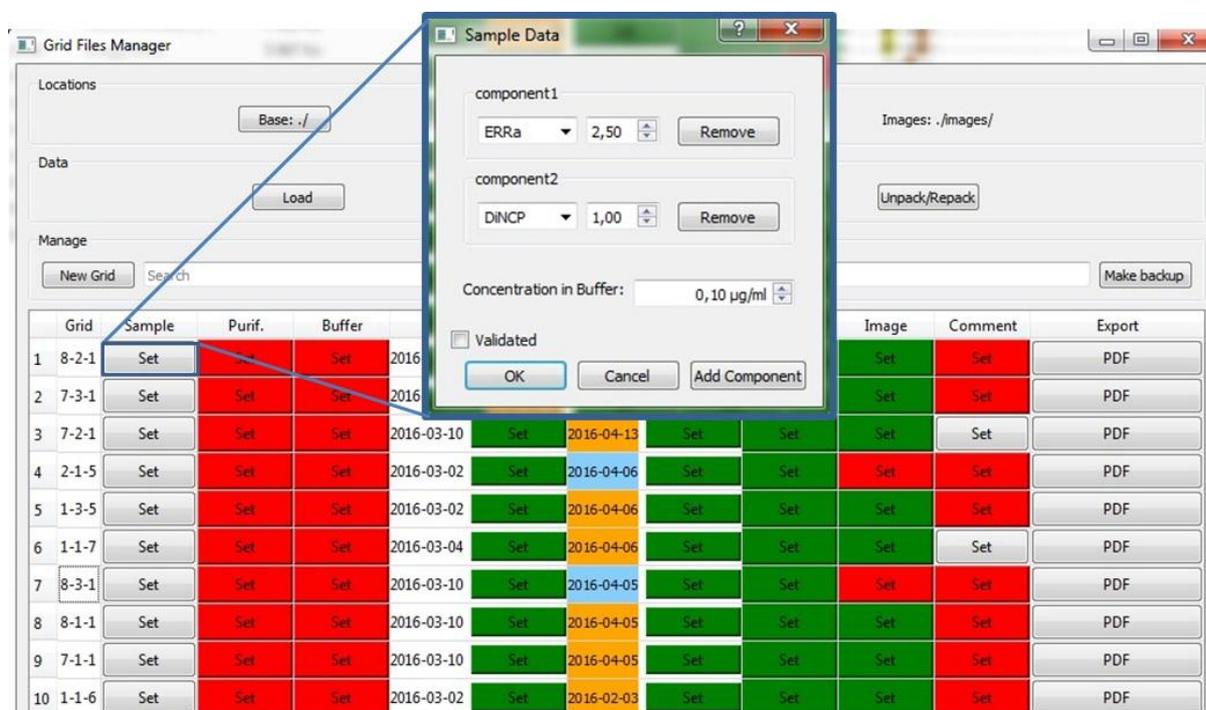
### 2.2 Résultats

Le logiciel qui en résulte, appelé "Grid Files Manager", permet d'organiser et de recueillir une série de données relatives à la préparation, la cryogénéisation et l'observation des échantillons. L'interface est divisée en trois parties. La première partie (partie "Locations" sur la figure 110), en haut de

L'interface permet de modifier le chemin relatif de la base de données. Par exemple s'il y a plusieurs bases de données en parallèle, on peut changer de dossier de référence sans redémarrer ni relancer une nouvelle instance. La deuxième partie du logiciel (partie "Data" sur la figure 110) permet les opérations simples sur la base de données, à savoir la sauvegarde, le rechargement, et la décompression et compression de la base de données s'il y a nécessité d'intervenir directement dessus manuellement. Ce dernier point sera vu plus en détails dans la partie discussion de la présentation de Grid Files Manager. La troisième partie (partie "Manage" sur la figure 110) permet la gestion et la visualisation des données. Cette dernière est composée d'une première interface qui permet de créer une nouvelle grille dans la base de données en lui donnant un nom en fonction de la préférence et des habitudes de chacun. Une fois créée, une nouvelle ligne apparaît où l'on peut renseigner la composition de l'échantillon dans la colonne "Sample", les détails et remarques pertinentes sur les différentes étapes de purification dans la colonne "Purif.", la composition du tampon final utilisé avant cryogénéisation dans la colonne "Buffer", les conditions de cryogénéisation dans la colonne "Cryo. Cond.", les paramètres de microscopie utilisés dans la colonne "Microscope", l'image de l'atlas de la grille dans la colonne "Atlas", une image à haut grossissement représentatif de l'échantillon dans la colonne "Image" ainsi que des commentaires divers et variés pour toutes informations n'ayant pas de case appropriée. La colonne "Export" quant à elle permet d'exporter l'ensemble des données relatives à une grille dans un fichier PDF composé de 2 pages A4, utile pour l'archivage ou le partage sous forme électronique ou papier de ses expériences et résultats.



**Figure 110 :** Interface principale dans Grid Files Manager. De manière générale l'ensemble des boutons "Set" peuvent adopter 3 états visualisés par un code couleur. Les boutons rouges indiquent les cellules pour lesquelles aucune information n'est saisie. Les boutons gris indiquent les cellules pour lesquelles une partie ou la totalité des champs sont saisis mais qui ne sont pas jugé complets ou en version finale pour l'utilisateur. Les boutons verts indiquent les cellules pour lesquelles toutes les informations sont saisies de manière définitive (un retour en arrière est possible à tout moment).



**Figure 111 :** Interface de saisie de la composition de l'échantillon dans Grid Files Manager. On notera la présence de la case à cocher "Validated" qui permet le passage au vert du bouton "Set".

La colonne "Sample" permet de renseigner la composition de l'échantillon, c'est à dire la liste des protéines/ARNs/ADNs/complexes présents en fonction du cas et leur stœchiométrie. Les valeurs peuvent être changées avec le clavier ou avec les petites flèches prévues à cet effet. Le nom de l'entité biologique est présenté dans un menu déroulant qui se complète au fur et à mesure. En effet, dès que j'écris un nom qui n'est pas encore dans la base de données, il est automatiquement ajouté au menu déroulant. Nous pouvons également renseigner la concentration du complexe dans cette fenêtre (Figure 111). L'interface de la composition du tampon, colonne "Buffer" est similaire, seule une troisième case sur chaque ligne est disponible pour indiquer la quantité de chaque composé, qu'elle soit indiquée en gramme, en molaire ou dans une autre unité. Dans ce contexte, les unités sont également dans une interface à menu déroulant qui se complète au fur et à mesure des saisies. La colonne "Purif." n'a pas de champ particulier, il s'agit d'une grande zone de texte où l'on peut écrire les grandes étapes de la purification et tous les détails que l'on souhaite conserver pour la comparaison des expériences.

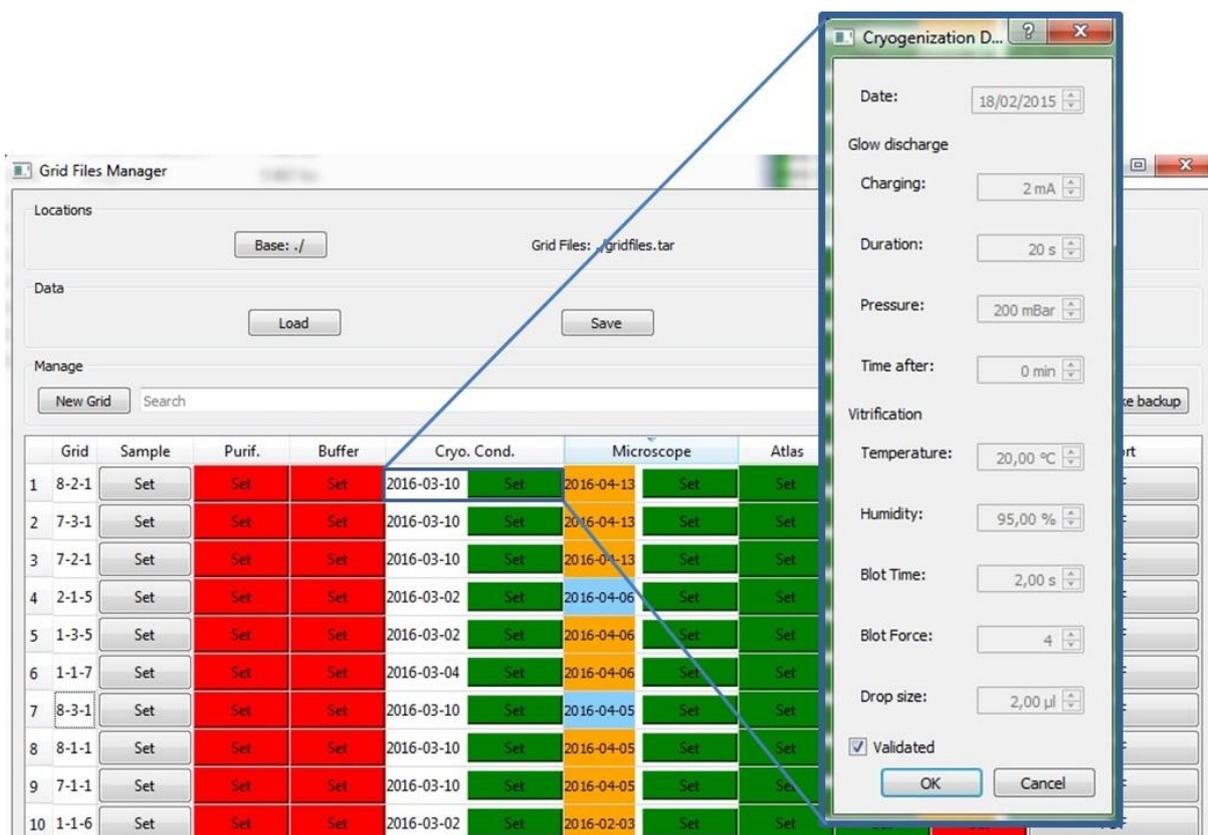


Figure 112 : Interface de saisie des détails de cryogénéisation des échantillons dans Grid Files Manager.

La colonne "Cryo. Cond." permet de renseigner toutes les informations relatives aux détails de cryogénéisation des échantillons. On peut y renseigner la date de cryogénéisation, les paramètres de décharge lumineuse (glow discharge) utilisés ainsi que les paramètres de vitrification utilisés qui

dans mon cas sont relatifs à la machine utilisée, c'est à dire un Vitrobot (FEI). Nous pouvons également renseigner la quantité d'échantillon déposé sur grille.

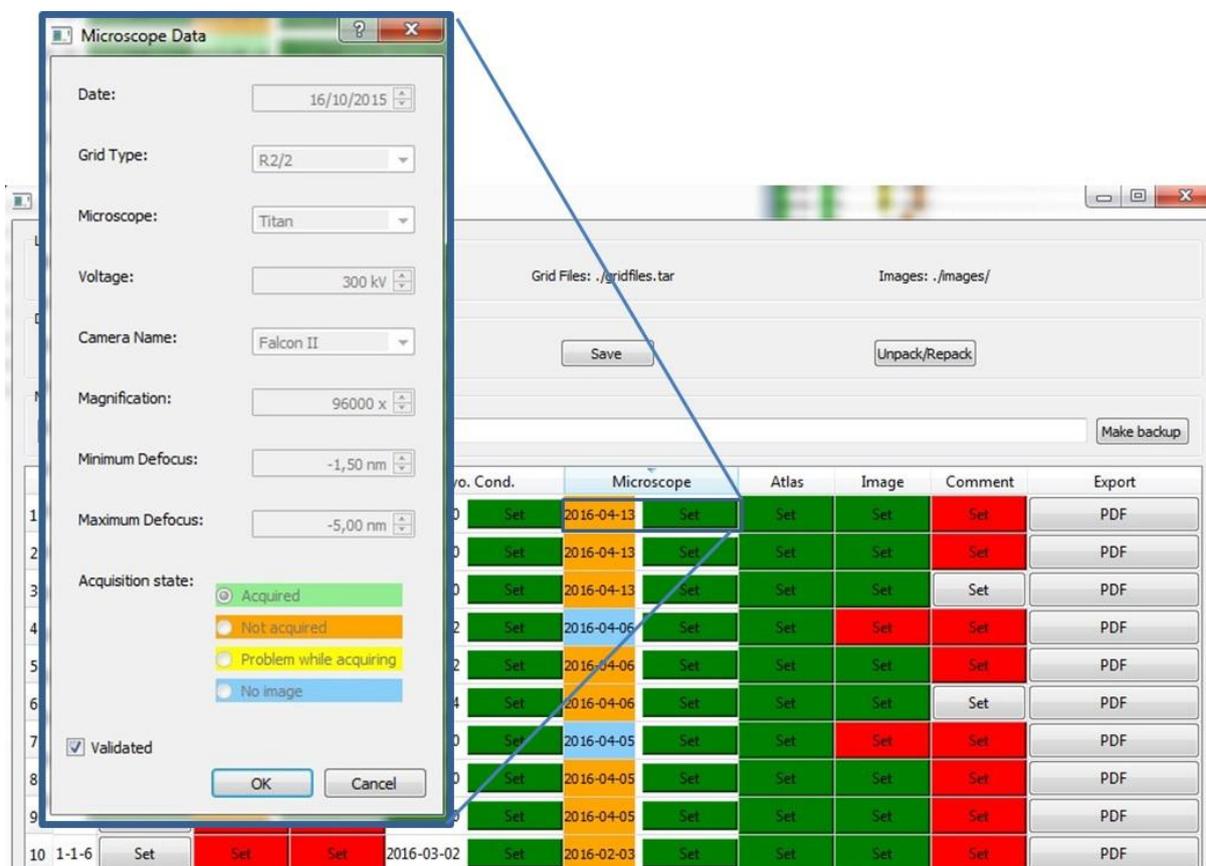
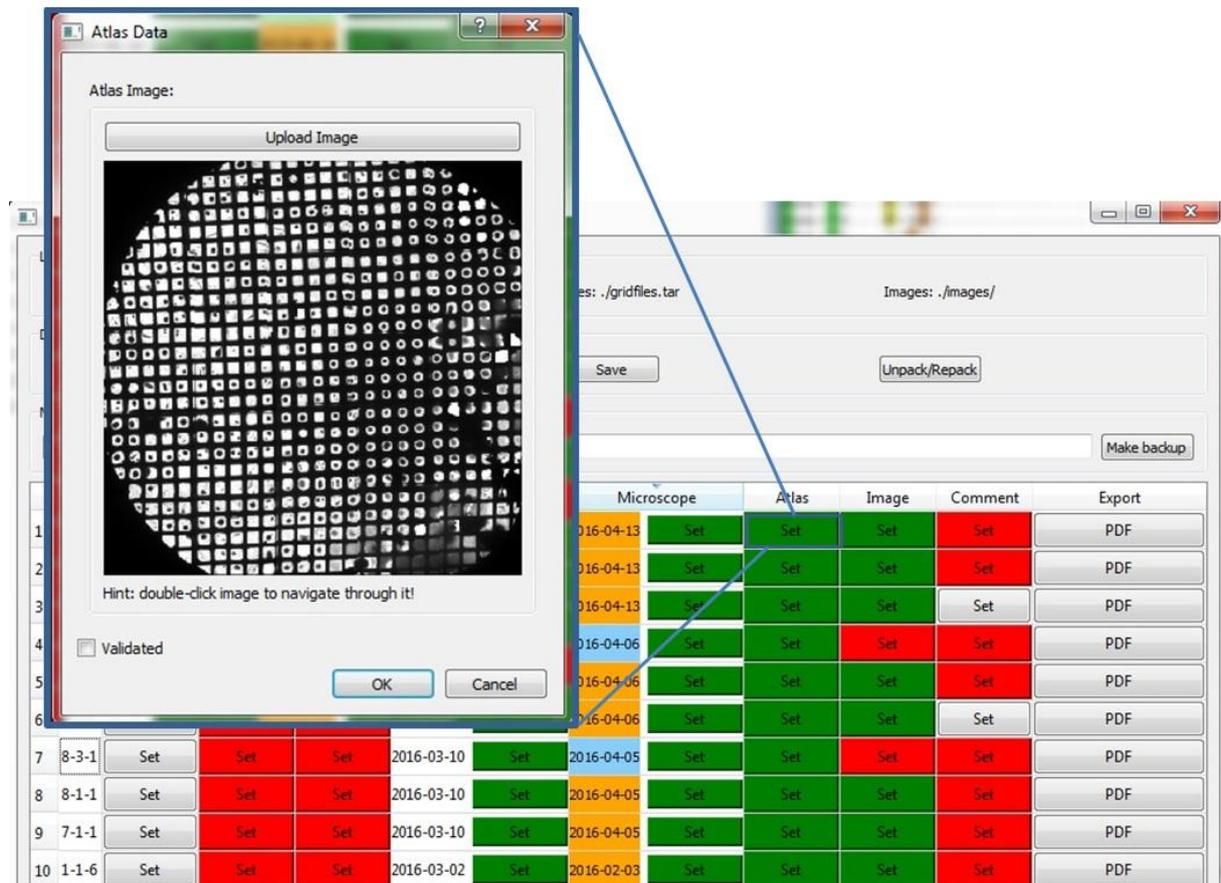


Figure 113 : Interface de saisie des détails d'acquisition et du microscope dans Grid Files Manager.

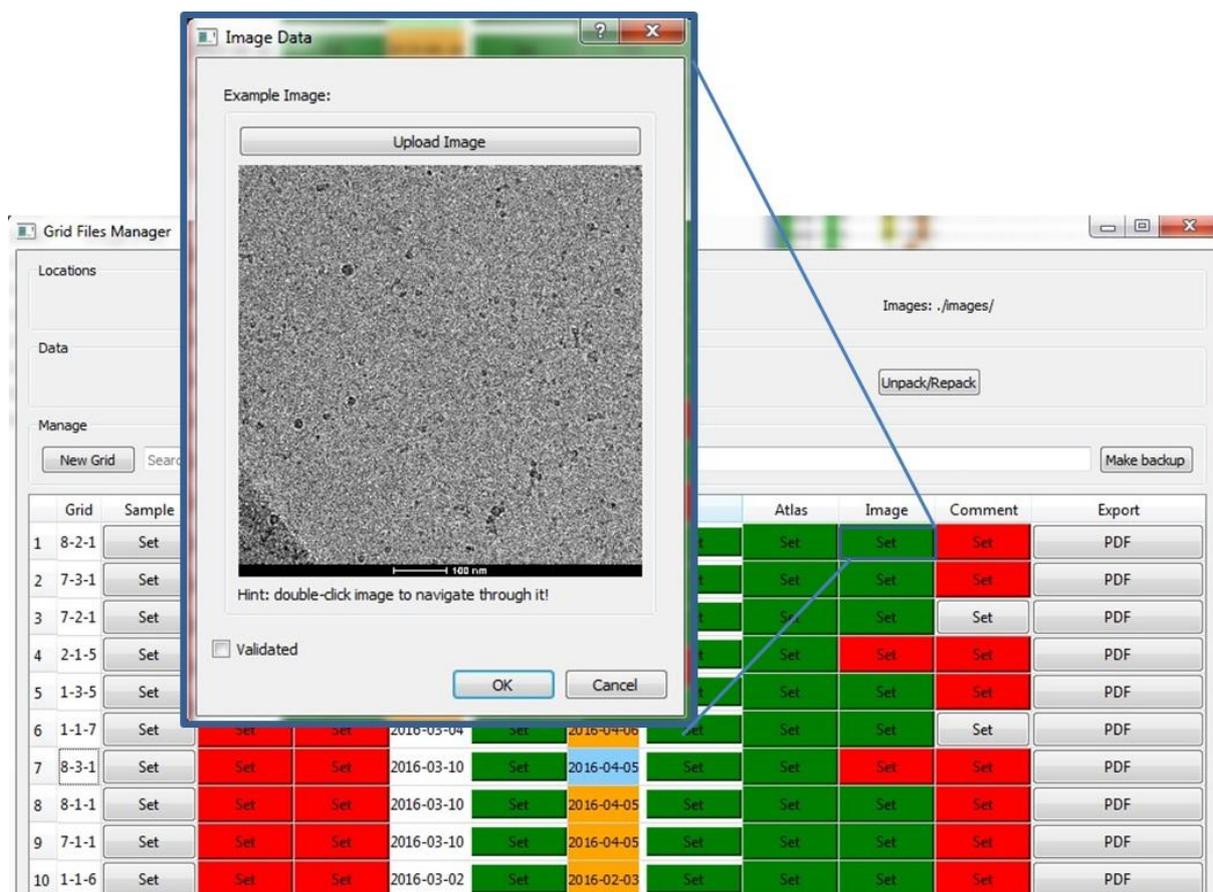
La colonne "Microscopie" permet de renseigner les détails relatifs aux paramètres d'acquisition, notamment la date et le type de grille visualisée (or, carbone, taille des trous, grille commerciale, carbone flotté etc.). On ajoute également le microscope utilisé, la caméra choisie, le grandissement d'acquisition et la plage de défocalisation choisie. Enfin nous pouvons renseigner l'état générale de l'acquisition suivant un code couleur, lui-même reporté directement dans la cellule "Microscope" de l'interface principale pour une vue d'ensemble complète de l'acquisition. Si la date est sur fond vert, la grille a été acquise en vue des étapes de traitement d'images décrites au chapitre 2 de cette thèse. Orange veut dire que certaines images ont été acquises, dans le cadre d'un criblage de grilles par exemple. Jaune indique un problème quelconque lors de l'acquisition, par exemple un arrêt non remarqué lors d'une acquisition automatique ou encore tout autres problèmes optiques non remarqués rendant les données inutilisables pour du traitement d'images. Et enfin bleu correspond à l'absence totale d'images à haut grandissement, par exemple à cause d'une glace trop épaisse sur l'ensemble de la grille.



**Figure 114 :** Interface d'ajout de l'image de l'atlas de la grille dans Grid Files Manager.

La colonne "Atlas" permet d'ajouter un fichier image de format JPEG ou PNG de l'atlas permettant d'avoir une vision globale de la qualité de la glace sur l'ensemble de la grille. Cela permet principalement de voir si les conditions de cryogénéisation doivent être modifiées, notamment celles de la machine qui exerce la pression pour absorber l'excès d'échantillons sur la grille.

Il est également possible d'ajouter une image des particules et de leur distribution, ceci dans le but d'avoir une idée de l'aspect général de la grille à un fort grossissement.



**Figure 115 :** Interface d'ajout d'une image exemple de l'échantillon à grandissement de travail dans Grid Files Manager.

### 2.3 Discussion et perspectives

Le logiciel a été développé avec le langage Python 3 pour plusieurs raisons. C'est un langage que beaucoup de personnes maîtrisent dans la communauté scientifique, il est plutôt simple et léger à utiliser pour une activité qui ne nécessite pas de calculs intenses. De plus il permet de créer des GUI (Graphical User Interface) efficaces relativement rapidement via un interfaçage avec la bibliothèque graphique Qt de bonne qualité.

L'utilisation de fichiers simples comme base de données permet d'éditer manuellement et très rapidement les données au besoin, par exemple pour compléter une liste de microscopes préenregistrée pour une plateforme différente ou autre. Cela permet également d'éviter de recourir à un système de gestion de base de données (SGBD) extérieur comme cela est souvent le cas. Nous pouvons également modifier les ajouts automatiques aux menus déroulants ; par exemple en cas d'erreur on peut supprimer un item ou le corriger.

### Chapitre 3 - Grid Files Manager un logiciel utilitaire de suivis d'échantillons en Cryo-ME

Ces fichiers sont formatés en YAML, un langage de balisage très intuitif. L'ensemble des fichiers est ensuite stocké sous forme d'archive TAR, lue et générée à la volée. Ainsi toutes les informations sont regroupées dans un seul fichier facile à centraliser, sauvegarder ou partager.

Le logiciel est déjà distribué sous forme de fichier exécutable unique pour Windows et Linux ne nécessitant aucune installation. Toutes les dépendances sont directement incluses dans le fichier.

Le projet est disponible via cette URL : <https://cbi-dev.igbmc.fr/cbi/gridfiles>

Ce projet s'intègre dans une logique plus large au niveau de la plateforme d'imagerie électronique au CBI-IGBMC. Il sera une base pour l'intégration au pipeline de microscopie électronique développé en interne, qui débute dès l'acquisition, et pour lequel je suis partie prenante. Cette intégration permet de suivre et d'avoir une base statistique sur l'ensemble des expériences dans leur intégralité de la purification à la structure.

Dans ce cadre un développement pour une utilisation multi utilisateurs et multi projets a débuté pour une intégration aux besoins de la plateforme et ainsi faciliter la gestion des utilisateurs externes dans le cadre des infrastructures FRISBI et Instruct-ERIC. Par exemple un formulaire a été ajouté permettant de renseigner les détails utilisateur et l'institut concerné pour un échantillon donné.

Une réflexion pour ajouter plusieurs items comme par exemple choisir le type de machine utilisée pour la cryogénéisation permettrait de rendre cet outil plus flexible et ainsi compatible pour une utilisation plus large avec de futurs équipements ou dans d'autres instituts de recherche qui ont des équipements différents. De plus pour faciliter le comparatif des données, nous réfléchissons également à une sortie standard pour comparer l'ensemble des paramètres ou seulement pour certaines étapes, plusieurs échantillons sous forme d'un tableau.

### 3. BackPhylo, un logiciel d'évolution et de phylogénie

#### 3.1 Introduction et contexte

Les études évolutives demandent la manipulation et le croisement de grandes quantités de données. Dans mon cas, l'étude évolutive porte sur l'évolution d'une famille de protéines, les récepteurs nucléaires, près de 70 membres répartis à travers l'ensemble des métazoaires. L'ensemble des connaissances actuelles représente plusieurs dizaines de milliers de séquences disponibles réparties entre plusieurs centaines d'espèces animales. Devant une telle quantité d'informations, il est nécessaire de hiérarchiser et de trier l'ensemble de ces données pour pouvoir les comparer les unes aux autres le plus facilement possible. On peut se trouver devant plusieurs problèmes dans ce type d'étude. Le premier étant la qualité des données récupérées dans les grandes banques de données telles que le NCBI. En effet, pour les organismes très étudiés tel que *Homo sapiens*, il arrive de trouver plusieurs fois la même séquence avec différents identifiants, ceci biaisant les alignements de séquences et allongeant inutilement la durée du calcul si un tri n'est pas réalisé. Il est aussi possible d'avoir une protéine mal annotée, car annotée automatiquement lors du séquençage à haut débit. Ce dernier point, s'il n'est pas corrigé, entrainera aussi des problèmes dans l'alignement de séquences comme la création de gaps dus uniquement à une ou plusieurs séquences erronées ne devant pas être présentes. Un autre problème pouvant être rencontré est la variété de noms donnés à une seule et même protéine. En effet, chaque récepteur nucléaire peut avoir plusieurs noms différents, et le même nom n'est pas toujours utilisé par défaut dans les bases de données. Ceci est essentiellement dépendant de la date de séquençage et du groupe qui l'a annoté. Par exemple, certaines protéines sont renommées pour uniformiser les bases et la nomenclature, mais les anciens noms restent présents dans les habitudes et dans les anciennes entrées de la base de données. Ainsi ne pouvant pas me contenter des séquences disponibles dans les bases de données annotées manuellement telles qu'UniProt, et afin d'avoir une vue d'ensemble sur cet ensemble de données important et afin d'exécuter plusieurs opérations de tri, classement et traitement de données, il a fallu trouver une solution intégrative.

J'ai ainsi mis au point avec l'aide d'un collègue informaticien (Jonathan Michalon), un outil d'aide à l'étude évolutive d'un groupe de protéines.

### 3.2 Résultats

Le logiciel qui résulte de ce travail que l'on a appelé "BackPhylo" permet d'organiser et recueillir une série de données liées à l'évolution d'une famille de protéines. Il permet aussi de déterminer à partir des séquences des organismes et taxons disponibles qui ne sont pas exhaustifs, la séquence probable de l'ancêtre commun à chaque division de la taxonomie. Il faut néanmoins garder à l'esprit dans ce genre d'approche que le nombre important de séquences encore manquantes peut biaiser en partie certains résultats qui sont susceptibles de changer après l'ajout de certains taxons et organismes nouvellement séquencés. L'autre point à prendre en compte est que les arbres phylogénétiques actuellement utilisés ne sont pas figés et peuvent encore évoluer avec de futures découvertes, surtout pour la partie basale de l'arbre et est sujette à débat et parfois remise en cause.

Pour utiliser ce logiciel, il est nécessaire de faire un premier travail préalable de récupération des numéros d'accèsion (ID) de toutes les protéines que l'on veut inclure dans notre étude. Il est facile de récupérer des listes de protéines en tant que résultats d'une recherche sur le NCBI ou l'EBI par exemple. Le système de recherche n'étant pas fiable à 100%, surtout quand la protéine ou famille de protéines n'a pas fait l'objet d'un regroupement attentif car ne représentant pas une protéine majeure, il est parfois nécessaire de faire des recherches séparées avec plusieurs mots-clés différents. On récupère et regroupe l'ensemble des ID. C'est là le point de départ de l'utilisation de BackPhylo.

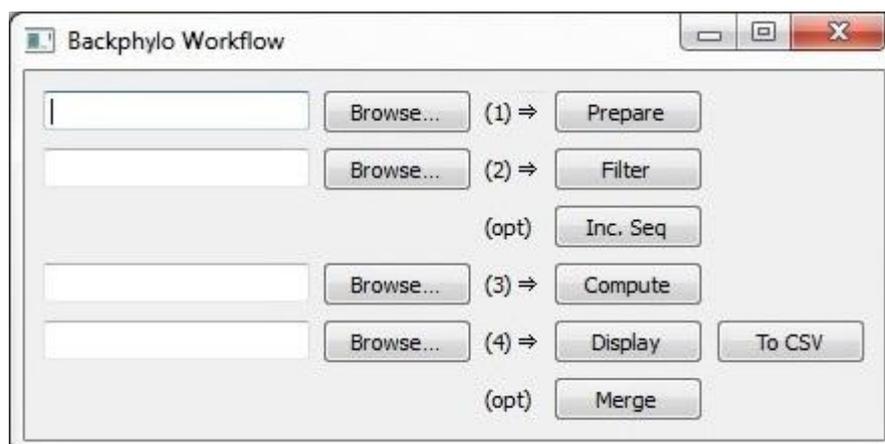
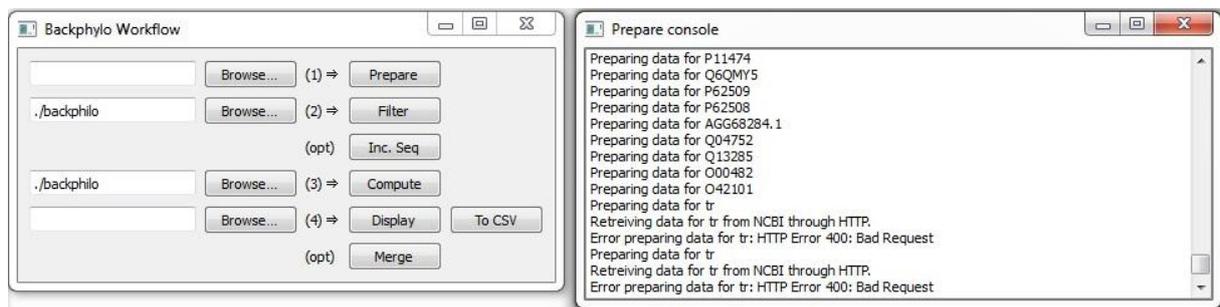


Figure 116 : Interface générale de BackPhylo.

Le logiciel est divisé en 4 parties distinctes indiquées sur l'interface. La première partie est la préparation des données. On donne comme entrée au logiciel un fichier contenant l'ensemble des ID préalablement récupérées. BackPhylo se charge de récupérer automatiquement via internet la fiche

XML de chaque protéine comme on peut le voir sur la figure 117. Le logiciel crée alors tous les petits fichiers dont il aura besoin pour les étapes suivantes. Pour chaque ID il va créer 4 fichiers : un fichier contenant la définition de l'ID, un fichier de lignée qui liste l'ensemble des taxons auquel l'organisme appartient, un fichier nom qui reprend le nom de l'entrée dans la base de données pour cet ID et enfin un fichier séquence qui contient la séquence protéique correspondant à l'ID. Suivant le nombre de séquences, cette étape nécessitant une requête HTTP pour chaque ID, cela peut prendre un peu de temps, dans mon cas environ 2 heures pour environ 70 000 séquences. A noter que dans le cas où l'on utilise plusieurs listes d'ID, une entrée déjà téléchargée ne sera pas traitée une seconde fois.



**Figure 117** : Interface générale de BackPhylo, phase de préparation.

Quand toutes les données sont récupérées, on peut commencer à utiliser la partie 2 de BackPhylo : la partie filtrage. Elle permet de trier facilement l'ensemble des données récupérées. L'interface est composée de plusieurs parties. Une première barre de recherche permet de filtrer l'ensemble des données sur l'ensemble des informations traitées, et non uniquement sur ce qui est affiché. Ainsi si on veut filtrer à partir d'un taxon ou d'un nom de protéine présent dans le titre on peut le faire facilement. En dessous se trouve un tableau comprenant l'ensemble des données. Dans ce tableau on peut voir de gauche à droite, une colonne avec le nom de l'organisme où chaque changement de couleurs indique le changement d'espèce, une colonne avec le nom de la séquence dans la base de données, une colonne avec l'ID de la séquence, une colonne avec la longueur de la séquence et enfin une colonne avec une case à cocher pour spécifier les séquences qui seront conservées pour l'étape suivante. Il est possible de sélectionner des plages de lignes avec la touche Majuscule du clavier et de faire une sélection multiple avec la touche Contrôle. La troisième partie de l'interface tout en bas contient les boutons de contrôle. On peut cocher ou décocher l'ensemble des données ou une plage sélectionnée. On peut ajouter une ou plusieurs marques (tag) aux lignes sélectionnées. Ce tag se matérialise par une nouvelle colonne. Lorsque l'on ajoute un tag on peut définir sa position, car si une ligne a 3 tags, il y aura 3 colonnes supplémentaires. Dans mon cas je l'utilise pour annoter rapidement les récepteurs nucléaires une fois identifiés (par exemple RXR, ERR etc.). Dans certains

cas, j'utilise plusieurs colonnes de tags pour définir des sous-groupes de plusieurs récepteurs nucléaires. Chaque fois qu'un tag est créé, un fichier "tag" se crée et contient l'ensemble des données de chaque colonne de tags. Comme chaque colonne peut être utilisée pour un tri alphanumérique, ces tags me permettent de regrouper tous les récepteurs d'un même type, par exemple ERR, même ceux qui sont mal annotés par exemple. La dernière ligne comprend 4 fonctions de filtrages. Le premier bouton "filter" écrit un fichier qui contient l'ensemble des ID pour lesquels la case à cocher "keep" est sélectionnée. C'est à partir de ce fichier que débute l'étape suivante. Deux boutons "Move Selec. to Bad" et "Move Selec. to Unused" permettent de déplacer des fichiers. A la racine du projet, il y a la création d'un dossier "bad" et un dossier "unused" où sont déplacés les fichiers correspondant aux lignes sélectionnées lorsque l'on appuie sur un de ses boutons. Les fichiers ne sont pas supprimés et si l'on fait une erreur il suffit de déplacer les fichiers dans le dossier d'origine pour qu'ils réapparaissent dans le tableau. Cela permet d'éliminer les faux-positifs de la recherche de départ (dossier "bad") ou les protéines en double (dossier "unused"), mais on garde la possibilité de revenir dessus à tout moment. Pour cela il suffit de relancer la partie filtrage depuis l'interface générale en modifiant le chemin des fichiers à chercher, c'est-à-dire modifier ./backphylo par ./bad par exemple. Le dernier bouton permet d'extraire un sous-ensemble de données, comme par exemple si on veut extraire tous les ERR ou tous les RXR.

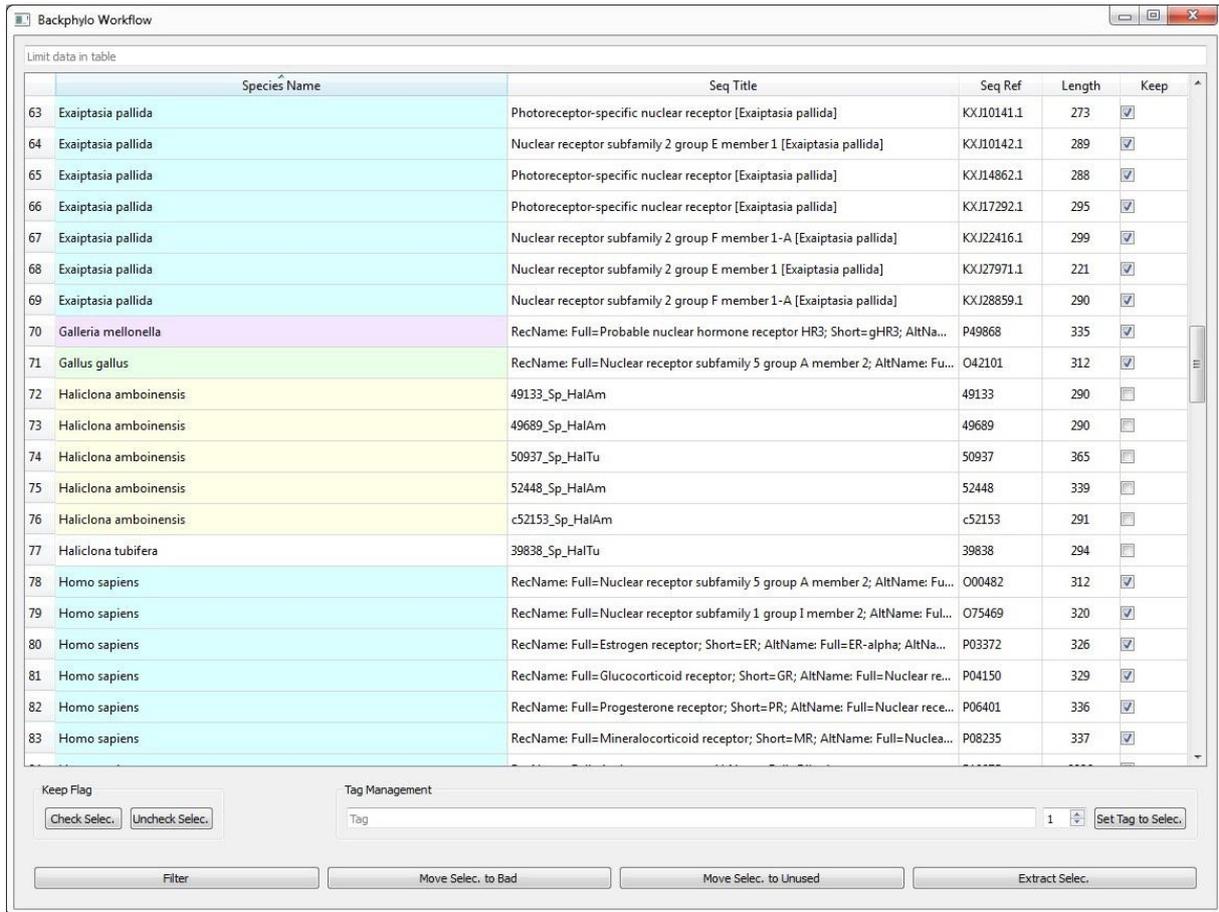
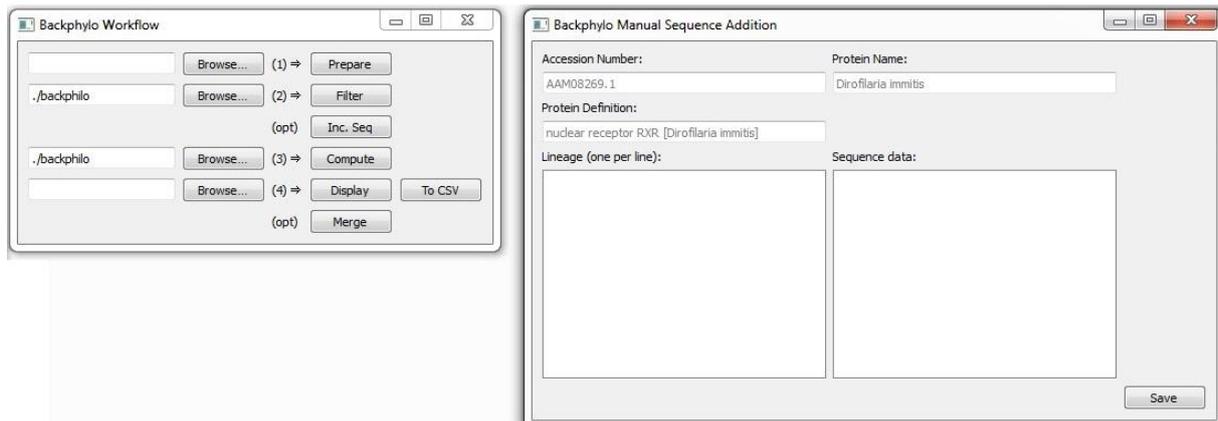


Figure 118 : Interface générale de BackPhylo, partie filtrage.

Dans le cadre d'un projet collaboratif ou si l'utilisateur produit lui-même de nouvelles séquences par séquençage, il se peut qu'on ait à manipuler des séquences non publiées et donc non disponibles dans les banques de données. Dans ce cas, l'étape optionnelle avant la partie 2 de BackPhylo permet d'ajouter manuellement des séquences non publiées. Il suffit de remplir le formulaire présenté dans la figure 119 et les informations renseignées seront intégrées de la même manière que les données provenant de bases de données en ligne.



**Figure 119 :** Interface générale de BackPhylo, ajout manuel de séquence.

Pour l'étape suivante, une fois que le jeu de données est prêt, il est nécessaire de faire une étape en dehors du logiciel. A partir du bouton "filter" cité précédemment, un fichier de séquences au format Fasta est créé avec l'ensemble des séquences. Il est alors nécessaire d'aligner ces données via un logiciel tiers. Dans mon cas, j'utilise le logiciel ClustalX. Une fois le fichier d'alignement généré, il suffit de le placer dans le dossier de travail de BackPhylo, il sera automatiquement détecté grâce à son extension ".aln".

Une fois placé à la racine du projet, on appuie sur le bouton "Compute" de la partie 3 de Backphylo. Cette partie est écrite en langage C et va générer un fichier au format Newick à partir de l'alignement de séquences. Il s'agit d'un format de fichier pour linéariser dans un fichier texte un arbre phylogénétique. Lors de cette étape l'arbre phylogénétique de l'ensemble des organismes présents va se créer. Les feuilles de cet arbre phylogénétique ne seront pas les organismes eux-mêmes contrairement à un arbre phylogénétique classique, mais ce seront les séquences (de récepteurs nucléaires dans le cas présent) de ces organismes. Ces séquences sont reliées à l'organisme puis continuent jusqu'au taxon des métazoaires, commun à toutes les espèces dans ce projet d'étude. Ici plusieurs cas de figures se présentent. En effet, pour chaque taxon de l'arbre, il y a également une séquence théorique ancestrale qui est générée. L'algorithme principal de calcul des séquences ancestrales probables est décrit ci-après. Premièrement un arbre est modélisé en mémoire, chaque taxon étant un nœud possédant une séquence d'acides aminés pondérés par une proportion d'apparition pour chaque acide aminé (sous forme de tableau à double entrée de taille (longueur-séquence)  $\times$  (24, nombre de lettres / acides aminés possibles)). Au chargement, pour chaque feuille possédant donc une séquence donnée, la proportion de 1.0 est attribuée à la case correspondant à la position de l'acide aminée dans la séquence et à sa lettre dans la liste de poids. Deuxièmement, un parcours en profondeur de l'arbre est effectué. Pour chaque nœud, la séquence probable est

déterminée en fonction des pondérations de tous les enfants. On parcourt la séquence, puis pour chaque lettre possible (les 24) on calcule la pondération (somme des poids de la lettre chez chaque enfant, divisé par le nombre d'enfants). S'il n'existe pas déjà une lettre avec un poids aussi haut on suppose que ce sera la lettre probable. S'il en existe déjà une avec le même poids, on marque la lettre comme étant indéterminée (par un 'X'). Dans le cas où un acide aminé est déterminé de cette manière, on "purge" la liste de pondération pour éviter de propager des données qui ne seront plus utiles chez les ancêtres. Ainsi une séquence est attribuée à chaque nœud ancestral, un 'X' était positionné là où il n'a pas été possible de déterminer un acide aminé prépondérant chez les enfants de ce nœud. Actuellement, ce mécanisme fonctionne pour une famille de récepteurs, c'est-à-dire si l'on isole tous les RXR et que l'on crée le fichier de format Newick à partir de ce sous-groupe. Il est prévu de permettre dans le futur la prise en compte par cette partie d'évolution des tags attribués précédemment afin de gérer l'ensemble des séquences en une fois et ainsi de détecter les duplications et délétions de gènes qui ont eu lieu au cours de l'évolution d'une superfamille de protéines.

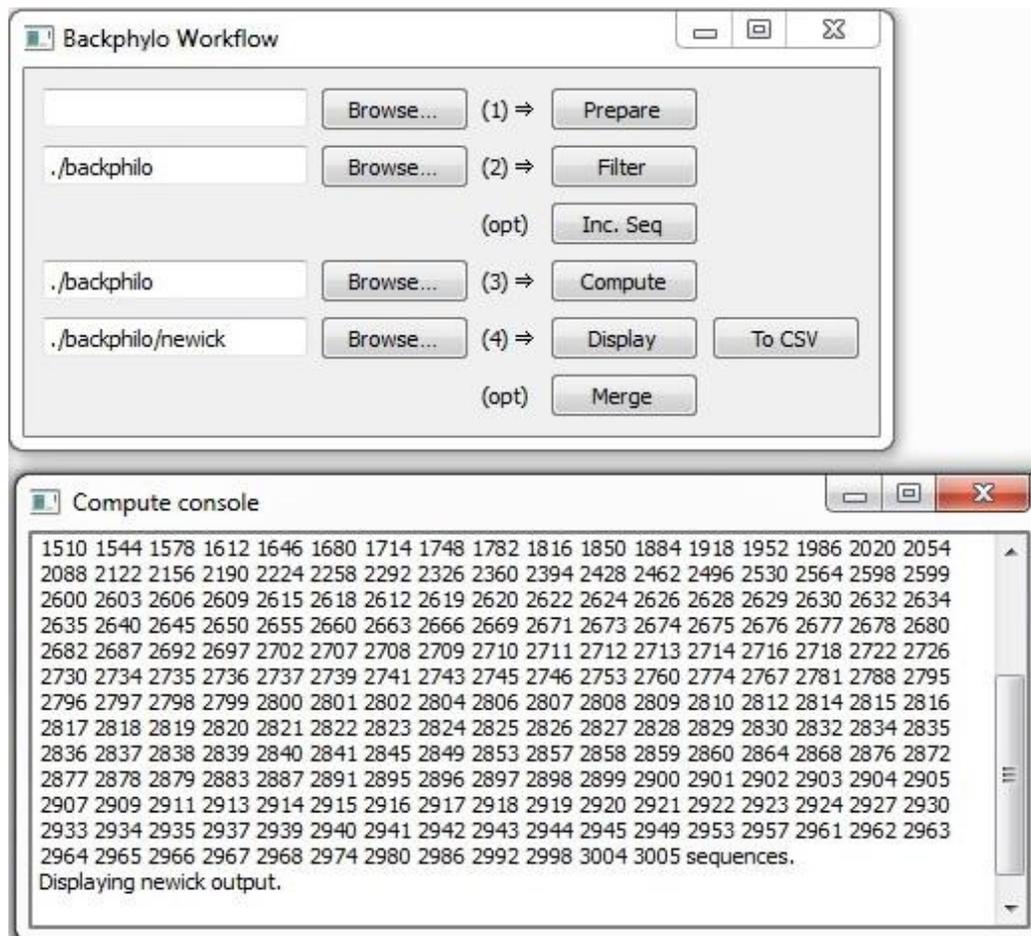
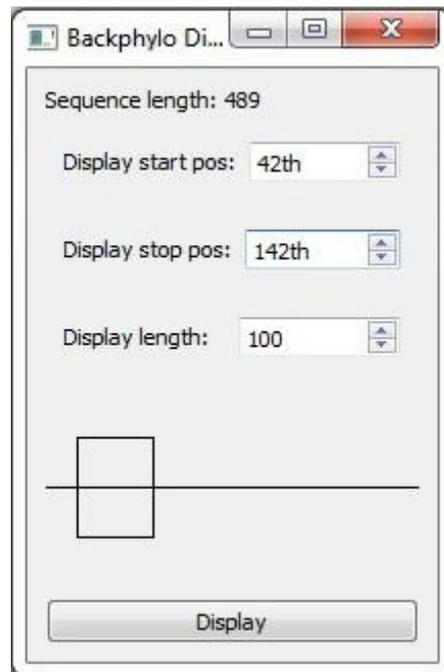


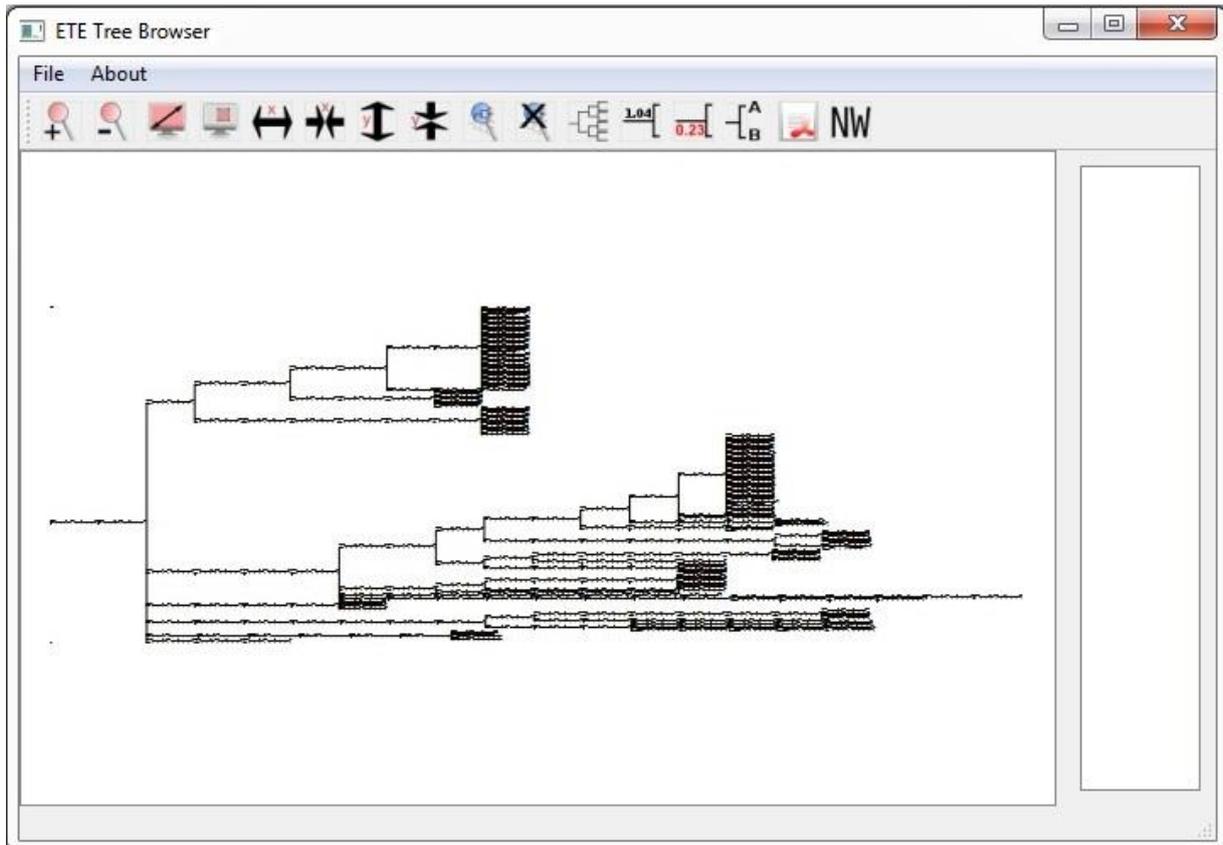
Figure 120 : Interface générale de BackPhylo, Calcul du fichier newick.

La partie 4 de BackPhylo est consacrée à la partie visualisation et analyse du fichier généré via la partie 3. Il y a deux types de sortie distinct : le bouton "display" laisse apparaître une boîte de dialogue qui demande quel résidu ou quelle plage de résidus on veut visualiser dans l'arbre phylogénique. En effet, avec un alignement de séquence de protéines de plus de 1000 acides aminés, une fois tous les GAP pris en compte, il n'est pas possible de s'y retrouver en visionnant l'ensemble des données. Ainsi, la sélection de résidus facilite l'analyse.



**Figure 121** : Interface générale de BackPhylo, interface de sélection du display du newick. La ligne en bas représente la longueur de l'alignement de séquences et le carré représente la portion que l'on souhaite afficher dans l'arbre.

Une fois la région d'intérêt sélectionnée, l'arbre phylogénétique s'affiche grâce à la bibliothèque python ETE2 (Huerta-Cepas et al., 2010, 2016).



**Figure 122** : Interface générale de BackPhylo, interface de phylogénie, display du newick.

On peut naviguer facilement à travers cet arbre phylogénétique et analyser la zone d'intérêt des séquences ancestrales générées.

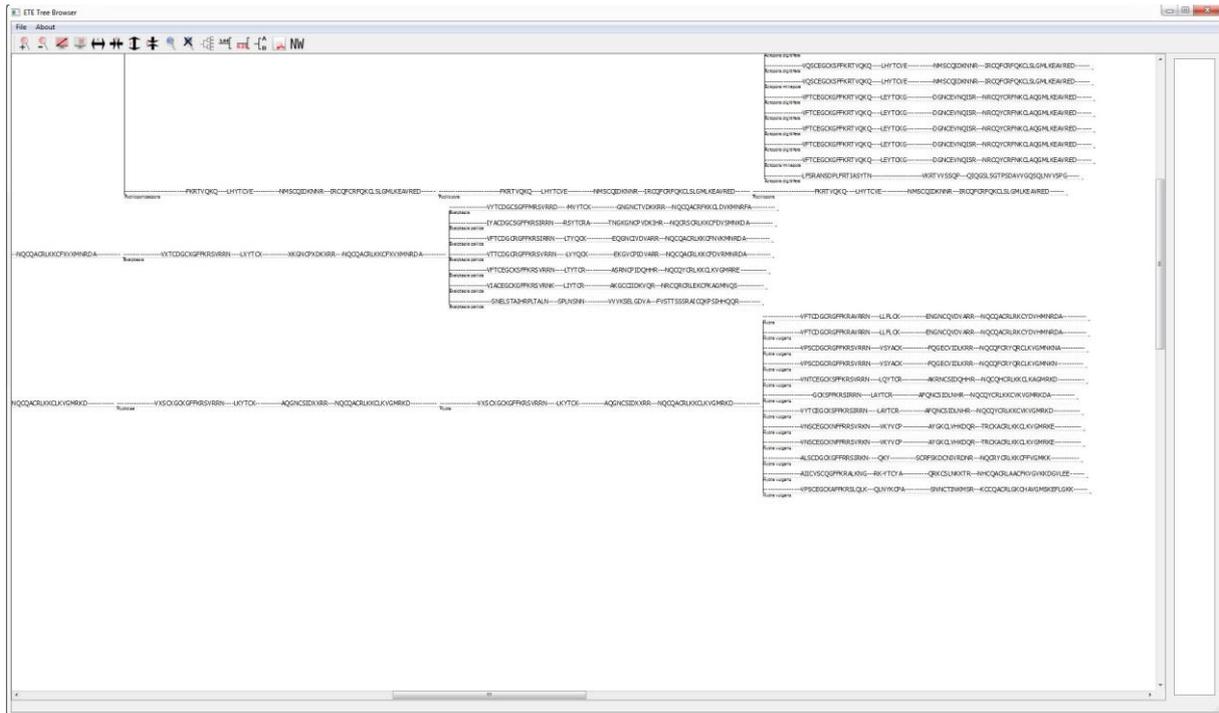


Figure 123 : Interface générale de BackPhylo, interface de phylogénie, zoom display du newick.

Il existe également la possibilité de fusionner et afficher avec des couleurs distinctes plusieurs arbres au format Newick calculés indépendamment. Il faut cependant que les séquences proviennent du même alignement pour ne pas générer d'erreur d'interprétation.



Figure 124 : Interface générale de BackPhylo, interface fusion de deux newick de 2 taxons proche, zoom display du newick.

La deuxième possibilité d'exploitation est un export au format CSV. Cependant ceci n'est pas un export basique identique à ce qui a été visualisé dans l'interface. Dans les CSV il n'y a pas la présence des séquences ancestrales théoriques par exemple. Lorsque l'on sélectionne cette option, une interface apparaît et propose de renseigner une ou plusieurs zones d'intérêts. Dans l'exemple figure

125 j'ai indiqué les intervalles de l'alignement qui correspondent aux différentes hélices du LBD des récepteurs nucléaires, ainsi que la région du  $\pi$ -turn.



**Figure 125 :** Interface générale de BackPhylo, interface CSV export.

Les résultats obtenus, figures 126 et 127, montrent un exemple de fichier de travail utilisé pour le chapitre 1 de cette thèse et pour lequel plusieurs ajouts ont été faits. La partie colorée en bas de la première figure correspond à l'arbre phylogénétique des organismes pour lesquels on a un ou plusieurs récepteurs nucléaires représentés. Les couleurs ne peuvent pas être générées dans un fichier CSV. Il s'agit donc d'un ajout dans le logiciel de visualisation que j'ai utilisé (Excel). La colonne suivante représente le nom de chaque récepteur nucléaire et le code couleur bleu ou vert indique respectivement les récepteurs nucléaires de classe I et de classe II (Brelivet et al., 2004). Les colonnes colorées suivantes en vert, orange, cyan, jaune indiquent des colonnes pour les résidus marqueurs de classe I et de classe II indiqués grâce à l'interface pour l'export CSV vu précédemment. Les cases

blanches indiquent le remplacement du marqueur de classe. La colonne rouge indique les récepteurs pour lesquels le motif du  $\pi$ -turn est présent. La série de colonnes blanches et larges qui suivent correspondent aux intervalles de séquences indiqués pour les hélices H1 à H12 du LBD. Pour la dernière partie, qui est une matrice avec des couleurs allant du vert au rouge, un tableau d'identité de séquence basé sur le LBD est calculé à partir de l'alignement de séquences. Ce tableau qui continue sur la seconde figure n'est pas généré directement par BackPhylo, mais par un autre outil en ligne ([imed.med.ucm.es/Tools/SIAS](http://imed.med.ucm.es/Tools/SIAS)). Si on utilise l'alignement pour générer ce tableau, l'ordre des séquences est le même et un simple copier-coller dans le fichier Excel permet alors d'obtenir une bonne correspondance des séquences.

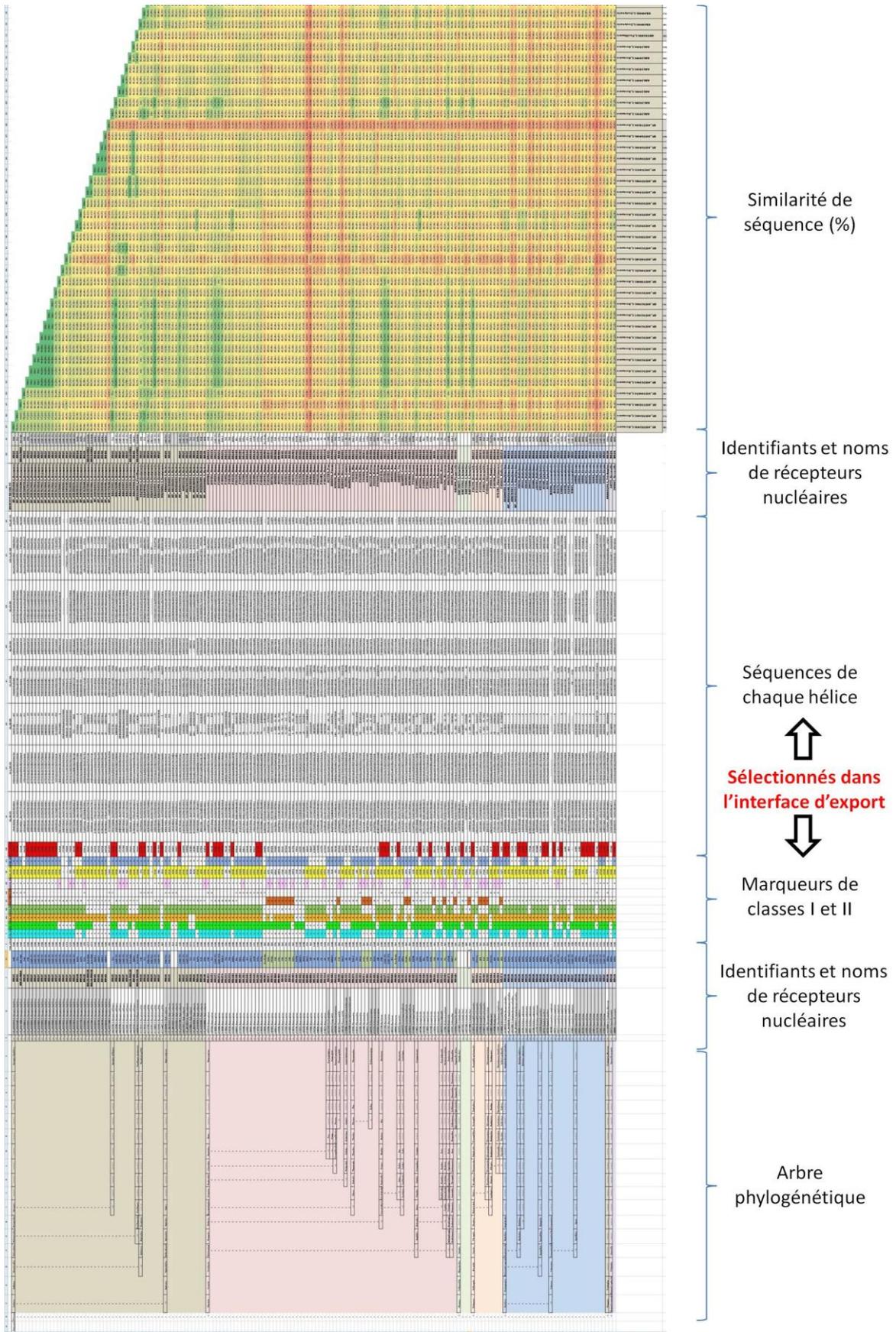


Figure 126 : Interface générale de BackPhylo, résultat CSV export (première moitié).

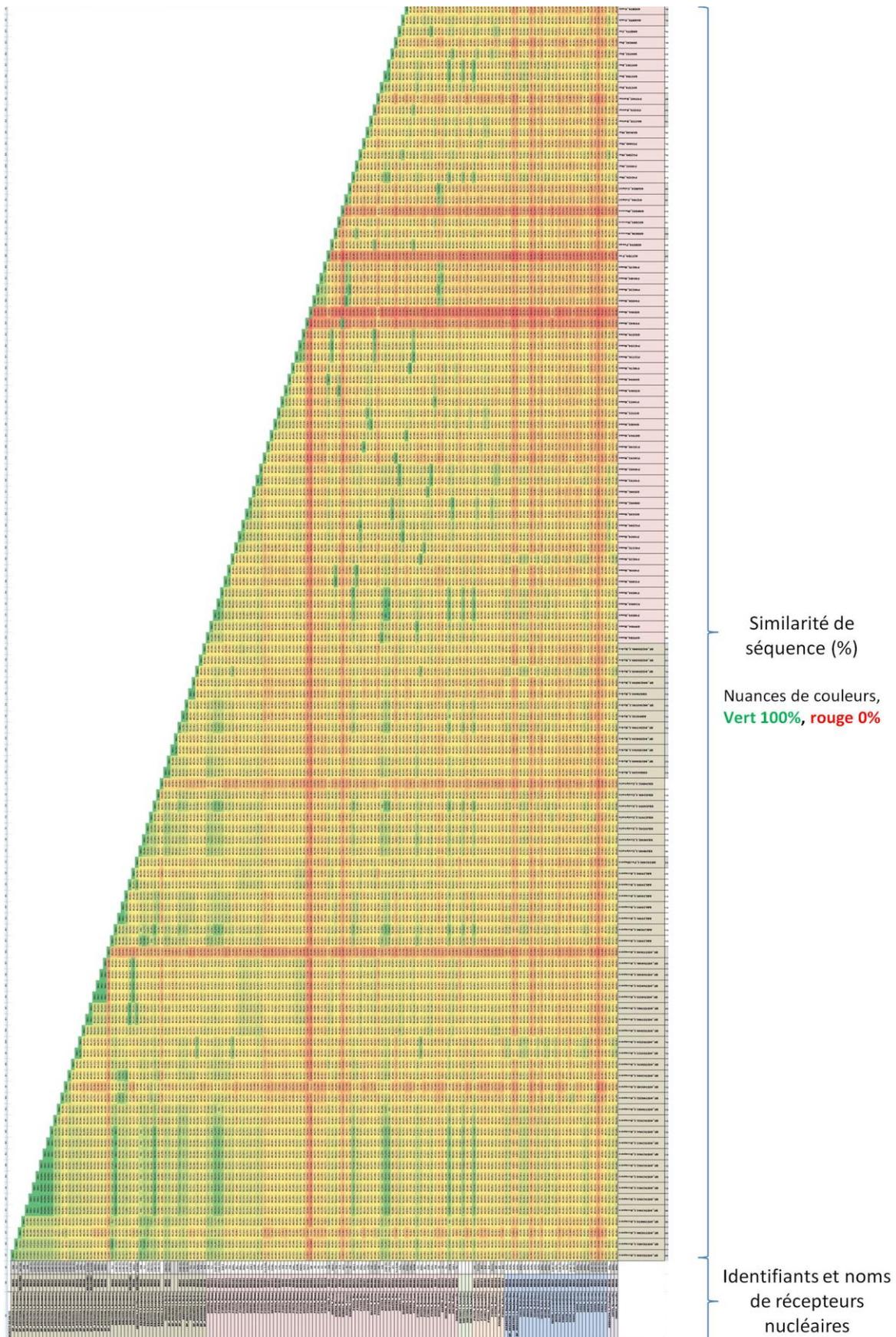


Figure 127 : Interface générale de BackPhylo, résultat CSV export (seconde moitié).

### 3.3 Discussion et perspectives

Le logiciel a été développé en python 3 comme pour Grid Files Manager présenté précédemment et pour les mêmes raisons. La partie calcul de séquences et traitement a été écrite en C pour des raisons de performance et de gestions fines de la mémoire (l'algorithme retenu nécessitant de garder de grandes quantités d'informations tout au long du traitement, notamment pour les arbres de plusieurs milliers de séquences). Le nom BackPhylo a été choisi car il permet de revenir en arrière dans l'évolution en calculant les séquences ancestrales théoriques lorsque c'est possible. Ce logiciel m'a permis de générer plusieurs fichiers de travail très utiles dans le cadre de la partie évolution des récepteurs nucléaires développé au chapitre 1 de ma thèse.

Parmi les prochains développements, il est envisagé de permettre le calcul d'une séquence théorique qui se regroupe par tags de différentes protéines d'une même famille. Cela, permettrait de créer les séquences intermédiaires de chaque tag, pour les protéines sans tag (car indéterminée, au niveau basal sans appartenance claire affichée avec les séquences "modernes") Le but de cette démarche serait de pouvoir déterminer avec fiabilité, dans quel taxon se sont probablement passé des évènements de duplications, différenciation ou encore délétion de gènes et retracer ainsi l'ensemble du chemin évolutif d'une famille. Dans le chapitre 1, nous avons vu qu'elles sont les descendants les plus proches des récepteurs nucléaires ancestraux en se basant sur leurs descendants chez les Poriferas. Avec la fonctionnalité énoncée, il serait alors plus facile d'effectuer le même travail pour l'ensemble de la famille en se basant sur toutes les séquences actuellement disponibles, provenant de génomes séquencés toujours plus nombreux. La mise à jour de l'arbre phylogénétique lorsque de nouvelles séquences deviennent disponibles serait aussi facilité.

Le logiciel est déjà distribué sous forme de fichier exécutable unique pour Windows et Linux ne nécessitant aucune installation. Toutes les dépendances sont directement incluses dans le fichier.

Après avoir implémenté les dernières fonctionnalités encore en développement et après publication le moment venu, le projet sera disponible via cette URL : <https://cbi-dev.igbmc.fr/cbi/backphylo>

## **Discussion générale**

---

Par la mise en œuvre de deux approches complémentaires, ma thèse a permis d'une part d'aborder l'origine et l'évolution des récepteurs nucléaires avec un aspect structural, d'autre part d'apporter de nouvelles connaissances structurales sur ERR, récepteur nucléaire à l'origine du groupe des récepteurs stéroïdiens. Ce récepteur nucléaire a été choisi comme récepteur nucléaire modèle du groupe des récepteurs stéroïdiens pour plusieurs raisons. ERR $\alpha$  est apparenté aux récepteurs des œstrogènes (ER $\alpha$  et ER $\beta$ ) et lie des séquences ADN similaires à celles liées par les ERs sous forme d'homodimères. D'autre part, du point de vue biochimique, ERR est plus simple à purifier et à manipuler que ER. Il nous permet donc d'étudier l'architecture moléculaire et l'organisation topologique des récepteurs nucléaires stéroïdiens homodimériques, le but étant d'analyser le complexe entier avec plusieurs ADN, et en présence ou en absence de coactivateurs.

La première approche est une analyse globale de l'ensemble des récepteurs nucléaires à partir d'une analyse bioinformatique des récepteurs nucléaires et à partir des bases de données de séquences et de structures afin de positionner ERR $\alpha$  dans la famille des récepteurs nucléaires d'un point de vue évolutif.

Dans ce contexte j'ai commencé par comparer l'ensemble des structures et des séquences de récepteurs nucléaires connus afin de mieux appréhender la famille de protéines dans son intégralité et ainsi mieux comprendre les particularités de ERR. Au cours de cette étude, je suis remonté jusqu'aux origines des récepteurs nucléaires. Un détail structural a attiré mon attention : un tour d'hélice  $\pi$  ( $\pi$ -turn) présent dans l'hélice H7 de RXR et de HNF-4. C'est un tour d'hélice qui comporte 5 acides aminés au lieu de 4 pour une hélice  $\alpha$  classique. Cette structure secondaire n'est pas banale et est énergétiquement défavorable. Sa présence et sa conservation stricte suggèrent donc qu'elle est importante pour le fonctionnement de ces deux récepteurs, d'autant plus que ce sont les deux seuls récepteurs nucléaires ayant cette modification dans le repliement conservé des LBD. Ce  $\pi$ -turn est accompagné d'un motif RxxxE au sein de H7 qui permet son maintien structural. L'observation est d'autant plus surprenante que ces deux récepteurs sont des exceptions, chacun à leur manière. En effet HNF-4 a deux acides aminés marqueurs de classe absents, tandis que RXR, est le seul récepteur nucléaire de classe I, à former des hétérodimères.

Il est communément admis que HNF-4 est le premier récepteur nucléaire, résultat que j'ai également pu constater en réalisant un calcul d'identité de séquence entre les deux récepteurs nucléaires présents dans NR1 et NR2, les seuls récepteurs nucléaires présents dans le plus ancien des taxons à

la base des métazoaires, celui des éponges (Porifera). J'ai également constaté que le motif RxxxE nécessaire au maintien du  $\pi$ -turn est également présent dans les séquences de NR1, et est donc présent dès le taxon basal des métazoaires. Cependant, ce motif est absent chez NR2, mais présent chez ces descendants, il s'agit donc d'une délétion tardive. Une autre observation a permis de remettre en cause le fait que HNF-4 soit le premier récepteur nucléaire. En effet, bien que NR1 et NR2 soient plus proches de la séquence de HNF-4, il n'y a que 4% de différences supplémentaires avec la séquence de RXR. Pour départager les deux récepteurs nucléaires nous avons utilisé une autre particularité de HNF-4 : ce récepteur n'a pas tous les acides aminés marqueurs de classe I contrairement aux autres récepteurs de classe I. Nous avons alors constaté que NR1 a tous les marqueurs de classe I contrairement à HNF-4, il est donc peu probable qu'il s'agisse d'un HNF-4 ancestral, d'autant plus que le récepteur nucléaire NR2 des Poriferas a quant à lui les mêmes exceptions de marqueurs de classes que HNF-4. Il est donc probable que le premier récepteur nucléaire soit globalement plus proche de HNF-4. Par contre nous ne pouvons pas affirmer ni infirmer quels marqueurs de classes sont présents à l'origine. Nous ne pouvons donc pas discriminer entre les deux hypothèses suivantes. La première est que le premier récepteur nucléaire est HNF-4, et qu'après duplication, deux nouveaux marqueurs de classes sont apparus par mutation, créant ainsi la lignée du récepteur nucléaire RXR puis les autres membres de classe I. La seconde hypothèse est le cheminement inverse, c'est à dire que le premier récepteur nucléaire ait été RXR, puis suite à une duplication puis à la perte de deux marqueurs de classe I par mutation, il y a eu création de la lignée de HNF-4. Le problème qu'il reste alors à élucider est l'absence du motif du  $\pi$ -turn chez NR2 qui est une forme ancestrale de HNF-4. Nous avons pu déterminer qu'il s'agit d'évènements de mutations qui ont totalement modifié l'hélice H7 de ce récepteur après séparation avec les taxons suivants. En effet, l'équivalent de NR2 dans le taxon suivant, les cnidaires, a quant à lui les mêmes exceptions que HNF-4 et NR2, et possède le motif du  $\pi$ -turn comme HNF-4. Chez les Cnidaires, une autre duplication permet l'apparition d'un troisième récepteur nucléaire. Ayant tous les marqueurs de classe I, il s'agit d'une duplication du RXR ancestral et de premières observations laissent penser qu'il s'agit là d'un premier évènement ayant permis l'émergence du groupe des récepteurs nucléaires stéroïdiens avec comme premier membre, ERR, le récepteur nucléaire d'intérêt de cette thèse.

Mon travail a donc permis de préciser l'origine des récepteurs nucléaires et d'y placer RXR en plus de HNF-4, lui déjà présent. Il a également permis de proposer un scénario évolutif concernant les premiers évènements génétiques de diversification de la famille des récepteurs nucléaires. Ces duplications engendrent les groupes de récepteurs nucléaires des HNF-4, des RXR/COUP-TF/PNR dès le premier taxon où est apparue la famille des récepteurs nucléaires. Le groupe des récepteurs stéroïdiens suit dans le taxon suivant des Cnidaires.

Nous montrons que quelques autres récepteurs nucléaires proches de RXR ont le motif de séquence RxxxE, mais seuls HNF-4 et RXR présentent la conformation du  $\pi$ -turn à cet endroit. L'analyse structurale de ces récepteurs suggère que le  $\pi$ -turn est stabilisé par des interactions intramoléculaires conservées aux deux extrémités de l'hélice H7, maintenant une conformation intra-hélicoïdale qui serait sinon instable à l'emplacement de RxxxE.

Plus tard, au cours de l'évolution, RXR a été sélectionné comme partenaire d'hétérodimérisation, probablement en partie à cause de la présence du  $\pi$ -turn, mais aussi à cause de différences subtiles dans la composition en acides aminés par rapport à HNF-4. Ces différences favorisent les interactions complémentaires entre RXR et ses récepteurs nucléaires partenaires. L'utilisation du  $\pi$ -turn peut donc être considérée comme une exaptation (c'est-à-dire une adaptation sélective opportuniste, privilégiant des caractères qui sont utiles à une nouvelle fonction). Ceci a probablement conduit à la nouvelle fonction cruciale d'hétérodimérisation et à la régulation stricte et spécifique de l'expression des gènes. En effet, les deux récepteurs nucléaires sont présents dans le taxon des Poriferas (spongiaires) et dans ce contexte, en absence de récepteurs nucléaires de classe II, leur activité biologique est susceptible de dépendre uniquement de la forme homodimérique et/ou monomérique.

Pour mener à bien cette étude j'ai été amené à traiter une grande quantité de données. Ne trouvant pas de logiciels ou d'outils adaptés à ce que je voulais faire, avec l'aide d'un collègue informaticien, nous avons développé un logiciel "Backphylo" avec pour but de pouvoir corréler un grand nombre d'informations de séquences dans un contexte évolutif. En effet, la famille des récepteurs nucléaires, avec près de 70 membres répartis à travers l'ensemble des métazoaires, représente plusieurs dizaines de milliers de séquences disponibles réparties entre plusieurs centaines d'espèces animales. Devant une telle quantité d'informations, il est nécessaire de hiérarchiser et de trier l'ensemble de ces données pour pouvoir les comparer les unes aux autres le plus facilement possible. On peut se trouver devant plusieurs problèmes dans ce type d'étude. Le premier étant la qualité des données récupérées dans les grandes banques de données telles que le NCBI. Ce développement m'a donc permis de mener à bien cette étude plus efficacement. Malgré le fait que certaines fonctions soient encore absentes, ce logiciel pourrait être très utile non seulement pour la suite de cette étude mais également pour d'autres études du même type, quelque soit la famille de protéines étudiée.

Parmi les prochains développements, il est envisagé de permettre le calcul d'une séquence théorique des intermédiaires de duplications pour identifier les ancêtres de chaque récepteur nucléaire dans les taxons basaux. Le but de cette démarche serait de pouvoir déterminer avec fiabilité dans quel

taxon se sont probablement passés des évènements de duplications, différenciations ou encore délétions de gènes et retracer ainsi l'ensemble du chemin évolutif d'une famille. Dans le chapitre 1, nous avons vu quels sont les descendants les plus proches des récepteurs nucléaires ancestraux en se basant sur leurs descendants chez les Poriferas. Avec la fonctionnalité énoncée, il serait alors plus facile d'effectuer le même travail pour l'ensemble de la famille en se basant sur toutes les séquences actuellement disponibles, provenant de génomes séquencés toujours plus nombreux. La mise à jour de l'arbre phylogénétique lorsque de nouvelles séquences deviennent disponibles serait aussi facilité.

La détermination de la structure tridimensionnelle du récepteur nucléaire ERR $\alpha$  en complexe avec son ADN et un coactivateur PGC-1 $\alpha$  constitue la seconde partie de ma thèse à utiliser les approches intégratives de la biologie structurale, et plus particulièrement la cryo-microscopie électronique.. L'objectif est l'analyse des relations structure-fonction dans le but de mieux comprendre les mécanismes d'action moléculaire. Il n'y a à l'heure actuelle, aucune structure à haute résolution connue d'un récepteur nucléaire en complexe avec un grand fragment de coactivateur. Il n'y a pas non plus de complexe à haute résolution connu pour un récepteur nucléaire stéroïdien lié à son ADN cible et comprenant les domaines DBD, charnière et LBD.

La singularité du récepteur ERR se traduit également par son mode d'interaction avec l'ADN. En effet, la quasi-totalité des récepteurs nucléaires dimérisent et se lient à des éléments de réponses dimériques (2x6 nucléotides espacés par 1 ou plusieurs acides nucléiques). ERR est également homodimérique, en raison des fortes interactions entre les deux LBD, cependant le dimère se lie préférentiellement à un élément de réponse monomérique comprenant un demi-site étendu (ERRE). La raison moléculaire de ce mode de fixation n'est actuellement pas connue et constitue l'un des points d'intérêts de l'étude structurale que j'ai effectué. En effet, cette spécificité d'interaction devrait permettre de comprendre le décodage de l'information contenue dans les séquences d'ADN des éléments de réponse des promoteurs, à partir de l'analyse de l'interface avec l'ADN dans la présente structure et sa comparaison avec celles de complexes liés à des séquences consensus.

Outre la connaissance fondamentale qu'il peut apporter sur la famille des récepteurs stéroïdiens, son étude est aussi intéressante pour ses applications biomédicales puisqu'ERR est une importante cible pharmaceutique, notamment dans le cas de plusieurs cancers. En effet, bien que le groupe des ERR n'ait pas de ligand naturel connu, ce sont des récepteurs nucléaires orphelins, plusieurs ligands synthétiques permettant de moduler leur activité sont connus.

Le premier défi à relever était l'optimisation des échantillons : la préparation d'échantillon en cryo-microscopie électronique est un point limitant important pour débiter une étude structurale, surtout quand le but est d'atteindre une résolution atomique. Ceci est d'autant plus vrai dans le cas d'un complexe relativement petit (entre 100 et 160 kDa) difficile à visualiser par cryo-ME où s'ajoute en plus de la difficulté de préparation, des difficultés liées à la technologie elle-même. En effet comme il a été montré dans la partie résultats du chapitre 2, au début de ma thèse, et durant les deux premières années, il n'était techniquement pas possible de visualiser les particules sur le Titan Krios de l'Institut en raison de la taille trop petite du complexe, entraînant un faible contraste en utilisant un voltage de 300 kV. Par la suite, de nouveaux détecteurs plus performants, et surtout l'ajout d'un Volta phase plate a changé la donne. Bien que le traitement d'images du dernier jeu de données ne soit pas encore terminé actuellement, et ne m'a donc pas permis d'atteindre pour l'instant la haute résolution, le processus est en bonne voie.

Cette partie de ma thèse peut se diviser en deux temps liés aux équipements scientifiques alors à disposition.

Les deux premières années, ont principalement été consacrées à l'optimisation des échantillons pour une étude par cryo-ME. J'ai réalisé ces optimisations sur le microscope Polara de l'Institut, utilisé à un voltage de 100 kV, ce qui permet d'amplifier significativement le contraste et donc d'avoir une observation des échantillons. La taille relativement petite des échantillons pose plusieurs problèmes lors d'un début de projet. Outre le fait qu'ils sont moins visibles qu'un virus ou encore qu'un ribosome, ils peuvent surtout se confondre avec des contaminants ou des résidus présents dans le tampon, notamment à cause de l'utilisation de détergents. Il devient alors difficile de tirer des conclusions sur ce qu'on observe sans effectuer d'expériences contrôles. Pour ces raisons, il est préférable d'effectuer quelques observations de tampon seul, pour bien se rendre compte des artefacts potentiellement présents. Ce sont ces observations, faites assez tôt au cours du projet qui nous ont poussé à changer de tampon et à remplacer le tampon Tris utilisé par du tampon HEPES. Le tampon HEPES montrant beaucoup moins d'artefacts de tailles similaires à un récepteur nucléaire. Plusieurs ADN ont été utilisés, de différentes tailles et différentes séquences. Bien qu'il y ait sûrement des changements sur la topologie des complexes non observables pour le moment avec les résolutions obtenues au cours de cette thèse, il ne semble pas influencer le comportement de l'échantillon. Le type de grilles et la façon de les préparer ont des effets plus ou moins importants en fonction des complexes. Sur ce point-là, le changement est radicalement différent dans le cas des complexes comprenant un ou deux nucléosomes pour lesquels des grilles Quantifoil R1.2/1.3 aident

pour permettre une distribution homogène des particules par rapport aux grilles R2/2. Dans les cas des complexes  $ERR\alpha$ -ADN et  $ERR\alpha$ -ADN-PGC-1 $\alpha$ , la différence est peu visible.

Au fur et à mesure des échecs et des réussites, nous avons isolé des conditions favorables pour le complexe  $ERR\alpha$ -ADN-PGC-1 $\alpha$  que je pourrai appliquer à d'autres complexes précédemment testés afin d'améliorer les complexes, comme les  $ERR\alpha$ -ADN et les complexes avec les nucléosomes dans une optique d'étude structurale à haute résolution. Par exemple, l'utilisation du détergent tel que le DDM à faible concentration et la présence de nouveaux équipements pour la préparation des grilles, notamment pour le glow discharge permettent de faire des grilles plus reproductibles. De plus, l'utilisation de DDM pourrait améliorer la stabilité des complexes avec le nucléosome en conditions cryogéniques. Pour ces complexes une autre solution a été envisagée mais non réalisée pour le moment : l'utilisation d'autres conditions de réticulation, avec par exemple un mixte de glutaraldéhyde avec du formaldéhyde, ou l'utilisation d'autres agents de réticulation tel que le BS3 (bis(sulfosuccinimidyl)suberate) (Shi et al., 2017).

Pour mener ce travail d'optimisation qui est à l'origine de la production de plus de 100 conditions différentes, avec l'aide d'un collègue informaticien (Jonathan Michalon), nous avons développé un logiciel "Grid files manager", ayant pour but de faciliter la corrélation entre les conditions de préparations des échantillons, allant de la purification jusqu'à la congélation des échantillons avec le résultat obtenu à l'observation par cryo-microscopie électronique. Il est alors possible de comparer facilement et rapidement plusieurs expériences, en faisant varier un ou plusieurs paramètres et déterminer ainsi les conditions les plus avantageuses.

Le logiciel est disponible via cette URL : <https://cbi-dev.igbmc.fr/cbi/gridfiles>

Ce projet s'intègre dans une logique plus large au niveau de la plateforme d'imagerie électronique au CBI-IGBMC. Il sera une base pour l'intégration au pipeline de microscopie électronique développé en interne, qui débute dès l'acquisition, et pour lequel je suis également partie prenante. Cette intégration permet de suivre et d'avoir une base statistique sur l'ensemble des expériences dans leur intégralité de la purification à la structure. Dans ce cadre, un développement pour une utilisation multi-utilisateurs et multi-projets a débuté pour une intégration aux besoins de la plateforme et ainsi faciliter la gestion des utilisateurs externes dans le cadre des infrastructures FRISBI et Instruct-ERIC.

Une réflexion pour ajouter plusieurs items comme par exemple le choix du type de machine utilisée pour la cryogénéisation permettrait de rendre cet outil plus flexible et ainsi plus compatible pour une utilisation plus large avec de futurs équipements ou dans d'autres instituts de recherche qui ont des

équipements différents. De plus, pour faciliter le comparatif des données, nous réfléchissons également à une sortie standard afin de pouvoir comparer l'ensemble des paramètres ou seulement pour certaines étapes, plusieurs échantillons sous forme d'un tableau.

Ces études m'ont permis d'avoir un premier aperçu de la topologie générale des complexes avec ERR. Pour les complexes avec le nucléosome, il n'a pas été possible de déterminer de structure avec ERR, faute d'un jeu de données où le complexe entre le nucléosome et ERR est formé en proportion suffisante (complexe hétérogène avec dissociation partielle de l'ADN). Cependant, les micrographes permettent d'avoir une première idée de l'orientation du récepteur nucléaire par rapport au nucléosome, à savoir qu'il serait orthogonal par rapport au plan du nucléosome dans le cas du complexe avec un seul nucléosome. Pour les complexes de ERR $\alpha$  - ADN - PGC-1 $\alpha$  on note l'asymétrie du complexe et il semble que PGC-1 $\alpha$  soit fixé en un seul exemplaire sur les complexes, en parfait accord avec les données biophysiques et de SAXS produites dans l'équipe. Il y a pourtant deux sites de liaisons possibles pour PGC-1 $\alpha$ , un sur chaque LBD, ce qui suggère un mécanisme allostérique dans la liaison du coactivateur sur le dimère de ERR. L'autre point est que PGC-1 $\alpha$  se fixe toujours du même côté dans le cas de ce complexe, pour cet ADN précis (BE29-embedded ERRE). De plus, la position de PGC-1 $\alpha$  correspond bien à ce qui est observé dans les structures cristallographiques de l'homodimère de ERR LBD liant un peptide court de PGC-1 $\alpha$  d'une vingtaine d'acides aminés. Dans ce contexte deux peptides de PGC-1 $\alpha$  sont fixés au récepteur, un sur chaque LBD. Par contre pour un fragment plus important de PGC-1 un seul coactivateur est lié au ERR homodimérique (ref Takacs 2013). Cette observation intéressante permet de faire des premières hypothèses qui seront validées ou réfutées lorsque des complexes seront affinés à haute résolution. Si un seul PGC-1 est lié à l'homodimère, il est possible que la fixation du PGC-1 $\alpha$  soit toujours du même côté sur ce complexe à cause d'un effet allostérique lié à la fixation à l'ADN. Cette fixation entraîne un remaniement différent entre le ERR fixé en 5' et celui en 3' de l'ADN, sachant que selon l'élément de réponse, le DBD en 3' pourrait être moins stable (Mohideen-Abdul et al., 2017). Cet effet indirect pourrait entraîner une asymétrie expliquant pourquoi une des deux sous-unités est préférée dans le recrutement du coactivateur PGC-1 $\alpha$ .

Pour répondre à ces questions il reste encore à dépasser certains obstacles liés au traitement d'images, afin d'atteindre la haute résolution. Compte tenu de ces difficultés, j'ai testé de très nombreux logiciels et algorithmes de traitement d'images, chacun avec leurs avantages et leurs inconvénients. Pour le moment aucun de ces logiciels (Relion 2, cisTEM, IMAGIC, EMAN2, cryoSPARC) ne m'a permis de mener à bien ce projet difficile, considérant la petite taille de l'objet, de bout à bout, il faut les combiner pour avancer dans le traitement d'images. La principale difficulté alors

réside dans le manque de standard de format en traitement d'images de cryo-ME. Chaque logiciel utilise son propre format de données, quelques fonctions d'import et d'export existent mais ne sont pas pour le moment suffisantes, et il faut régulièrement écrire de petits scripts pour continuer le traitement d'un logiciel à l'autre. Les dernières nouveautés dans ce domaine montrent une vraie volonté de la part des développeurs de fournir des logiciels aussi performants et complets que ceux qu'on retrouve pour les études en cristallographie aux rayons-X.

Les nouvelles technologies qui sont apparues au cours de ma thèse permettent à présent de faire des acquisitions de micrographes et des reconstructions à haute résolution de petits objets à 300 kV sur un microscope Titan Krios grâce aux caméras à détection directe d'électrons et à la Volta phase plate qui augmente fortement le contraste. Avec ces équipements à la pointe de la technologie et les dernières avancées en traitement d'images, plusieurs exemples apparaissent dans la littérature (Khoshouei et al., 2016, 2017). Cependant la tâche reste encore difficile, surtout pour les petits objets ne présentant pas de symétrie tel que ERR. Au contraire la pseudo-symétrie du complexe ERR peut constituer un problème à surmonter lors du traitement d'images.

Mon travail de thèse a permis d'établir les meilleures conditions de préparations de grilles pour les complexes  $ERR\alpha$ /ADN,  $ERR\alpha$ /ADN/PGC-1 $\alpha$ ,  $ERR\alpha$ /NCP etc. Il m'a permis également de procéder à l'enregistrement de données à haute résolution avec de l'équipement de pointe. J'ai donc pu étudier la structure à moyenne résolution et la topologie de plusieurs complexes, établir les bases sine qua non pour que l'étude structurale du complexe puisse être réalisée à haute résolution dans l'avenir. Les implications, basées sur l'analyse bioinformatique et structurale, sont fondamentaux pour le domaine des récepteurs nucléaire et pourra avoir des retombées en biomédecine.

# Publications

---

## **Publications:**

- **B. Beinsteiner**, J. Michalon & B. P. Klaholz. IBISS, a versatile and interactive tool for integrated sequence and 3D structure analysis of large macromolecular complexes. *Bioinformatics*, 2015, 31, 3339-3344. doi: 10.1093/bioinformatics/btv347.
  
- **B. Beinsteiner**, D. Moras. Structural Analysis of Heterodimeric Nuclear Receptors. Chapter: Nuclear Receptors: From Structure to the Clinic, pp.119-133. 2015 - IJ McEwan, R Kumar Springer doi: 10.1007/978-3-319-18729-7
  
- J. G. Arnez, **B. Beinsteiner** & D. Moras. Aminoacyl-tRNA Synthetases. *Encyclopedia of Life Sciences*, 2015, 10.1002/9780470015902.a0000530.pub3
  
- M. I. Valencia-Sánchez, A. Rodríguez-Hernández, R. Ferreira, HA. Santamaría-Suárez, M. Arciniega, A. C. Dock-Bregeon, D. Moras, **B. Beinsteiner**, H. Mertens, D. Svergun, L. G. Brieba, M. Grøtli, A. Torres-Larios. Structural Insights into the Polyphyletic Origins of Glycyl tRNA Synthetases. *J Biol Chem* 291, 14430-14446., 2016. Doi: 10.1074/jbc.M116.730382
  
- I. Orlov, A. G. Myasnikov, L. Andronov, S. K. Natchiar, H. Khatter, **B. Beinsteiner**, J-F. Ménétret, I. Hazemann, K. Mohideen, K. Tazibt, R. Tabaroni, H. Kratzat, N. Djabeur, T. Bruxelles, F. Raivoniaina, L. di Pompeo, M. Torchy, I. Billas, A. Urzhumtsev & B. P. Klaholz. The integrative role of cryo electron microscopy in molecular and cellular structural biology. *Biol Cell.*, 2017, 109, 81-93. doi: 10.1111/boc.201600042; invited review.

## **Publication en préparation :**

- **B. Beinsteiner**, A. Mc Ewen, G, Markov, V.Laudet, I. Billas, D. Moras. A structural signature motif enlighten the origin of NHRs
  
- Claire Lecroisey, Amel Yamoun, Alastair G. McEwen, **Brice Beinsteiner**, Guillaume Holzer, Gabriel V. Markov, Jenifer C. Croce, Shinja Yoo, Vanessa R. Flores, David A. Weisblat, Hector Escrivá, Michael Schubert, Isabelle M.L. Billas, Dino Moras, Vincent Laudet. NR7 is a missing link for understanding the origin of nuclear receptor heterodimerization

# Bibliographie

---

## Bibliographie

- Adrian, M., Dubochet, J., Lepault, J., and McDowell, A.W. (1984). Cryo-electron microscopy of viruses. *Nature* *308*, 32–36.
- Albani, A.E., Bengtson, S., Canfield, D.E., Bekker, A., Macchiarelli, R., Mazurier, A., Hammarlund, E.U., Boulvais, P., Dupuy, J.-J., Fontaine, C., et al. (2010). Large colonial organisms with coordinated growth in oxygenated environments 2.1 Gyr ago. *Nature* *466*, 100–104.
- Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J. (1990). Basic local alignment search tool. *J. Mol. Biol.* *215*, 403–410.
- Apweiler, R., Bairoch, A., Wu, C.H., Barker, W.C., Boeckmann, B., Ferro, S., Gasteiger, E., Huang, H., Lopez, R., Magrane, M., et al. (2004). UniProt: the Universal Protein knowledgebase. *Nucleic Acids Res.* *32*, D115–D119.
- Araki, M., and Motojima, K. (2006). Identification of ERRalpha as a specific partner of PGC-1alpha for the activation of PDK4 gene expression in muscle. *FEBS J.* *273*, 1669–1680.
- Ariazi, E.A., Clark, G.M., and Mertz, J.E. (2002). Estrogen-related receptor alpha and estrogen-related receptor gamma associate with unfavorable and favorable biomarkers, respectively, in human breast cancer. *Cancer Res.* *62*, 6510–6518.
- Aumais, J.P., Lee, H.S., DeGannes, C., Horsford, J., and White, J.H. (1996). Function of Directly Repeated Half-sites as Response Elements for Steroid Hormone Receptors. *J. Biol. Chem.* *271*, 12568–12577.
- Bairoch, A., and Apweiler, R. (1996). The SWISS-PROT protein sequence data bank and its new supplement TREMBL. *Nucleic Acids Res.* *24*, 21–25.
- Bairoch, A., and Boeckmann, B. (1991). The SWISS-PROT protein sequence data bank. *Nucleic Acids Res.* *19*, 2247–2249.
- Beinsteiner, B., and Moras, D. (2015). Structural Analysis of Heterodimeric Nuclear Receptors. In *Nuclear Receptors: From Structure to the Clinic*, I.J. McEwan, and R. Kumar, eds. (Springer International Publishing), pp. 119–133.
- Beinsteiner, B., Michalon, J., and Klaholz, B.P. (2015). IBISS, a versatile and interactive tool for integrated sequence and 3D structure analysis of large macromolecular complexes. *Bioinformatics* *31*, 3339–3344.
- Bernstein, F.C., Koetzle, T.F., Williams, G.J.B., Meyer Jr., E.F., Brice, M.D., Rodgers, J.R., Kennard, O., Shimanouchi, T., and Tasumi, M. (1977). The protein data bank: A computer-based archival file for macromolecular structures. *J. Mol. Biol.* *112*, 535–542.
- Bertrand, S., Brunet, F.G., Escriva, H., Parmentier, G., Laudet, V., and Robinson-Rechavi, M. (2004). Evolutionary Genomics of Nuclear Receptors: From Twenty-Five Ancestral Genes to Derived Endocrine Systems. *Mol. Biol. Evol.* *21*, 1923–1937.

- Billas, I.M., Moulinier, L., Rochel, N., and Moras, D. (2001). Crystal structure of the ligand-binding domain of the ultraspiracle protein USP, the ortholog of retinoid X receptors in insects. *J. Biol. Chem.* *276*, 7465–7474.
- Bledsoe, R.K., Montana, V.G., Stanley, T.B., Delves, C.J., Apolito, C.J., McKee, D.D., Consler, T.G., Parks, D.J., Stewart, E.L., Willson, T.M., et al. (2002). Crystal Structure of the Glucocorticoid Receptor Ligand Binding Domain Reveals a Novel Mode of Receptor Dimerization and Coactivator Recognition. *Cell* *110*, 93–105.
- Blondel, A., Renaud, J.-P., Fischer, S., Moras, D., and Karplus, M. (1999). Retinoic acid receptor: a simulation analysis of retinoic acid binding and the resulting conformational changes<sup>11</sup>Edited by T. Richmond. *J. Mol. Biol.* *291*, 101–115.
- Bookout, A.L., Jeong, Y., Downes, M., Yu, R.T., Evans, R.M., and Mangelsdorf, D.J. (2006). Anatomical Profiling of Nuclear Receptor Expression Reveals a Hierarchical Transcriptional Network. *Cell* *126*, 789–799.
- Bourguet, W., Ruff, M., Chambon, P., Gronemeyer, H., and Moras, D. (1995). Crystal structure of the ligand-binding domain of the human nuclear receptor RXR- $\alpha$ . *Nature* *375*, 377–382.
- Bourguet, W., Vivat, V., Wurtz, J.-M., Chambon, P., Gronemeyer, H., and Moras, D. (2000). Crystal Structure of a Heterodimeric Complex of RAR and RXR Ligand-Binding Domains. *Mol. Cell* *5*, 289–298.
- Bray, D., and Duke, T. (2004). Conformational Spread: The Propagation of Allosteric States in Large Multiprotein Complexes. *Annu. Rev. Biophys. Biomol. Struct.* *33*, 53–73.
- Brelivet, Y., Kammerer, S., Rochel, N., Poch, O., and Moras, D. (2004). Signature of the oligomeric behaviour of nuclear receptors at the sequence and structural level. *EMBO Rep.* *5*, 423–429.
- Brenner, S., and Horne, R.W. (1959). A negative staining method for high resolution electron microscopy of viruses. *Biochim. Biophys. Acta* *34*, 103–110.
- Bridgham, J.T., Eick, G.N., Larroux, C., Deshpande, K., Harms, M.J., Gauthier, M.E.A., Ortlund, E.A., Degan, B.M., and Thornton, J.W. (2010). Protein Evolution by Molecular Tinkering: Diversification of the Nuclear Receptor Superfamily from a Ligand-Dependent Ancestor. *PLoS Biol* *8*, e1000497.
- Brilot, A.F., Chen, J.Z., Cheng, A., Pan, J., Harrison, S.C., Potter, C.S., Carragher, B., Henderson, R., and Grigorieff, N. (2012). Beam-induced motion of vitrified specimen on holey carbon film. *J. Struct. Biol.* *177*, 630–637.
- Bulyanko, Y.A., and O'Malley, B.W. (2011). Nuclear receptor coactivators: structural and functional biochemistry. *Biochemistry* *50*, 313–328.
- Cai, Q., Lin, T., Kamarajugadda, S., and Lu, J. (2013). Regulation of glycolysis and the Warburg effect by estrogen-related receptors. *Oncogene* *32*, 2079–2086.
- Cammer, S. (2007). SChISM2: creating interactive web page annotations of molecular structure models using Jmol. *Bioinformatics* *23*, 383–384.

## Bibliographie

Caputi, F.F., and Romualdi, S.C. and P. (2017). Epigenetic Approaches in Neuroblastoma Disease Pathogenesis. *Neuroblastoma - Curr. State Recent Updat.*

Cartoni, R., Léger, B., Hock, M.B., Praz, M., Crettenand, A., Pich, S., Ziltener, J.-L., Luthi, F., Dériaz, O., Zorzano, A., et al. (2005). Mitofusins 1/2 and ERR $\alpha$  expression are increased in human skeletal muscle after physical exercise. *J. Physiol.* *567*, 349–358.

Cavallini, A., Notarnicola, M., Giannini, R., Montemurro, S., Lorusso, D., Visconti, A., Minervini, F., and Caruso, M.G. (2005). Oestrogen receptor-related receptor alpha (ERR $\alpha$ ) and oestrogen receptors (ER $\alpha$  and ER $\beta$ ) exhibit different gene expression in human colorectal tumour progression. *Eur. J. Cancer* *41*, 1487–1494.

Chalbos, D., Westley, B., May, F., Alibert, C., and Rochefort, H. (1986). Cloning of cDNA sequences of a progestin-regulated mRNA from MCF7 human breast cancer cells. *Nucleic Acids Res.* *14*, 965–982.

Chandra, V., Huang, P., Hamuro, Y., Raghuram, S., Wang, Y., Burriss, T.P., and Rastinejad, F. (2008). Structure of the intact PPAR- $\gamma$ -RXR- $\alpha$  nuclear receptor complex on DNA. *Nature* *456*, 350–356.

Chandra, V., Huang, P., Potluri, N., Wu, D., Kim, Y., and Rastinejad, F. (2013). Multidomain integration in the structure of the HNF-4 $\alpha$  nuclear receptor complex. *Nature* *495*, 394–398.

Chandra, V., Wu, D., Li, S., Potluri, N., Kim, Y., and Rastinejad, F. (2017). The quaternary architecture of RAR $\beta$ -RXR $\alpha$  heterodimer facilitates domain-domain signal transmission. *Nat. Commun.* *8*.

Chen, J.D., and Evans, R.M. (1995). A transcriptional co-repressor that interacts with nuclear hormone receptors. *Nature* *377*, 454–457.

Chen, D., Riedl, T., Washbrook, E., Pace, P.E., Coombes, R.C., Egly, J.-M., and Ali, S. (2000). Activation of Estrogen Receptor  $\alpha$  by S118 Phosphorylation Involves a Ligand-Dependent Interaction with TFIID and Participation of CDK7. *Mol. Cell* *6*, 127–137.

Clamp, M., Cuff, J., Searle, S.M., and Barton, G.J. (2004). The Jalview Java alignment editor. *Bioinformatics* *20*, 426–427.

Dahlman-Wright, K., and McEwan, I.J. (1996). Structural Studies of Mutant Glucocorticoid Receptor Transactivation Domains Establish a Link between Transactivation Activity in Vivo and  $\alpha$ -Helix-Forming Potential in Vitro. *Biochemistry* *35*, 1323–1327.

Dandey, V.P., Wei, H., Zhang, Z., Tan, Y.Z., Acharya, P., Eng, E.T., Rice, W.J., Kahn, P.A., Potter, C.S., and Carragher, B. (2018). Spotiton: New features and applications. *J. Struct. Biol.* *202*, 161–169.

Danev, R., and Baumeister, W. (2016). Cryo-EM single particle analysis with the Volta phase plate. *ELife* *5*.

Danev, R., Buijsse, B., Khoshouei, M., Plitzko, J.M., and Baumeister, W. (2014). Volta potential phase plate for in-focus phase contrast transmission electron microscopy. *Proc. Natl. Acad. Sci.* *111*, 15635–15640.

- Deblois, G., Hall, J.A., Perry, M.-C., Laganière, J., Ghahremani, M., Park, M., Hallett, M., and Giguère, V. (2009). Genome-wide identification of direct target genes implicates estrogen-related receptor alpha as a determinant of breast cancer heterogeneity. *Cancer Res.* *69*, 6149–6157.
- DeLano, W. (2002). PyMOL: An Open-Source Molecular Graphics Tool.
- Detera-Wadleigh, S.D., and Fanning, T.G. (1994). Phylogeny of the Steroid Receptor Superfamily. *Mol. Phylogenet. Evol.* *3*, 192–205.
- Egea, P.F., Mitschler, A., Rochel, N., Ruff, M., Chambon, P., and Moras, D. (2000). Crystal structure of the human RXRalpha ligand-binding domain bound to its natural ligand: 9-cis retinoic acid. *EMBO J.* *19*, 2592–2601.
- Egea, P.F., Mitschler, A., and Moras, D. (2002). Molecular Recognition of Agonist Ligands by RXRs. *Mol. Endocrinol.* *16*, 987–997.
- Emmert, D.B., Stoehr, P.J., Stoesser, G., and Cameron, G.N. (1994). The European Bioinformatics Institute (EBI) databases. *Nucleic Acids Res.* *22*, 3445–3449.
- Escriva, H., Delaunay, F., and Laudet, V. Ligand binding and nuclear receptor evolution. *BioEssays* *22*, 717–727.
- Fischer, N., Konevega, A.L., Wintermeyer, W., Rodnina, M.V., and Stark, H. (2010). Ribosome dynamics and tRNA movement by time-resolved electron cryomicroscopy. *Nature* *466*, 329–333.
- Freedman, L.P., Luisi, B.F., Korszun, Z.R., Basavappa, R., Sigler, P.B., and Yamamoto, K.R. (1988). The function and structure of the metal coordination sites within the glucocorticoid receptor DNA binding domain. *Nature* *334*, 543–546.
- Fu, M., Wang, C., Reutens, A.T., Wang, J., Angeletti, R.H., Siconolfi-Baez, L., Ogryzko, V., Avantaggiati, M.-L., and Pestell, R.G. (2000). p300 and p300/cAMP-response Element-binding Protein-associated Factor Acetylate the Androgen Receptor at Sites Governing Hormone-dependent Transactivation. *J. Biol. Chem.* *275*, 20853–20860.
- Fujimoto, J., and Sato, E. (2009). Clinical implication of estrogen-related receptor (ERR) expression in uterine endometrial cancers. *J. Steroid Biochem. Mol. Biol.* *116*, 71–75.
- Fujimura, T., Takahashi, S., Urano, T., Kumagai, J., Ogushi, T., Horie-Inoue, K., Ouchi, Y., Kitamura, T., Muramatsu, M., and Inoue, S. (2007). Increased expression of estrogen-related receptor alpha (ERRalpha) is a negative prognostic predictor in human prostate cancer. *Int. J. Cancer* *120*, 2325–2330.
- Gaillard, E., Bruck, N., Brelivet, Y., Bour, G., Lalevée, S., Bauer, A., Poch, O., Moras, D., and Rochette-Egly, C. (2006). Phosphorylation by PKA potentiates retinoic acid receptor alpha activity by means of increasing interaction with and phosphorylation by cyclin H/cdk7. *Proc. Natl. Acad. Sci. U. S. A.* *103*, 9548–9553.

## Bibliographie

- Gampe, R.T., Montana, V.G., Lambert, M.H., Miller, A.B., Bledsoe, R.K., Milburn, M.V., Kliewer, S.A., Willson, T.M., and Xu, H.E. (2000). Asymmetry in the PPAR $\gamma$ /RXR $\alpha$  Crystal Structure Reveals the Molecular Basis of Heterodimerization among Nuclear Receptors. *Mol. Cell* 5, 545–555.
- Gangloff, M., Ruff, M., Eiler, S., Duclaud, S., Wurtz, J.M., and Moras, D. (2001). Crystal Structure of a Mutant hER $\alpha$  Ligand-binding Domain Reveals Key Structural Features for the Mechanism of Partial Agonism. *J. Biol. Chem.* 276, 15059–15065.
- Gao, H., Valle, M., Ehrenberg, M., and Frank, J. (2004). Dynamics of EF-G interaction with the ribosome explored by classification of a heterogeneous cryo-EM dataset. *J. Struct. Biol.* 147, 283–290.
- Gewering, T., Januliene, D., Ries, A.B., and Moeller, A. (2018). Know your detergents: A case study on detergent background in negative stain electron microscopy. *J. Struct. Biol.*
- Giguère, V. (1999). Orphan Nuclear Receptors: From Gene to Function. *Endocr. Rev.* 20, 689–725.
- Giguère, V. (2008). Transcriptional Control of Energy Homeostasis by the Estrogen-Related Receptors. *Endocr. Rev.* 29, 677–696.
- Giguère, V., Yang, N., Segui, P., and Evans, R.M. (1988). Identification of a new class of steroid hormone receptors. *Nature* 331, 91–94.
- Glass, C.K. (1994). Differential Recognition of Target Genes by Nuclear Receptor Monomers, Dimers, and Heterodimers. *Endocr. Rev.* 15, 391–407.
- Glass, C.K., and Rosenfeld, M.G. (2000). The coregulator exchange in transcriptional functions of nuclear receptors. *Genes Dev.* 14, 121–141.
- Grant, T., and Grigorieff, N. Measuring the optimal exposure for single particle cryo-EM using a 2.6 Å reconstruction of rotavirus VP6. *ELife* 4.
- Greschik, H., Wurtz, J.-M., Sanglier, S., Bourguet, W., van Dorsselaer, A., Moras, D., and Renaud, J.-P. (2002). Structural and Functional Evidence for Ligand-Independent Transcriptional Activation by the Estrogen-Related Receptor 3. *Mol. Cell* 9, 303–313.
- Greschik, H., Althage, M., Flaig, R., Sato, Y., Chavant, V., Peluso-Iltis, C., Choulier, L., Cronet, P., Rochel, N., Schüle, R., et al. (2008). Communication between the ERR $\alpha$  Homodimer Interface and the PGC-1 $\alpha$  Binding Surface via the Helix 8–9 Loop. *J. Biol. Chem.* 283, 20220–20230.
- Gronemeyer, H., and Laudet, V. (1995). Transcription factors 3: nuclear receptors. *Protein Profile* 2, 1173–1308.
- Heard, D.J., Norby, P.L., Holloway, J., and Vissing, H. (2000). Human ERR $\gamma$ , a Third Member of the Estrogen Receptor-Related Receptor (ERR) Subfamily of Orphan Nuclear Receptors: Tissue-Specific Isoforms Are Expressed during Development and in the Adult. *Mol. Endocrinol.* 14, 382–392.
- van Heel, M., and Keegstra, W. (1981). IMAGIC: A fast, flexible and friendly image analysis software system. *Ultramicroscopy* 7, 113–129.

- van Heel, M., Harauz, G., Orlova, E.V., Schmidt, R., and Schatz, M. (1996). A New Generation of the IMAGIC Image Processing System. *J. Struct. Biol.* *116*, 17–24.
- Helsen, C., Dubois, V., Verfaillie, A., Young, J., Trekels, M., Vancaenenbroeck, R., Maeyer, M.D., and Claessens, F. (2012). Evidence for DNA-Binding Domain–Ligand-Binding Domain Communications in the Androgen Receptor. *Mol. Cell. Biol.* *32*, 3033–3043.
- Henderson, R. (2013). Avoiding the pitfalls of single particle cryo-electron microscopy: Einstein from noise. *Proc. Natl. Acad. Sci.* *110*, 18037–18041.
- Herráez, A. (2006). Biomolecules in the computer: Jmol to the rescue. *Biochem. Mol. Biol. Educ. Bimon. Publ. Int. Union Biochem. Mol. Biol.* *34*, 255–261.
- Herzog, B., Cardenas, J., Hall, R.K., Villena, J.A., Budge, P.J., Giguère, V., Granner, D.K., and Kralli, A. (2006). Estrogen-related Receptor  $\alpha$  Is a Repressor of Phosphoenolpyruvate Carboxykinase Gene Transcription. *J. Biol. Chem.* *281*, 99–106.
- Hollenberg, S.M., Weinberger, C., Ong, E.S., Cerelli, G., Oro, A., Lebo, R., Thompson, E.B., Rosenfeld, M.G., and Evans, R.M. (1985). Primary structure and expression of a functional human glucocorticoid receptor cDNA. *Nature* *318*, 635–641.
- Hörlein, A.J., Näär, A.M., Heinzl, T., Torchia, J., Gloss, B., Kurokawa, R., Ryan, A., Kamei, Y., Söderström, M., Glass, C.K., et al. (1995). Ligand-independent repression by the thyroid hormone receptor mediated by a nuclear receptor co-repressor. *Nature* *377*, 397–404.
- Horwitz, K.B., Jackson, T.A., Bain, D.L., Richer, J.K., Takimoto, G.S., and Tung, L. (1996). Nuclear receptor coactivators and corepressors. *Mol. Endocrinol.* *10*, 1167–1177.
- Hosokawa, F., Tomita, T., Naruse, M., Honda, T., Hartel, P., and Haider, M. (2003). A spherical aberration-corrected 200 kV TEM. *Microscopy* *52*, 3–10.
- Hudson, W.H., Youn, C., and Ortlund, E.A. (2014). Crystal Structure of the Mineralocorticoid Receptor DNA Binding Domain in Complex with DNA. *PLOS ONE* *9*, e107000.
- Huerta-Cepas, J., Dopazo, J., and Gabaldón, T. (2010). ETE: a python Environment for Tree Exploration. *BMC Bioinformatics* *11*, 24.
- Huerta-Cepas, J., Serra, F., and Bork, P. (2016). ETE 3: Reconstruction, Analysis, and Visualization of Phylogenomic Data. *Mol. Biol. Evol.* *33*, 1635–1638.
- Iwema, T., Billas, I.M., Beck, Y., Bonneton, F., Nierengarten, H., Chaumot, A., Richards, G., Laudet, V., and Moras, D. (2007). Structural and functional characterization of a novel type of ligand-independent RXR-USP receptor. *EMBO J.* *26*, 3770–3782.
- Jakób, M., Kołodziejczyk, R., Orłowski, M., Krzywda, S., Kowalska, A., Dutko-Gwózdź, J., Gwózdź, T., Kochman, M., Jaskólski, M., and Ozyhar, A. (2007). Novel DNA-binding element within the C-terminal extension of the nuclear receptor DNA-binding domain. *Nucleic Acids Res.* *35*, 2705–2718.

## Bibliographie

Jeltsch, J.M., Krozowski, Z., Quirin-Stricker, C., Gronemeyer, H., Simpson, R.J., Garnier, J.M., Krust, A., Jacob, F., and Chambon, P. (1986). Cloning of the chicken progesterone receptor. *Proc. Natl. Acad. Sci. U. S. A.* *83*, 5424–5428.

Jensen, E.V. (1962). On the mechanism of estrogen action. *Perspect. Biol. Med.* *6*, 47–59.

Jiang, W., and Chiu, W. (2001). Web-based Simulation for Contrast Transfer Function and Envelope Functions. *Microsc. Microanal.* *7*, 329–334.

Jiang, W., Baker, M.L., Jakana, J., Weigele, P.R., King, J., and Chiu, W. (2008). Backbone structure of the infectious  $\epsilon$ 15 virus capsid revealed by electron cryomicroscopy. *Nature* *451*, 1130–1134.

Kallen, J., Schlaeppli, J.-M., Bitsch, F., Filipuzzi, I., Schilb, A., Riou, V., Graham, A., Strauss, A., Geiser, M., and Fournier, B. (2004). Evidence for Ligand-independent Transcriptional Activation of the Human Estrogen-related Receptor  $\alpha$  (ERR $\alpha$ ) CRYSTAL STRUCTURE OF ERR $\alpha$  LIGAND BINDING DOMAIN IN COMPLEX WITH PEROXISOME PROLIFERATOR-ACTIVATED RECEPTOR COACTIVATOR-1 $\alpha$ . *J. Biol. Chem.* *279*, 49330–49337.

Kallen, J., Lattmann, R., Beerli, R., Blechschmidt, A., Blommers, M.J.J., Geiser, M., Ottl, J., Schlaeppli, J.-M., Strauss, A., and Fournier, B. (2007). Crystal Structure of Human Estrogen-related Receptor  $\alpha$  in Complex with a Synthetic Inverse Agonist Reveals Its Novel Molecular Mechanism. *J. Biol. Chem.* *282*, 23231–23239.

Khoshouei, M., Radjainia, M., Phillips, A.J., Gerrard, J.A., Mitra, A.K., Plitzko, J.M., Baumeister, W., and Danev, R. (2016). Volta phase plate cryo-EM of the small protein complex Prx3. *Nat. Commun.* *7*, 10534.

Khoshouei, M., Radjainia, M., Baumeister, W., and Danev, R. (2017). Cryo-EM structure of haemoglobin at 3.2 Å determined with the Volta phase plate. *Nat. Commun.* *8*, 16099.

Klaholz, B.P. (2015). Structure Sorting of Multiple Macromolecular States in Heterogeneous Cryo-EM Samples by 3D Multivariate Statistical Analysis. *Open J. Stat.* *05*, 820.

Klaholz, B.P., Myasnikov, A.G., and Heel, M. van (2004). Visualization of release factor 3 on the ribosome during termination of protein synthesis. *Nature* *427*, 862–865.

Kliwer, S.A., Lehmann, J.M., and Willson, T.M. (1999). Orphan Nuclear Receptors: Shifting Endocrinology into Reverse. *Science* *284*, 757–760.

Koshland, D.E., Némethy, G., and Filmer, D. (1966). Comparison of Experimental Binding Data and Theoretical Models in Proteins Containing Subunits\*. *Biochemistry* *5*, 365–385.

Krust, A., Green, S., Argos, P., Kumar, V., Walter, P., Bornert, J.M., and Chambon, P. (1986). The chicken oestrogen receptor sequence: homology with v-erbA and the human oestrogen and glucocorticoid receptors. *EMBO J.* *5*, 891–897.

Kühlbrandt, W. (2014). Cryo-EM enters a new era. *ELife* *3*, e03678.

- Laudet, V. (1997). Evolution of the nuclear receptor superfamily: early diversification from an ancestral orphan receptor. *J. Mol. Endocrinol.* *19*, 207–226.
- Lawson, C.L., Baker, M.L., Best, C., Bi, C., Dougherty, M., Feng, P., van Ginkel, G., Devkota, B., Lagerstedt, I., Ludtke, S.J., et al. (2011). EMDDataBank.org: unified data resource for CryoEM. *Nucleic Acids Res.* *39*, D456–D464.
- Leeuwen, I.J.D.V., Pereira, D. da C., Flach, K.D., Piersma, S.R., Haase, C., Bier, D., Yalcin, Z., Michalides, R., Feenstra, K.A., Jiménez, C.R., et al. (2013). Interaction of 14-3-3 proteins with the Estrogen Receptor Alpha F domain provides a drug target interface. *Proc. Natl. Acad. Sci.* *110*, 8894–8899.
- Leschziner, A. (2010). Chapter Nine - The Orthogonal Tilt Reconstruction Method. In *Methods in Enzymology*, G.J. Jensen, ed. (Academic Press), pp. 237–262.
- Leschziner, A.E., and Nogales, E. (2006). The orthogonal tilt reconstruction method: An approach to generating single-class volumes with no missing cone for ab initio reconstruction of asymmetric particles. *J. Struct. Biol.* *153*, 284–299.
- Li, J., Wang, J., Wang, J., Nawaz, Z., Liu, J.M., Qin, J., and Wong, J. (2000). Both corepressor proteins SMRT and N-CoR exist in large protein complexes containing HDAC3. *EMBO J.* *19*, 4342–4350.
- Li, X., Mooney, P., Zheng, S., Booth, C., Braunfeld, M.B., Gubbens, S., Agard, D.A., and Cheng, Y. (2013). Electron counting and beam-induced motion correction enable near atomic resolution single particle cryoEM. *Nat. Methods* *10*, 584–590.
- Li, Y., Lambert, M.H., and Xu, H.E. (2003). Activation of Nuclear Receptors: A Perspective from Structural Genomics. *Structure* *11*, 741–746.
- Liao, H.Y., Hashem, Y., and Frank, J. (2015). Efficient Estimation of Three-Dimensional Covariance and its Application in the Analysis of Heterogeneous Samples in Cryo-Electron Microscopy. *Structure* *23*, 1129–1137.
- Lou, X., Toresson, G., Benod, C., Suh, J.H., Philips, K.J., Webb, P., and Gustafsson, J.-A. (2014). Structure of the retinoid X receptor  $\alpha$ -liver X receptor  $\beta$  (RXR $\alpha$ -LXR $\beta$ ) heterodimer on DNA. *Nat. Struct. Mol. Biol.* *21*, 277–281.
- Lu, D., Kiriya, Y., Lee, K.Y., and Giguère, V. (2001). Transcriptional Regulation of the Estrogen-inducible pS2 Breast Cancer Marker Gene by the ERR Family of Orphan Nuclear Receptors. *Cancer Res.* *61*, 6755–6761.
- Luger, K., Mäder, A.W., Richmond, R.K., Sargent, D.F., and Richmond, T.J. (1997). Crystal structure of the nucleosome core particle at 2.8 Å resolution. *Nature* *389*, 251–260.
- Luisi, B.F., Xu, W.X., Otwinowski, Z., Freedman, L.P., Yamamoto, K.R., and Sigler, P.B. (1991). Crystallographic analysis of the interaction of the glucocorticoid receptor with DNA. *Nature* *352*, 497–505.

## Bibliographie

Maletta, M., Orlov, I., Roblin, P., Beck, Y., Moras, D., Billas, I.M.L., and Klaholz, B.P. (2014). The palindromic DNA-bound USP/EcR nuclear receptor adopts an asymmetric organization with allosteric domain positioning. *Nat. Commun.* 5.

Mastronarde, D.N. (2005). Automated electron microscope tomography using robust prediction of specimen movements. *J. Struct. Biol.* 152, 36–51.

May, F.E. (2014). Novel drugs that target the estrogen-related receptor alpha: their therapeutic potential in breast cancer. *Cancer Manag. Res.* 6, 225–252.

McMullan, G., Faruqi, A.R., Clare, D., and Henderson, R. (2014). Comparison of optimal performance at 300keV of three direct electron detectors for use in low dose electron microscopy. *Ultramicroscopy* 147, 156–163.

Meijsing, S.H., Pufall, M.A., So, A.Y., Bates, D.L., Chen, L., and Yamamoto, K.R. (2009). DNA binding site sequence directs glucocorticoid receptor structure and activity. *Science* 324, 407–410.

Meinke, G., and Sigler, P.B. (1999). DNA-binding mechanism of the monomeric orphan nuclear receptor NGFI-B. *Nat. Struct. Mol. Biol.* 6, 471–477.

Merk, A., Bartesaghi, A., Banerjee, S., Falconieri, V., Rao, P., Davis, M.I., Pragani, R., Boxer, M.B., Earl, L.A., Milne, J.L.S., et al. (2016). Breaking Cryo-EM Resolution Barriers to Facilitate Drug Discovery. *Cell* 165, 1698–1707.

Meyer, E.F. (1997). The first years of the Protein Data Bank. *Protein Sci. Publ. Protein Soc.* 6, 1591–1597.

Mizwicki, M.T., Keidel, D., Bula, C.M., Bishop, J.E., Zanello, L.P., Wurtz, J.-M., Moras, D., and Norman, A.W. (2004). Identification of an alternative ligand-binding pocket in the nuclear vitamin D receptor and its functional importance in  $1\alpha,25(\text{OH})_2$ -vitamin D<sub>3</sub> signaling. *Proc. Natl. Acad. Sci.* 101, 12876–12881.

Mohideen-Abdul, K., Tazibt, K., Bourguet, M., Hazemann, I., Lebars, I., Takacs, M., Cianféroni, S., Klaholz, B.P., Moras, D., and Billas, I.M.L. (2017). Importance of the Sequence-Directed DNA Shape for Specific Binding Site Recognition by the Estrogen-Related Receptor. *Front. Endocrinol.* 8.

Monod, J., Wyman, J., and Changeux, J.-P. (1965). On the nature of allosteric transitions: A plausible model. *J. Mol. Biol.* 12, 88–118.

Monsalve, M., Wu, Z., Adelmant, G., Puigserver, P., Fan, M., and Spiegelman, B.M. (2000). Direct Coupling of Transcription and mRNA Processing through the Thermogenic Coactivator PGC-1. *Mol. Cell* 6, 307–316.

Moras, D., and Gronemeyer, H. (1998). The nuclear receptor ligand-binding domain: structure and function. *Curr. Opin. Cell Biol.* 10, 384–391.

Nagy, L., Kao, H.Y., Love, J.D., Li, C., Banayo, E., Gooch, J.T., Krishna, V., Chatterjee, K., Evans, R.M., and Schwabe, J.W. (1999). Mechanism of corepressor binding and release from nuclear hormone receptors. *Genes Dev.* 13, 3209–3216.

- O'Malley, B.W. (1971). Mechanisms of Action of Steroid Hormones. *N. Engl. J. Med.* *284*, 370–377.
- Orlov, I., Rochel, N., Moras, D., and Klaholz, B.P. (2012). Structure of the full human RXR/VDR nuclear receptor heterodimer complex with its DR3 target DNA. *EMBO J.* *31*, 291–300.
- Orlov, I., Myasnikov, A.G., Andronov, L., Natchiar, S.K., Khatter, H., Beinsteiner, B., Ménétret, J.-F., Hazemann, I., Mohideen, K., Tazibt, K., et al. (2016). The integrative role of cryo electron microscopy in molecular and cellular structural biology. *Biol. Cell.*
- Orlova, E.V., and Saibil, H.R. (2010). Chapter Twelve - Methods for Three-Dimensional Reconstruction of Heterogeneous Assemblies. In *Methods in Enzymology*, G.J. Jensen, ed. (Academic Press), pp. 321–341.
- Orlova, E.V., and Saibil, H.R. (2011). Structural Analysis of Macromolecular Assemblies by Electron Microscopy. *Chem. Rev.* *111*, 7710–7748.
- Oro, A.E., McKeown, M., and Evans, R.M. (1990). Relationship between the product of the *Drosophila* ultraspiracle locus and the vertebrate retinoid X receptor. *Nature* *347*, 298–301.
- Ostell, J.M., and Kans, J.A. (1998). The NCBI data model. *Methods Biochem. Anal.* *39*, 121–144.
- Osz, J., McEwen, A.G., Poussin-Courmontagne, P., Moutier, E., Birck, C., Davidson, I., Moras, D., and Rochel, N. (2015). Structural basis of natural promoter recognition by the retinoid x nuclear receptor. *Sci. Rep.* *5*, 8216.
- Patch, R.J., Searle, L.L., Kim, A.J., De, D., Zhu, X., Askari, H.B., O'Neill, J.C., Abad, M.C., Rentzeperis, D., Liu, J., et al. (2011). Identification of Diaryl Ether-Based Ligands for Estrogen-Related Receptor  $\alpha$  as Potential Antidiabetic Agents. *J. Med. Chem.* *54*, 788–808.
- Peitsch, M.C., Wells, T.N.C., Stampf, D.R., and Sussman, J.L. (1995). The Swiss-3DImage collection and PDB-Browser on the World-Wide Web. *Trends Biochem. Sci.* *20*, 82–84.
- Penczek, P.A., Frank, J., and Spahn, C.M.T. (2006). A method of focused classification, based on the bootstrap 3D variance analysis, and its application to EF-G-dependent translocation. *J. Struct. Biol.* *154*, 184–194.
- Pérez-Schindler, J., Summermatter, S., Salatino, S., Zorzato, F., Beer, M., Balwierz, P.J., Nimwegen, E. van, Feige, J.N., Auwerx, J., and Handschin, C. (2012). The Corepressor NCoR1 Antagonizes PGC-1 $\alpha$  and Estrogen-Related Receptor  $\alpha$  in the Regulation of Skeletal Muscle Function and Oxidative Metabolism. *Mol. Cell. Biol.* *32*, 4913–4924.
- Petkovich, M., Brand, N.J., Krust, A., and Chambon, P. (1987). A human retinoic acid receptor which belongs to the family of nuclear receptors. *Nature* *330*, 444–450.
- Petoukhov, M.V., Billas, I.M.L., Takacs, M., Graewert, M.A., Moras, D., and Svergun, D.I. (2013). Reconstruction of quaternary structure from X-ray scattering by equilibrium mixtures of biological macromolecules. *Biochemistry* *52*, 6844–6855.

## Bibliographie

- Pettersen, E.F., Goddard, T.D., Huang, C.C., Couch, G.S., Greenblatt, D.M., Meng, E.C., and Ferrin, T.E. (2004). UCSF Chimera—A visualization system for exploratory research and analysis. *J. Comput. Chem.* *25*, 1605–1612.
- Puigserver, P., Wu, Z., Park, C.W., Graves, R., Wright, M., and Spiegelman, B.M. (1998). A Cold-Inducible Coactivator of Nuclear Receptors Linked to Adaptive Thermogenesis. *Cell* *92*, 829–839.
- Puigserver, P., Adelmant, G., Wu, Z., Fan, M., Xu, J., O'Malley, B., and Spiegelman, B.M. (1999). Activation of PPAR $\gamma$  Coactivator-1 Through Transcription Factor Docking. *Science* *286*, 1368–1371.
- Puigserver, P., Rhee, J., Lin, J., Wu, Z., Yoon, J.C., Zhang, C.-Y., Krauss, S., Mootha, V.K., Lowell, B.B., and Spiegelman, B.M. (2001). Cytokine Stimulation of Energy Expenditure through p38 MAP Kinase Activation of PPAR $\gamma$  Coactivator-1. *Mol. Cell* *8*, 971–982.
- Radermacher, M., Wagenknecht, T., Verschoor, A., and Frank, J. (1987). Three-dimensional reconstruction from a single-exposure, random conical tilt series applied to the 50S ribosomal subunit of *Escherichia coli*. *J. Microsc.* *146*, 113–136.
- Rastinejad, F., Wagner, T., Zhao, Q., and Khorasanizadeh, S. (2000). Structure of the RXR-RAR DNA-binding complex on the retinoic acid response element DR1. *EMBO J.* *19*, 1045–1054.
- Raven, P.H., Evert, R.F., and Eichhorn, S.E. (2005). *Biology of Plants* (W.H. Freeman).
- Reece, J.B. (2011). *Campbell Biology* (Benjamin Cummings / Pearson).
- Renaud, J.-P., Rochel, N., Ruff, M., Vivat, V., Chambon, P., Gronemeyer, H., and Moras, D. (1995). Crystal structure of the RAR- $\gamma$  ligand-binding domain bound to all-trans retinoic acid. *Nature* *378*, 681–689.
- Robin, G., Baffet, H., Catteau-Jonard, S., and Dewailly, D. (2012). Déficits en 21-hydroxylase et fertilité féminine. *Médecine Reprod.* *14*, 226–235.
- Rochel, N., Ciesielski, F., Godet, J., Moman, E., Roessle, M., Peluso-Iltis, C., Moulin, M., Haertlein, M., Callow, P., Mély, Y., et al. (2011). Common architecture of nuclear receptor heterodimers on DNA direct repeat elements with different spacings. *Nat. Struct. Mol. Biol.* *18*, 564–570.
- Roemer, S.C., Donham, D.C., Sherman, L., Pon, V.H., Edwards, D.P., and Churchill, M.E.A. (2006). Structure of the Progesterone Receptor-Deoxyribonucleic Acid Complex: Novel Interactions Required for Binding to Half-Site Response Elements. *Mol. Endocrinol.* *20*, 3042–3052.
- Russo, C.J., and Passmore, L.A. (2014). Ultrastable gold substrates for electron cryomicroscopy. *Science* *346*, 1377–1380.
- Saibil, H.R. (2000). Macromolecular structure determination by cryo-electron microscopy. *Acta Crystallogr. D Biol. Crystallogr.* *56*, 1215–1222.
- Sap, J., Muñoz, A., Damm, K., Goldberg, Y., Ghysdael, J., Leutz, A., Beug, H., and Vennström, B. (1986). The c-erb-A protein is a high-affinity receptor for thyroid hormone. *Nature* *324*, 635–640.

- Saunders, M., and Shaw, J.A. (2014). Biological Applications of Energy-Filtered TEM. In *Electron Microscopy: Methods and Protocols*, J. Kuo, ed. (Totowa, NJ: Humana Press), pp. 689–706.
- Schreiber, S.N., Knutti, D., Brogli, K., Uhlmann, T., and Kralli, A. (2003). The transcriptional coactivator PGC-1 regulates the expression and activity of the orphan nuclear receptor estrogen-related receptor alpha (ERRalpha). *J. Biol. Chem.* *278*, 9013–9018.
- Schwabe, J.W., Chapman, L., Finch, J.T., and Rhodes, D. (1993). The crystal structure of the estrogen receptor DNA-binding domain bound to DNA: how receptors discriminate between their response elements. *Cell* *75*, 567–578.
- Shi, J.-M., Pei, J., Liu, E.-Q., and Zhang, L. (2017). Bis(sulfosuccinimidyl) suberate (BS3) crosslinking analysis of the behavior of amyloid- $\beta$  peptide in solution and in phospholipid membranes. *PLOS ONE* *12*, e0173871.
- Simonetti, A., Marzi, S., Myasnikov, A.G., Fabbretti, A., Yusupov, M., Gualerzi, C.O., and Klaholz, B.P. (2008). Structure of the 30S translation initiation complex. *Nature* *455*, 416–420.
- Skafar, D.F., and Zhao, C. (2008). The multifunctional estrogen receptor-alpha F domain. *Endocrine* *33*, 1–8.
- Sladek, R., Bader, J.A., and Giguère, V. (1997). The orphan nuclear receptor estrogen-related receptor alpha is a transcriptional regulator of the human medium-chain acyl coenzyme A dehydrogenase gene. *Mol. Cell. Biol.* *17*, 5400–5409.
- Snel, B., Lehmann, G., Bork, P., and Huynen, M.A. (2000). STRING: a web-server to retrieve and display the repeatedly occurring neighbourhood of a gene. *Nucleic Acids Res.* *28*, 3442–3444.
- Spurr, N.K., Solomon, E., Jansson, M., Sheer, D., Goodfellow, P.N., Bodmer, W.F., and Vennstrom, B. (1984). Chromosomal localisation of the human homologues to the oncogenes *erbA* and *B*. *EMBO J.* *3*, 159–163.
- Steven, A., and Belnap, D. (2005). *Electron Microscopy and Image Processing: An Essential Tool for Structural Analysis of Macromolecules*. *Curr. Protoc. Protein Sci.* *42*, 17.2.1-17.2.39.
- Subramaniam, S. (2013). Structure of trimeric HIV-1 envelope glycoproteins. *Proc. Natl. Acad. Sci.* *110*, E4172–E4174.
- Sun, P., Wei, L., Denkert, C., Lichtenegger, W., and Sehouli, J. (2006). The Orphan Nuclear Receptors, Estrogen Receptor-related Receptors: their Role as New Biomarkers in Gynecological Cancer. *Anticancer Res.* *26*, 1699–1706.
- Suzuki, T., Miki, Y., Moriya, T., Shimada, N., Ishida, T., Hirakawa, H., Ohuchi, N., and Sasano, H. (2004). Estrogen-Related Receptor  $\alpha$  in Human Breast Carcinoma as a Potent Prognostic Factor. *Cancer Res.* *64*, 4670–4676.
- Svensson, S., Östberg, T., Jacobsson, M., Norström, C., Stefansson, K., Hallén, D., Johansson, I.C., Zachrisson, K., Ogg, D., and Jendeborg, L. (2003). Crystal structure of the heterodimeric complex of LXR $\alpha$  and RXR $\beta$  ligand-binding domains in a fully agonistic conformation. *EMBO J.* *22*, 4625–4633.

## Bibliographie

Takacs, M., Petoukhov, M.V., Atkinson, R.A., Roblin, P., Ogi, F.-X., Demeler, B., Potier, N., Chebaro, Y., Dejaegere, A., Svergun, D.I., et al. (2013). The asymmetric binding of PGC-1 $\alpha$  to the ERR $\alpha$  and ERR $\gamma$  nuclear receptor homodimers involves a similar recognition mechanism. *PLoS One* 8, e67810.

Tan, Y.Z., Cheng, A., Potter, C.S., and Carragher, B. (2016). Automated data collection in single particle electron microscopy. *Microscopy* 65, 43–56.

Tanenbaum, D.M., Wang, Y., Williams, S.P., and Sigler, P.B. (1998). Crystallographic comparison of the estrogen and progesterone receptor's ligand binding domains. *Proc. Natl. Acad. Sci. U. S. A.* 95, 5998.

Tateno, Y., and Gojobori, T. (1997). DNA Data Bank of Japan in the age of information biology. *Nucleic Acids Res.* 25, 14–17.

Thompson, J.D., Higgins, D.G., and Gibson, T.J. (1994). CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.* 22, 4673–4680.

Thornton, J.W., Need, E., and Crews, D. (2003). Resurrecting the Ancestral Steroid Receptor: Ancient Origin of Estrogen Signaling. *Science* 301, 1714–1717.

Tremblay, A., Tremblay, G.B., Labrie, F., and Giguère, V. (1999). Ligand-Independent Recruitment of SRC-1 to Estrogen Receptor  $\beta$  through Phosphorylation of Activation Function AF-1. *Mol. Cell* 3, 513–519.

Tremblay, A.M., Wilson, B.J., Yang, X.-J., and Giguère, V. (2008). Phosphorylation-Dependent Sumoylation Regulates Estrogen-Related Receptor- $\alpha$  and - $\gamma$  Transcriptional Activity through a Synergy Control Motif. *Mol. Endocrinol.* 22, 570–584.

Van Heel, M. (2013). Finding trimeric HIV-1 envelope glycoproteins in random noise. *Proc. Natl. Acad. Sci.* 110, E4175–E4177.

Vanacker, J.M., Delmarre, C., Guo, X., and Laudet, V. (1998). Activation of the osteopontin promoter by the orphan nuclear receptor estrogen receptor related alpha. *Cell Growth Differ.* 9, 1007.

Vanacker, J.-M., Pettersson, K., Gustafsson, J.-Å., and Laudet, V. (1999). Transcriptional targets shared by estrogen receptor-related receptors (ERRs) and estrogen receptor (ER)  $\alpha$ , but not by ER $\beta$ . *EMBO J.* 18, 4270–4279.

Vesely, P.W., Staber, P.B., Hoefler, G., and Kenner, L. (2009). Translational regulation mechanisms of AP-1 proteins. *Mutat. Res. Mutat. Res.* 682, 7–12.

Wärnmark, A., Treuter, E., Wright, A.P.H., and Gustafsson, J.-Å. (2003). Activation Functions 1 and 2 of Nuclear Receptors: Molecular Strategies for Transcriptional Activation. *Mol. Endocrinol.* 17, 1901–1909.

Waterhouse, A.M., Procter, J.B., Martin, D.M.A., Clamp, M., and Barton, G.J. (2009). Jalview Version 2--a multiple sequence alignment editor and analysis workbench. *Bioinformatics* 25, 1189–1191.

- White, H.E., Saibil, H.R., Ignatiou, A., and Orlova, E.V. (2004). Recognition and Separation of Single Particles with Size Variation by Statistical Analysis of their Images. *J. Mol. Biol.* *336*, 453–460.
- Williams, D.B., and Carter, C.B. (2009). *Transmission Electron Microscopy: A Textbook for Materials Science* (Springer US).
- Wilson, B.J., Tremblay, A.M., Deblois, G., Sylvain-Drolet, G., and Giguère, V. (2010). An Acetylation Switch Modulates the Transcriptional Activity of Estrogen-Related Receptor  $\alpha$ . *Mol. Endocrinol.* *24*, 1349–1358.
- Wurtz, J.M., Bourguet, W., Renaud, J.P., Vivat, V., Chambon, P., Moras, D., and Gronemeyer, H. (1996). A canonical structure for the ligand-binding domain of nuclear receptors. *Nat. Struct. Biol.* *3*, 87–94.
- Wurtz, J.-M., Egner, U., Heinrich, N., Moras, D., and Mueller-Fahrnow, A. (1998). Three-Dimensional Models of Estrogen Receptor Ligand Binding Domain Complexes, Based on Related Crystal Structures and Mutational and Structure–Activity Relationship Data. *J. Med. Chem.* *41*, 1803–1814.
- Yamamoto, K.R., and Alberts, B. (1975). The interaction of estradiol-receptor protein with the genome: an argument for the existence of undetected specific sites. *Cell* *4*, 301–310.
- Yang, C., Zhou, D., and Chen, S. (1998). Modulation of Aromatase Expression in the Breast Tissue by ER $\alpha$ -1 Orphan Receptor. *Cancer Res.* *58*, 5695–5700.
- Yang, N., Shigeta, H., Shi, H., and Teng, C.T. (1996). Estrogen-related Receptor, hERR1, Modulates Estrogen Receptor-mediated Response of Human Lactoferrin Gene Promoter. *J. Biol. Chem.* *271*, 5795–5804.
- Yang, X., Downes, M., Yu, R.T., Bookout, A.L., He, W., Straume, M., Mangelsdorf, D.J., and Evans, R.M. (2006). Nuclear Receptor Expression Links the Circadian Clock to Metabolism. *Cell* *126*, 801–810.
- Yi, P., Wang, Z., Feng, Q., Pintilie, G.D., Foulds, C.E., Lanz, R.B., Ludtke, S.J., Schmid, M.F., Chiu, W., and O'Malley, B.W. (2015). The Structure of A Biologically Active Estrogen Receptor-Coactivator Complex on DNA. *Mol. Cell* *57*, 1047–1058.
- Yi, P., Wang, Z., Feng, Q., Chou, C.-K., Pintilie, G.D., Shen, H., Foulds, C.E., Fan, G., Serysheva, I., Ludtke, S.J., et al. (2017). Structural and Functional Impacts of ER Coactivator Sequential Recruitment. *Mol. Cell* *67*, 733-743.e4.
- Yoon, J.C., Puigserver, P., Chen, G., Donovan, J., Wu, Z., Rhee, J., Adelmant, G., Stafford, J., Kahn, C.R., Granner, D.K., et al. (2001). Control of hepatic gluconeogenesis through the transcriptional coactivator PGC-1. *Nature* *413*, 131–138.
- Yu, X., Jin, L., and Zhou, Z.H. (2008). 3.88 Å structure of cytoplasmic polyhedrosis virus by cryo-electron microscopy. *Nature* *453*, 415–419.
- Zhang, J., Kalkum, M., Chait, B.T., and Roeder, R.G. (2002). The N-CoR-HDAC3 Nuclear Receptor Corepressor Complex Inhibits the JNK Pathway through the Integral Subunit GPS2. *Mol. Cell* *9*, 611–623.

## Bibliographie

Zhang, X., Settembre, E., Xu, C., Dormitzer, P.R., Bellamy, R., Harrison, S.C., and Grigorieff, N. (2008). Near-atomic resolution using electron cryomicroscopy and single-particle reconstruction. *Proc. Natl. Acad. Sci.* *105*, 1867–1872.

Zhao, Y., Zhang, K., Giesy, J.P., and Hu, J. (2015). Families of Nuclear Receptors in Vertebrate Models: Characteristic and Comparative Toxicological Perspective. *Sci. Rep.* *5*.

(1999). A Unified Nomenclature System for the Nuclear Receptor Superfamily. *Cell* *97*, 161–163.



# Origine et évolution des récepteurs nucléaires et étude structurale du premier stéroïdien, ERR

## Résumé en français

Les récepteurs nucléaires (RNs) sont des facteurs de transcriptions se liant à des séquences spécifiques d'ADN et activant la transcription de gènes en réponse à la fixation de ligands spécifiques. Parmi tous les RNs impliqués dans l'étiologie des cancers, les récepteurs liés aux œstrogènes ERR jouent un rôle important dans les cancers du sein, de l'ovaire, du colon, de l'endomètre et la prostate. Ce RN est dit orphelin car il ne possède pas de ligand naturel connu à ce jour.

Par une approche de biologie structurale intégrative combinant cryo-microscopie électronique, bioinformatique et évolution, mon travail de thèse s'est focalisé sur l'étude structurale de ERR et sur l'origine et l'évolution des RNs. Dans ce contexte, 3 outils informatiques ont été développés.

Les résultats obtenus ont permis d'une part la révision des connaissances fondamentales sur l'origine des récepteurs nucléaires et leur évolution. D'autre part, l'étude structurale de ERR a permis d'acquérir de nouvelles données sur la topologie des récepteurs nucléaires stéroïdiens fixés sur un élément de réponse ERRE/ERE ainsi que sur le mécanisme allostérique de la liaison du coactivateur PGC-1 $\alpha$  sur le dimère de ERR.

La résolution du complexe à l'échelle atomique par cryo-microscopie électronique permettra d'ouvrir la voie vers la conception de nouvelles molécules thérapeutiques.

Mots-clés : ERR; Récepteur nucléaire; Régulation transcriptionnelle; Biologie structurale; Cryo-microscopie électronique; Bioinformatique; Evolution.

## Résumé en anglais

Nuclear receptors (NRs) are transcription factors which bind to specific DNA sequences and activate gene transcription in response to the binding of specific ligands. Among all of the RNs involved in the etiology of cancers, ERR estrogen receptors play an important role in breast, ovarian, colon, endometrial and prostate cancers. This NR is said to be orphan because it does not have a natural ligand known to date. Using an integrative structural biology approach combining cryo-electron microscopy, bioinformatics and evolution, my PhD work focused on the structural study of ERR and the origin and evolution of RNs. In this context, three informatic tools have been developed.

The results obtained allowed, on the one hand, the revision of fundamental knowledge on the origin of nuclear receptors and their evolution. On the other hand, structural study of ERR allow us to acquire new data on topology of steroid nuclear receptors fixed on an element of ERRE / ERE response as well as on the allosteric mechanism of the binding of the coactivator PGC-1 $\alpha$  on the dimer of ERR.

The resolution of the complex at the atomic scale by cryo-electron microscopy will open the way towards the design of new therapeutic molecules.

Key-words: ERR; Nuclear receptor; Transcriptional regulation; Structural biology; Cryo-electron microscopy; bioinformatics; Evolution.