UNIVERSITÉ DE STRASBOURG

ÉCOLE DOCTORALE des Sciences de la Vie et de la Santé

Centre de Biologie Intégrative, IGBMC, UMR7104, Illkirch

THÈSE présentée par :

Grigory SHAROV

soutenue le : **18 septembre 2015**

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

Spécialité : Biophysique et biologie structurale

Étude structurale du co-activateur transcriptionnel SAGA et de son module d'acétylation des histones

THÈSE dirigée par :

Dr. SCHULTZ Patrick	Directeur de recherche, IGBMC, Strasbourg

RAPPORTEURS :

Dr. TIMMERS Marc	Professeur, University Medical Center, Utrecht
Dr. SCHOEHN Guy	Directeur de recherche, Institut de Biologie Structurale, Grenoble

---

PRESIDENT DE JURY :

Dr. DEVYS Didier	Directeur de recherche, IGBMC, Strasbourg

Dédiée à Mylène

Acknowledgments

My thesis would be impossible without constant support and help of several people, and I would like to express my gratitude to some of them here.

First of all, I would like to sincerely thank my supervisor Dr. Patrick Schultz, for giving me an opportunity to work on a fascinating subject of transcription complexes. I'm very grateful for his constant guidance, support and patience during all four years of my PhD. Thank you for providing me an inspiration and motivation, without which I probably could not reach this point.

I would like to equally thank jury members Dr. Marc Timmers, Dr. Guy Schoehn and Dr. Didier Devys who accepted the invitation for my defense and agreed to review and evaluate my work.

I am thankful to all past and present members of my group, especially Christine for our small talks, creating an amazing atmosphere in the lab, always being supportive and teaching me all old-school methods of sample and grid preparation; Corinne, our best bowling player, for her constant questions I was always glad to answer ☺ and training me at F20; Gabor for teaching me the basics of EM and image processing, always showing me the right way of doing things. I'd like to thank Alexandre, who was patient enough to work with me for three years and taught me most of the biochemistry that I almost forgot, it was a pleasure to work together and to learn from you! I thank Adam for our extremely fruitful discussions and “10-min talks”, as well as his invaluable help with SAGA preparations. I am also grateful to Pélagie, our master student who was helping me with HAT purifications; and Ferdinand for our enthusiastic tea-time discussions. I also thank Céline for always bringing her positive spirit to our lab and solving numerous organizational issues.

I am thankful to all the members of Bruno Klaholz team, especially Jean-François for all the help with microscopes. Most importantly, I would like to thank Sasha for teaching me all the practical tricks of EM field, for our endless discussions and long microscope sessions, for constantly being supportive and motivated to share his knowledge and experience.

I would like to thank all my friends and colleagues in ESBS/CEBGS and now in CBI, as well as in IGBMC, who made these four years here a truly memorable experience. Anna, Claire Bedez and Claire Batisse, Firas, Judit, Heena, Karima, Laura, Martin, Nicolas, Valerie, Valeriya. Thanks to Alastair for his grumpiness ☻, my best friend Simone for sharing the food, and of course, and all the other members of X-ray team: Alexey, Daniya, David, Justine, Irina, Iskander, Mélanie, Mumin, Rustam, Nico, Sergey, Valerio and Yuzuru. Thank you for making this journey truly enjoyable!

I am thankful to my friends Alexander, Alexey, Igor and Vladislav for our everyday political discussions during lunch ☺

I am very grateful to our “service info” team for solving all our network / hardware / software problems and creating a powerful infrastructure so that I could use all the nodes ☺: Arnaud, Jonathan et Véronique.

Special thank to all the people of IGBMC platforms, especially to Isabelle for providing me infected insect cells every time I needed them; Edouard, Karim and Paola for teaching me some cloning and construction of the plasmids; Adeline for indispensable MS analysis; Catherine for all her help with using the proteomics platform and surveying if I use the centrifuges correctly.

I would like to take the opportunity to thank our collaborators from Laszlo Tora group, especially Anne Riss for introducing me into HAT story, Nikolaos for “secretly” providing me the plasmids, Elisabeth and Marjorie for teaching me to do immunoprecipitations and nuclear extract preparations. Also I thank Simon Trowitzsch from Imre Berger's team for tutoring me about GraFix and HAT preps; Nha-Thi and Noëlle Potier for all the work they've done on cross-linking/MS project.

Most importantly I would like to thank my family for always supporting me, when I was far away from home. Sincere thanks to letting me follow my scientific passion and encouraging me all the time during this long journey.

List of abbreviations

- CC – cross-correlation
- C_c – chromatic aberration
- CCD – charge-coupled device
- cryo-EM – cryo electron microscopy
- C_s – spherical aberration
- CTF – contrast transfer function
- CX/MS – cross-linking, coupled to mass spectrometry
- DDD – direct detection device
- DQE – detective quantum efficiency
- dUb – deubiquitination
- EM – electron microscopy
- ESI – electrospray ionization
- FEG – field emission gun
- FSC – Fourier shell correlation
- GraFix – gradient fixation
- GSD – ground state depletion microscopy
- GTF – general transcription factor
- HAT – histone acetyltransferase
- HFD – histone fold domain
- IMAC – immobilized metal ion affinity chromatography
- IR – isomorphous replacement
- LC-MS/MS – liquid chromatography, coupled to tandem mass spectrometry
- MAD – multiple-wavelength anomalous dispersion
- MALDI – matrix-assisted laser desorption ionization
- maxCC – maximum cross-correlation
- ML – maximum likelihood
- MR – molecular replacement
- MS, MS/MS – mass spectrometry, tandem mass spectrometry
- NER – nucleotide excision repair

- NMR – nuclear magnetic resonance
- NPC – nuclear pore complex
- PALM – photo-activated localization microscopy
- PDB – protein database
- PHD – plant homeodomain
- PIC – pre-initiation complex
- PolyQ – polyglutamine
- RCT – random conical tilt
- RP – regulatory particle
- SANS – small angle neutron scattering
- SAXS – small angle X-ray scattering
- SBP – streptavidin-binding peptide
- SNR – signal-to-noise ratio
- SPA – single particle analysis
- STED – stimulated emission depletion microscopy
- STORM – stochastic optical reconstruction microscopy
- TAP – tandem affinity purification
- TEM – transmission electron microscopy
- TSS – transcription start site
- UAS – upstream activation sequence
- XC – X-ray crystallography

Gene and proteins mentioned in this work are named according to the following convention:

- yeast gene symbols comprise three italic lowercase letters and an Arabic number (*gal4*)
- yeast proteins are referred to by the non-italic, initial letter uppercase symbols (Gal4)
- human gene and protein symbols are all uppercase: gene names are in italic (*ADA2A*), while protein names are not (ADA2A)

Species names are mentioned according to the binomial nomenclature and italicized (*Saccharomyces cerevisiae*)

Table of contents

List of figures

List of tables

Résumé

La transcription de gènes par l'ARN polymérase II est étroitement contrôlée et requière l'action concertée d'un grand nombre de protéines. Celles-ci déclenchent alors l'assemblage spécifique du Complexe de Pré-Initiation (PIC) au niveau du promoteur des gènes transcrits, conduisant au positionnement correct de l'ARN polymérase II sur le site de démarrage de la transcription. Les activateurs de transcription, séquence-spécifique ainsi que les modifications post-traductionnelles des histones du nucléosome contribuent au recrutement de complexes co-activateurs formés de multiples sous-unités, agissant comme agents de pontage entre ces activateurs de transcription et le PIC. Les co-activateurs modifient à leur tour la structure de la chromatine dans la région du promoteur des gènes. Ils coordonnent l'assemblage du PIC avec des modifications épigénétiques de la chromatine et des événements de signalisation médiée par les activateurs.

Le complexe co-activateur SAGA est un exemple parfait de cette classe de régulateurs de la transcription. SAGA régule l'expression de nombreux gènes dans la levure, la drosophile et les mammifères, et il a été démontré qu'il contribue à de nombreux processus fondamentaux: la prolifération cellulaire, le développement des métazoaires, la réponse à des signaux de stress. SAGA interagit avec de nombreux activateurs de transcription, tels que Gal4 ou Gcn4 chez la levure et les oncogènes MYC ou E2F chez l'homme. SAGA contient plusieurs domaines protéiques qui reconnaissent les modifications post-traductionnelles des histones ainsi que les deux activités enzymatiques. Le profil de l'expression de l'acétylation de l'histone H3K9 et de la déubiquitination de l'histone H2B suggère que SAGA acétyle les promoteurs et déubiquitine les régions transcrites de tous les gènes exprimés dans la levure et les cellules humaines, et est impliqué dans la transcription par l'ARN Pol II.

Chez la levure, le complexe SAGA, d'un poids moléculaire de 1.8 MDa, est composé de 19 sous-unités distinctes, qui ont toutes, sauf une, un homologue chez les eucaryotes supérieurs. D'importants efforts de recherche ont été consentis pour comprendre l'organisation moléculaire de SAGA mais à ce jour, aucune structure à haute résolution permettant le positionnement tridimensionnel de chaque sous-unité n'est disponible. Le complexe SAGA a été décrit comme ayant une organisation modulaire par des études génétiques de complémentation, de modèles en microscopie électronique, de protéomique quantitative sur des mutants de délétion ainsi que par des expériences de pontage chimique couplées à la spectrométrie de masse. Toutefois, en fonction de la méthode utilisée, les interactions des sous-unités et leurs positions relatives sont différentes.

SAGA contient deux activités enzymatiques permettant d'acétyler ou de deubiquitiner les extrémités N-terminales des histones nucléosomales. La déficience d'une de ces enzymes de SAGA conduit à une létalité embryonnaire et des mutations dans ces sous-unités provoquent des syndromes de neurodégénérescence ou contribuent à la tumorigenèse chez l'homme. L'activité Histone Acétyl-Transférase (HAT) est exercée par la sous-unité Gcn5 et modulée par les sous-unités Ada2 et Ada3. Avec Sgf29, ces 4 sous-unités forment le module HAT, un régulateur majeur de l'acétylation des histones H3 dans les cellules de levure. La localisation spatiale du module HAT au sein du complexe SAGA est en débat.

SAGA est nécessaire au recrutement de la machinerie de transcription sur les promoteurs SAGA-dépendants. Les sous-unités Spt3 et Spt8 ont toutes les deux été décrites comme interagissant directement avec TBP (TATA-box Binding Protein), indiquant ainsi le rôle de SAGA dans la régulation de l'interaction promoteur-TBP. TBP peut se lier au complexe SAGA, mais contrairement aux études d'interaction réalisées sur des sous-unités isolées, cette interaction nécessite Spt8 et Ada3 et est indépendante de Spt3. Les positions des sous-unités Spt3, Spt8 et le site d'interaction de TBP ont été cartographiées dans les vues 2D obtenues par microscopie électronique, cependant des résultats contradictoires ont été rapportés.

Ainsi, à ce jour, la position du module HAT et de plusieurs sous-unités clés tels que Gcn5, Spt3 et Spt20 reste controversée. L'objectif de cette étude est de comprendre l'organisation structurelle du co-activateur transcriptionnel SAGA, permettant d'interpréter les fonctions du complexe. Au cours de ce travail, les questions suivantes seront examinées:

- **Quelles sont l'architecture du module HAT, les interactions des ses différentes sous-unités et leur localisation dans le complexe SAGA?**
- **Quelle est l'organisation moléculaire du complexe SAGA chez la levure ?**
- **Comment SAGA interagit avec TBP?**

Pour répondre à ces questions, différentes approches ont été utilisées: la biologie moléculaire, la purification de protéines, la spectrométrie de masse, la modélisation par homologie ainsi que la microscopie électronique associée à l'analyse de particules isolées.

La spectrométrie de masse, couplée à des expériences de pontage chimique a été utilisée en combinaison avec la modélisation par homologie afin de construire un modèle structural à faible résolution du module HAT. Nous avons exprimé une forme recombinante du module HAT humain, composée de 4 sous-unités, dans des cellules d'insectes en utilisant le système d'expression du baculovirus, et développé un protocole de purification qui nous a permis de produire ce module HAT recombinant à une concentration de 1-2 mg/ml. Après pontage, une masse moléculaire unique de 241 kDa a été détectée par spectrométrie de masse correspondant à une stœchiométrie de 1:1:1:1 pour chacune des quatre sous-unités du module. Afin d'obtenir un aperçu de l'organisation topologique du module HAT de SAGA, le complexe ponté a été protéolysé et le mélange de peptides résultant analysé sur un spectromètre de masse. Une carte de liaison a été produite pour visualiser les liens intra- et inter-sous-unités (fig. 1).

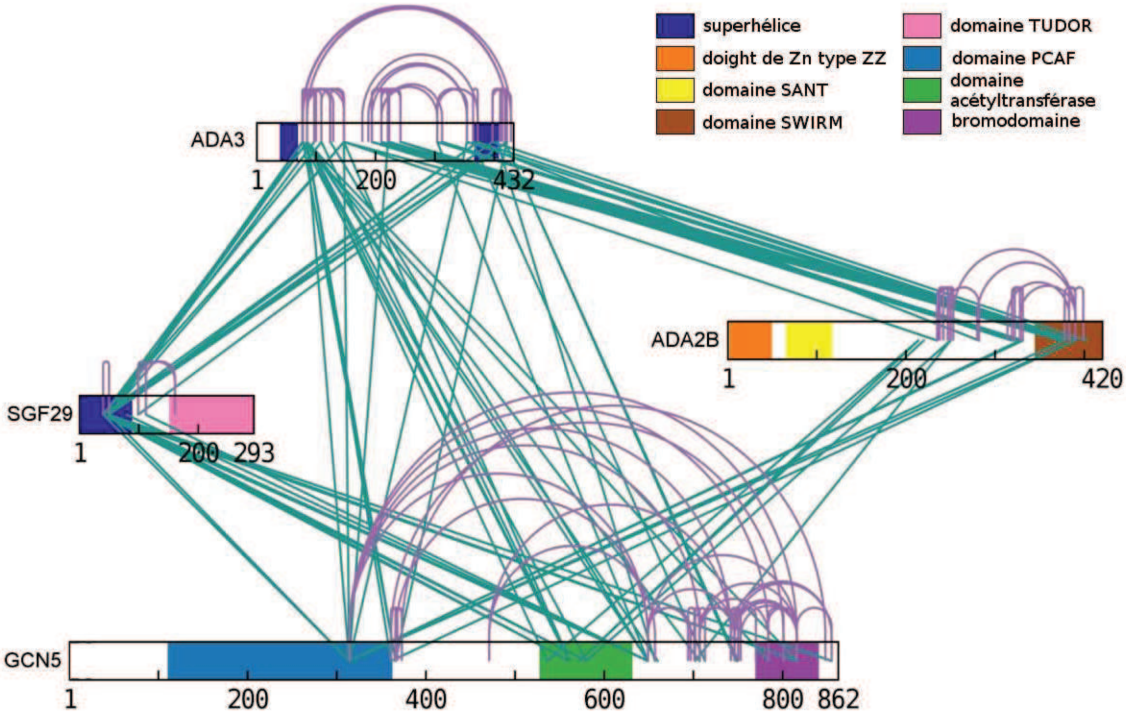

Fig. 1. Carte de liaison entre les sous-unités du module HAT recombinant humain.

Les liens inter-sous-unités sont verts, les liens intra-sous-unités sont violets. Les domaines structuraux de sous-unités HAT sont indiqués.

Cette carte met clairement en évidence plusieurs domaines interdépendants discrets: un pontage massif a été observé entre ADA3 et ADA2B suggérant un solide réseau de connexions entre ces deux sous-unités. Aucun lien croisé n’a pu être détecté entre SGF29 et ADA2B, ni sur la partie N-terminale de GCN5, ni sur le domaine TUDOR de SGF29 et les domaines ZZ et SANT de ADA2B. Curieusement, ces domaines non pontés sont souvent impliqués dans les interactions avec les histones modifiées. Ils sont susceptibles d'être exposés afin de se lier aux nucléosomes et ainsi interagissent moins avec les autres sous-unités du module. La carte topologique obtenue a été utilisée pour construire un modèle du module HAT sur la base de structures atomiques de domaines protéiques disponibles de HAT. Des modèles par homologie ont été construits pour les sous-unités n'ayant aucune information structurale disponible et en tenant compte des liens intra-sous-unités. Afin d'obtenir un modèle à faible résolution du sous-complexe HAT, les quatre sous-unités ont été positionnées l'une par rapport aux autres (fig. 2): les distances entre les lysines pontées chimiquement ont été réduites au minimum en évitant des conflits stériques entre les modèles atomiques. Plusieurs parties non structurées de sous-unités manquent dans le module HAT obtenu. L' application de méthodes de dynamique moléculaire pourra améliorer ce modèle issu d'une approche de l'amarrage moléculaire en corps rigide.

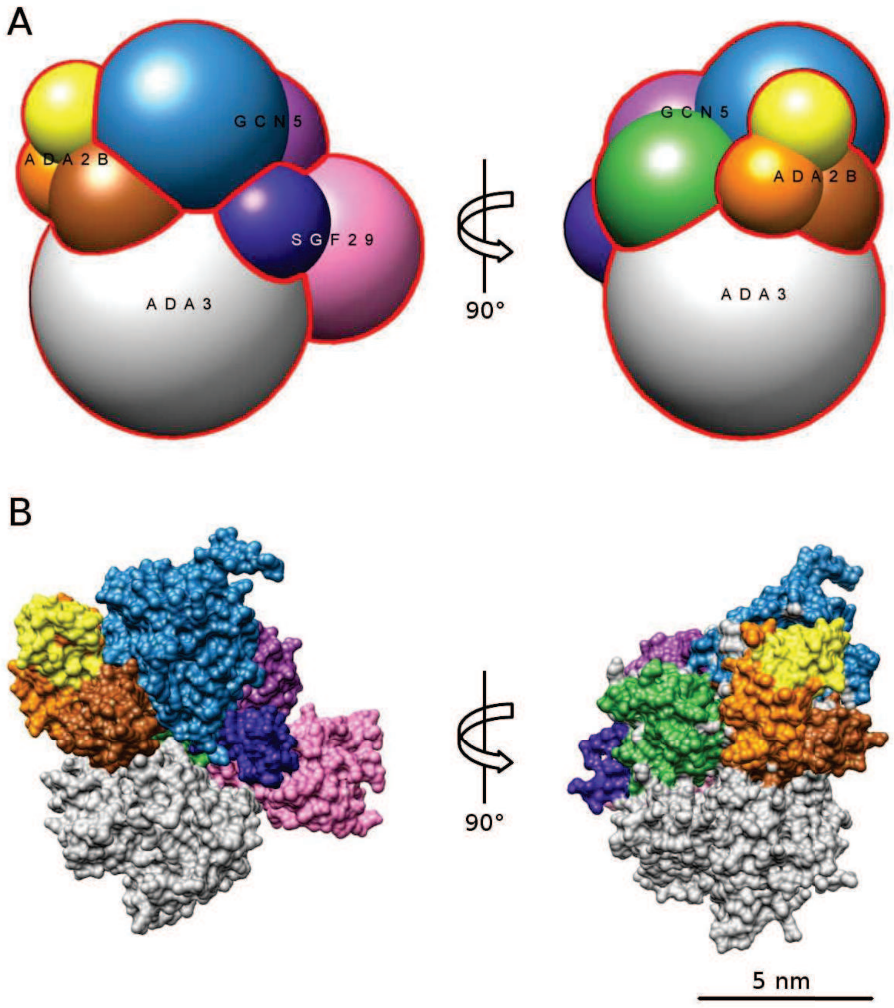

Fig. 2. Organisation quaternaire du module HAT.

(a) – modèle du module HAT obtenu en utilisant les données de pontage chimique, la modélisation par homologie et l'amarrage moléculaire; (b) – représentation de surface du modèle HAT composé des structures atomiques de domaines conservés. Le modèle représente 63% des résidus de HAT.

Nous avons comparé nos résultats avec ceux d'une carte de liaison récemment publiée (Han et al., 2014) pour le complexe SAGA complet de levure. Malgré une distribution de lysines distincte dans les séquences des deux organismes, le réseau de liaisons entre les domaines est hautement conservé à l'exception notable de Sgf29. Un grand nombre de liaisons intra-sous-unités n'interagit qu'avec Ada3 tandis que SGF29 dispose d'un réseau d'interaction fort avec la plupart des domaines de GCN5. L'interaction entre les sous-unités Sgf29 et Gcn5 chez la levure avait déjà été démontrée par des analyses systématiques par délétion. Les mêmes régions de la sous-unité GCN5 ont des liens croisés avec ADA3 tout au long de sa séquence tandis que chez la levure le lien entre ces deux sous-unités est limité à un seul domaine dans la partie C-terminale de Ada3. Plutôt que de refléter les différences dans la séquence primaire entre les deux organismes (levure et humain), nous suggérons que ces différences dans le réseau de liaisons reflètent un changement conformationnel du module HAT entre sa forme isolée et sa forme incorporée dans le complexe SAGA. De manière intéressante, les domaines qui montrent moins de liaisons transversales au sein du module HAT de levure sont fortement impliqués dans les interactions avec les autres sous-unités de SAGA. Ainsi, la sous-unité Ada3 de levure forme peu de liens avec les autres sous-unités du module HAT, mais forme de nombreux liens croisés avec d'autres protéines comme Taf5, Taf6, Taf12, Spt7, Sgf73 et Ada1. Nous avons montré que dans le sous-complexe humain, ADA3 est largement ponté avec les autres sous-unités du module HAT: ADA2B, SGF29 et GCN5. Ces résultats suggèrent que le module HAT est réorganisé lors de son intégration dans le complexe SAGA.

Pour positionner le module HAT au sein du complexe SAGA, nous avons purifié celui-ci à partir de souche de levure mutante, où l'ensemble du module HAT est détaché. Le complexe mutant a été coloré négativement et analysé par microscopie électronique, puis sa structure comparée à celle du complexe SAGA natif. Cette analyse a révélé deux densités manquantes (fig. 3): une correspondant au module de déubiquitination (dUb, domaine III), dont la position a été déterminée dans une étude précédente (Durand et al., 2014), et l'autre, située dans la partie la plus flexible du complexe – le domaine V du lobe B. Le complexe mutant, dans son ensemble, affiche une hétérogénéité encore plus élevée que dans le complexe natif, puisque le lobe B est plus flexible, ce qui limite l'analyse à haute résolution. Nous pensons que la perte du module dUb dans le complexe mutant découle du protocole de purification: l'élution avec de l'EGTA déstabilise les doigts de Zn présents dans plusieurs sous-unités du module dUb. L'emplacement du module HAT dans l'extrémité flexible du lobe B (domaine V) est confirmé dans une étude récente publiée par le groupe de C. Yip (Setiaputra et al., 2015).

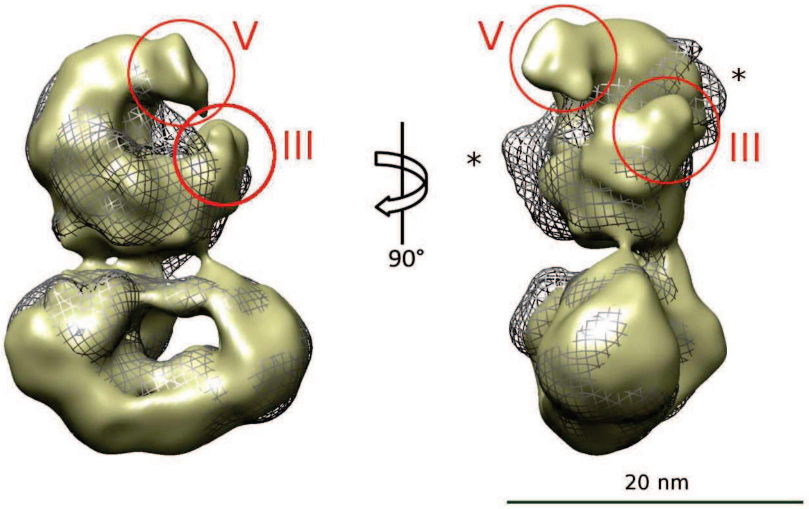

Fig. 3. Comparaison entre la structure de mutant SAGA sans HAT (mailles grises) et la structure de SAGA native (en jaune).

Densités supplémentaires sont visibles dans les domaines III et V. Plusieurs différences mineures (marquées par des astérisques) entre les cartes 3D sont également marquées.

Pour améliorer les connaissances actuelles du complexe SAGA, nous avons déterminé sa structure en utilisant la cryo-microscopie électronique (cryo-EM) dans des conditions quasi-natives. Les tentatives précédentes pour résoudre la structure de SAGA ont été limitées par un faible rendement de purification par TAP tag et l'utilisation de la coloration négative. Un nouveau protocole de purification a été développé au laboratoire et a permis d'obtenir des échantillons très homogènes et stables du complexe SAGA endogène de levure. Pour stabiliser le complexe purifié et empêcher sa dissociation durant la préparation des échantillons, nous avons effectué un pontage chimique avec la glutaraldéhyde. L'analyse des images du complexe SAGA totalement hydraté obtenues par cryo-EM nous a permis d'atteindre une résolution d'environ 15 Å et représente la première structure par cryo-EM de ce complexe co-activateur à une telle résolution. La structure révèle que le complexe est composé de deux grands lobes: le lobe A contenant les sous-unités Tra1 et certaines TAFs, et le lobe B, contenant aussi des TAFs, Ada1, les modules HAT et dUb, et les sous-unités SPT. En comparaison avec les études antérieures, nous avons analysé la flexibilité du lobe B et caractérisé deux types de mouvement: l'ouverture et la fermeture d'une pince moléculaire formée par le lobe B d'une part (fig. 4), la rotation complexe (tourbillonnant) du lobe B par rapport au lobe A d'autre part (fig. 5). Cette flexibilité représente un obstacle majeur à l'amélioration de la résolution de la structure de SAGA.

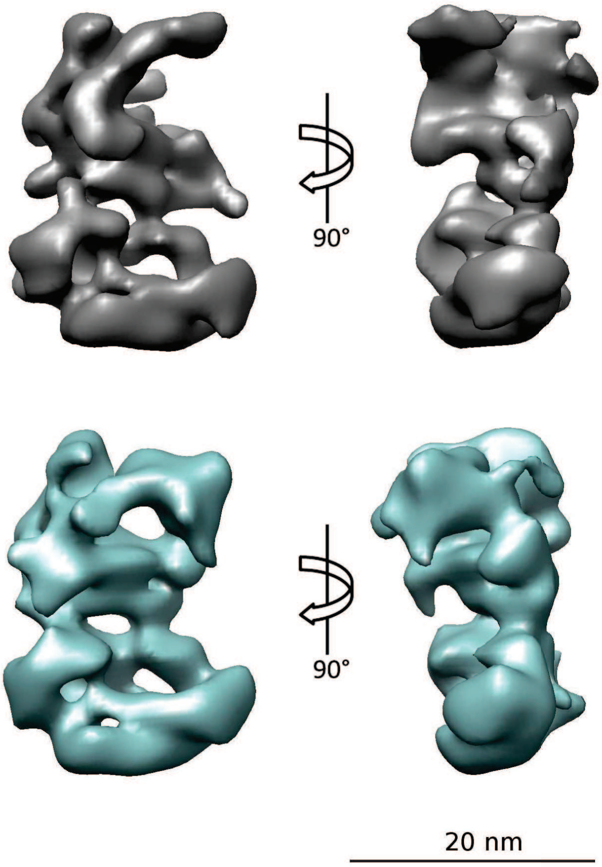

Fig. 4. Deux conformations distinctes du lobe B de SAGA obtenues par cryo-EM: ouverte (gris) et fermée (vert).

Les domaines du lobe B adoptent des positions différentes en facilitant la fermeture de la pince moléculaire.

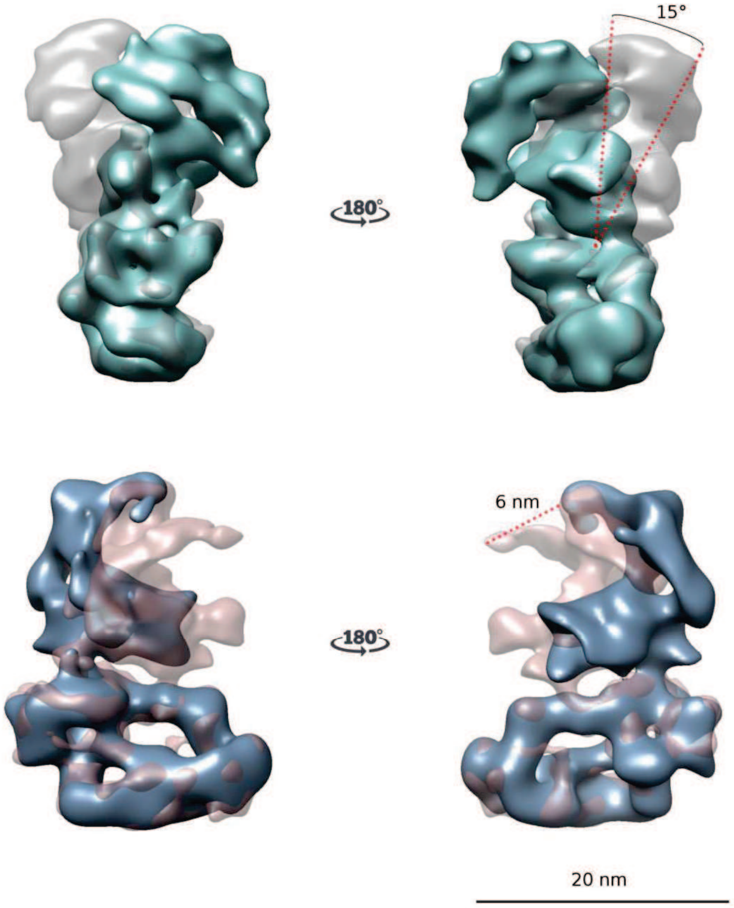

Fig. 5. Les conformations différentes de SAGA montrant le mouvement complexe du lobe B.

Pour étudier la liaison de TBP, SAGA a été incubé avec un excès de la protéine TBP purifiée et le complexe reconstitué a été analysé par microscopie électronique en coloration négative. Le contrôle témoin de SAGA sans TBP a été préparé de la même manière. L’analyse des deux jeux de données a abouti à des reconstitutions en 3D, qui ont été comparées par différence de densité. Dans les deux cas, nous avons pu identifier les conformations ouvertes et fermées du lobe B (fig. 6), indiquant que TBP n’est pas impliqué dans l'ouverture et la fermeture de la pince moléculaire. A une résolution d'environ 25 Å nous sommes en mesure d'identifier la position de TBP et du module dUb sur la carte de microscopie électronique (fig. 7). Fait intéressant, dans la conformation fermée de SAGA, on observe que lors de la liaison de TBP le lobe B subit un grand réarrangement de plusieurs domaines.

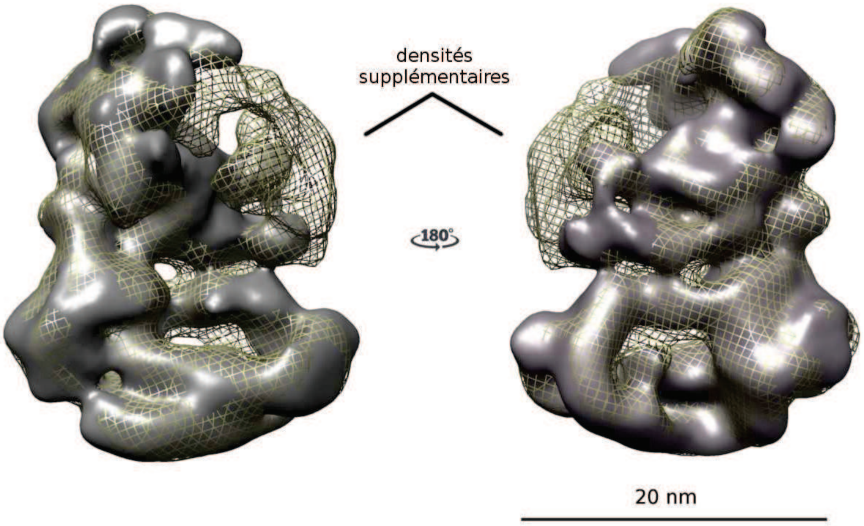

Fig. 6. Superposition entre les conformations ouvertes et fermées du complexe SAGA avec TBP.

L'ouverture et la fermeture de la pince moléculaire sont médiées par les extensions des domaines du lobe B. Les densités supplémentaires entre les conformations sont indiquées.

Au cours de ce travail, nous avons étudié par microscopie électronique l'organisation du complexe SAGA endogène chez la levure et avons obtenu la première structure à résolution intermédiaire (10-15 Å) du complexe entier. Nous avons pu analyser plus précisément et visualiser deux états conformationnels de SAGA: ouvert et fermé. Nous suggérons que cette flexibilité permet à SAGA de s'adapter rapidement aux interactions avec différents substrats et cofacteurs. Nous avons localisé le module HAT et démontré que sa suppression provoque une flexibilité conformationnelle encore plus grande de SAGA. Une grande mobilité du module HAT peut refléter la capacité de SAGA à interagir avec et modifier de nombreux substrats, tels que les histones nucléosomiques. Le module dUb est très sensible aux différentes étapes de purification, et nous avons développé un nouveau protocole de purification, qui assurait la rétention complète de ce module. Nous avons défini le site d'interaction de TBP avec SAGA et mis en évidence sa proximité avec le module dUb. Nous avons observé que TBP peut se lier aux conformations ouvertes et fermées de SAGA. Enfin, le réseau d'interactions entre sous-unités et leur stœchiométrie dans le module HAT a été décrit en analysant un sous-complexe recombinant humain. Sur la base des données de spectrométrie de masse et de la modélisation par homologie, nous avons déterminé le premier modèle à faible résolution du module HAT.

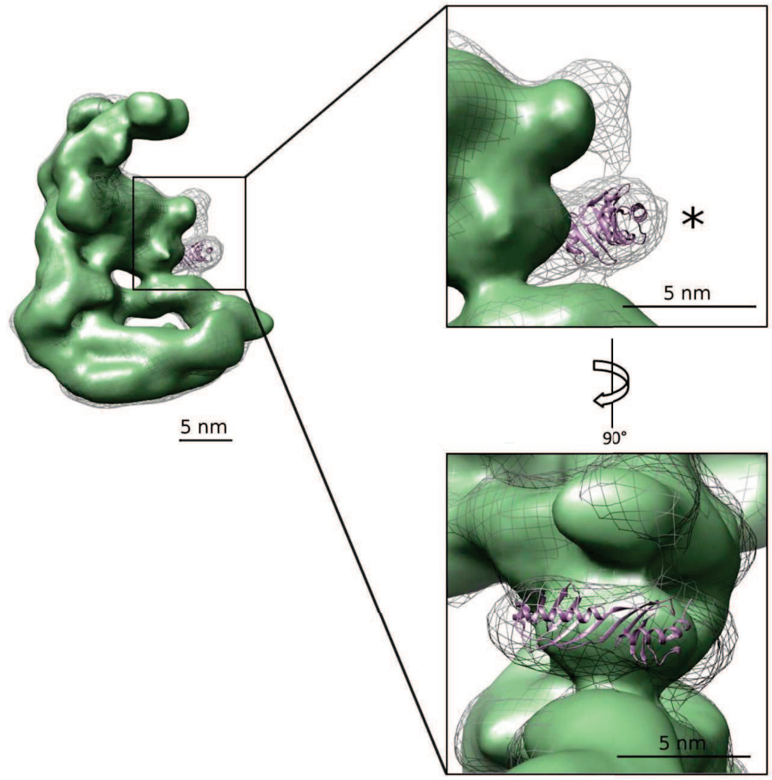

Fig. 7. Superposition des structures 3D de SAGA en conformation ouverte sans (vert) et avec (mailles grises) TBP.

L'astérisque (*) marque la position de TBP. Les coordonnées de TBP sont prises dans PDB 1YTF.

Le complexe SAGA est un acteur majeur de la machinerie transcriptionnelle et intervient dans de nombreuses étapes de régulation de la transcription. Ce travail apporte une nouvelle représentation de l'organisation moléculaire de ce complexe en utilisant principalement la microscopie électronique. Malgré nos efforts, nous n’avons pas encore obtenu une structure de SAGA à haute résolution qui permettrait de positionner sans ambiguïté toutes les sous-unités. L’obstacle principal reste la grande flexibilité du complexe, qui limite la résolution des données structurales. Néanmoins, nous sommes persuadés que l'analyse de grands jeux de données permettra de discriminer les conformations différentes de SAGA et de déterminer sa structure à haute résolution. Beaucoup d'efforts sont consentis pour stabiliser le lobe le plus labile du complexe, et nous explorons la possibilité de stabiliser une conformation unique en utilisant d'autres molécules tels que des nucléosomes.

Chapter 1. Introduction.

1.1. Nucleosomes and chromatin are fundamental structural units of a genome.

Within a cell the DNA, encoding the genetic information, is associated with proteins forming a so-called chromosome. All prokaryotic and eukaryotic cells, and even some viruses, follow this type of DNA organization. Such DNA packaging serves several means: it protects DNA from damage, it provides a unique form of compaction so that DNA can fit into the cell, it supports the processes of correct separation of the genetic information between the daughter cells during cell division and finally regulates DNA accessibility. In eukaryotes, nuclear DNA is associated with small basic proteins called histones, forming a chromatin. Less abundant chromatin-associated proteins are represented by numerous DNA-binding proteins, participating in replication, recombination, repair and transcription.

In total, the DNA of a diploid human cell contains about 6 billion base pairs. Given the rise of a single base pair is 3.4 Å in the DNA double helix (Watson and Crick, 1953), the length of complete linear DNA would be about 2 meters. Thus, in order to fit into 10 µm human cell nucleus, the DNA has to be compacted 20.000 times. The first level of the compaction is achieved by the regular association of DNA with histones into nucleosomes. Eukaryotic nucleosome consists of two copies of each of the 4 core histones H2A, H2B, H3 and H4, around which 147 bp of DNA is wrapped (fig. 8). However, the linker DNA region between the neighboring nucleosomes has a variable length of 20 to 60 bp depending on cell type and species. The core histones assemble pairwise: H3-H4 and H2A-H2B. Nucleosome assembly begins with the formation of a H3-H4 heterotetramer that binds DNA, followed by the association of two H2A-H2B heterodimers. The linker histone H1 is positioned at the diad axis of the nucleosome and binds both DNA linkers, connecting the nucleosome with two neighboring ones. All histones are relatively small 10-15 kDa proteins and 20% of their primary sequence contains either lysines or arginines – positively charged amino acids. These highly basic proteins allow effective binding to negatively charged DNA. Each core histone contains an N-terminal unstructured tail that is not required for nucleosome formation but instead is subjected to numerous post-translational modifications, such as methylation, acetylation, phosphorylation, ubiquitination etc. of mostly serine, lysine and arginine residues. These modifications can dramatically affect nucleosome functioning, as discussed later.

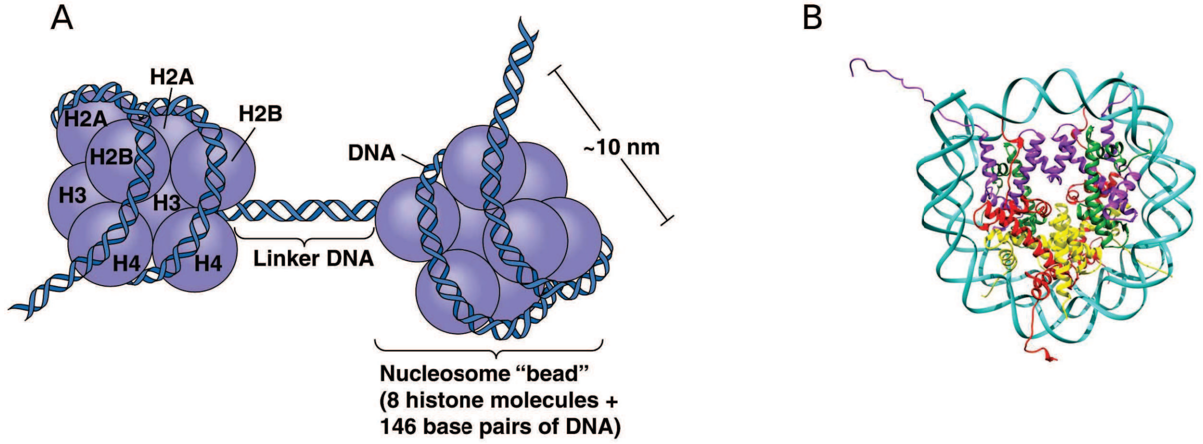

Fig. 8. DNA is packaged into nucleosomes.

(a) Schematic representation of the nucleosomal organization; (b) Crystal structure of the nucleosome core particle showing a histone octamer with a DNA wrapped around: red – H2A, yellow – H2B, purple – H3, green – H4. PDB 1AOI.

The nucleosomal level of chromatin organization provides approximately sixfold compaction. This is far short from the 20.000-fold DNA compaction that is observed in eukaryotic cells. Thus, most of the compaction is supplied by higher-order chromatin organization levels. Observation of the chromosomes in the cell by light and electron microscopy revealed two different types of chromatin: heterochromatin and euchromatin (fig. 9). Heterochromatin has a very dense appearance that is well stained by nucleophylic dyes suggesting a high DNA content. Functional studies indicate that heterochromatin shows very little gene expression suggesting that its highly compact structure represents a natural barrier for accessing DNA by many enzymes. On the contrary, euchromatin has more open structure that is poorly stained with the dyes, and displays high gene expression levels. DNA in both chromatin types is packaged into the nucleosomes, however the higher-order packing of nucleosomes differs in euchromatin and heterochromatin. These levels are shown in the fig. 10: (i) Addition of linker histone H1 promotes the formation of more compact nucleosomal assembly with zig-zag appearance, that can be further compressed into 30-nm fiber, stabilized by interactions of N-terminal histone tails with adjacent nucleosomal histones. These fibers provide about 40-fold compaction of DNA. (ii) Next, such fibers are believed to be further compacted by forming chromatin loops, stabilized by nuclear scaffold structures. (iii) The final level of compaction is the formation of the chromosome structure, still not well understood.

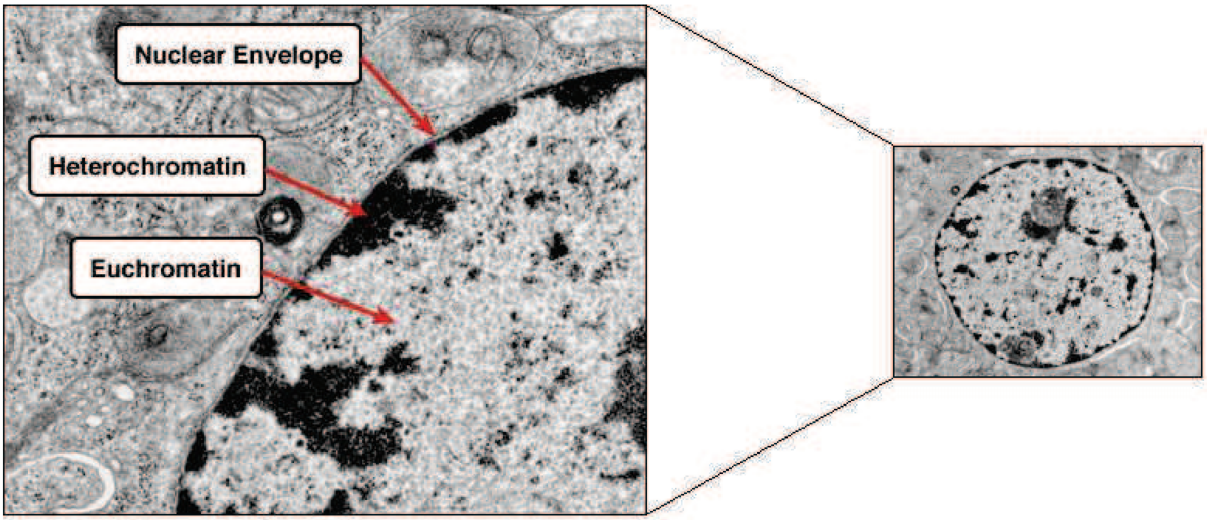

Fig. 9. DNA in the nucleus exists in two forms that reflect the level of cell activity.
Heterochromatin appears as darkly stained irregular particles scattered throughout the nucleus or adjacent to the nuclear envelope. Euchromatin is dispersed and not readily stainable. While active transcription is prevalent in euchromatin regions, heterochromatin regions are mostly inactive. Image is taken from http://medcell.med.yale.edu

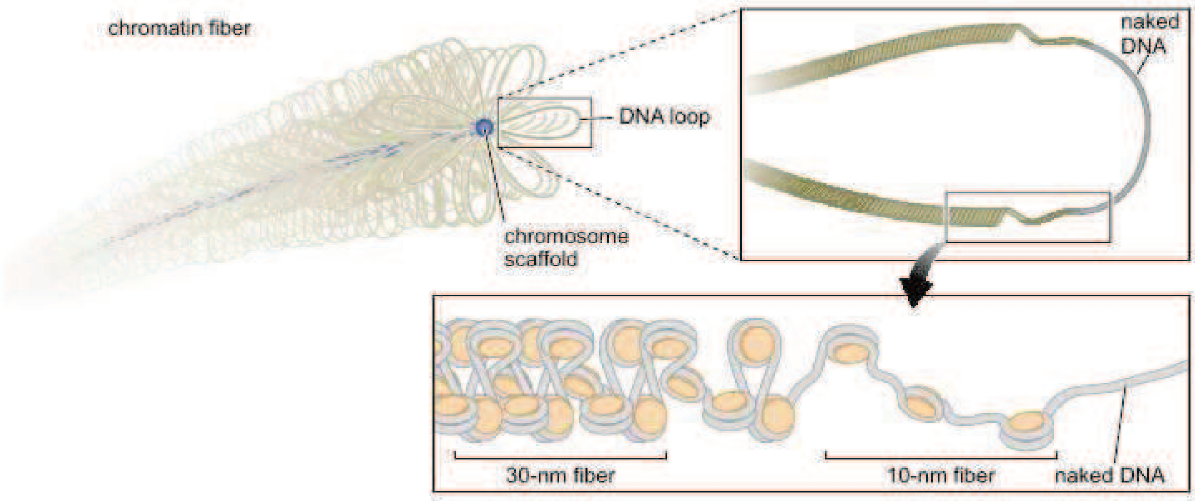

Fig. 10. High-order structure of chromatin.
A model shows how the nucleosomal DNA can be further compacted into a 30-nm fiber, forming large loops tethered by the nuclear scaffold, that are finally compacted into a chromosome. From Watson, 2014.

The dynamic interaction of histones with DNA has profound effects on genome expression. To provide access to non-histone DNA binding proteins, any particular site of DNA can be uncovered from the histone octamer by different mechanisms: nucleosomes can slide along the DNA, unwrapping necessary binding sites, or the histones can be evicted completely from the DNA by dedicated molecular machines. To facilitate this intrinsic ability of nucleosomes to move along the DNA, many chromatin-remodeling complexes exist in the cell. Most of such protein complexes can change interactions of DNA with histones by using the energy of ATP hydrolysis. Besides catalytic ATP-hydrolyzing subunit, these multiprotein complexes often contain specific subunits that target particular DNA sites. Targeting specific DNA regions by remodeling complexes can be mediated by DNA-bound transcription factors. Another possibility is to target specific histone post-translational modifications of nucleosomal histone tails via different protein domains (chromodomains, bromodomains etc.) present in the subunits of remodeling complexes.

1.2. Histone modifications can modulate chromatin accessibility.

Amino-terminal histone tails are often subjected to deposition of different post-translational modifications that are involved in different cellular events (fig. 11). For example, lysine acetylation of H3/H4 is a known hallmark of actively transcribed chromatin (Allfrey et al., 1964), whereas methylation of H3K9 and H3K27 correlates with transcriptional repression (Nakayama et al., 2001). Besides transcription regulation, histone modifications can influence DNA repair, nucleosome assembly, mitosis and meiosis. Multiple histone modifications occurring at specific chromatin sites have led to the hypothesis of a histone code, stating that the combinatorial modification of histone tails constitutes a biological code that can be read, written and erased by specific protein complexes in the cell. Specific acetylation or methylation marks can be recognized by dedicated proteins, that often carry themselves catalytic subunits that can modify histones further. However, it remains highly controversial whether combinations of modifications have distinct meanings. Site-specific DNA-binding proteins remain the most important providers of specificity, determining when a certain gene is expressed or not. Nonetheless, certain modifications are often correlated with a state of gene activity and it is not known whether the structure of chromatin is altered or whether this is simply the consequence of gene activity.

The mechanism of altering nucleosomal DNA structures is not yet completely understood. It is accepted that the deposition of acetylation or phosphorylation marks onto lysines of histone tails decrease their overall positive charge, reducing the affinity of the histones towards negatively charged DNA. Thereby, acetylation impedes formation of the 30-nm chromatin fibers from nucleosomal DNA, repressing the creation of high-order chromatin structures. It is important to note that similar to remodellers, nucleosome modification enzymes are often part of large multiprotein complexes. Both types of complexes can work together to increase accessibility of DNA. For example, certain modifications can recruit the enzymes that perform the same modification as well as remodeling enzymes that propagate changes of chromatin organization in particular DNA regions.

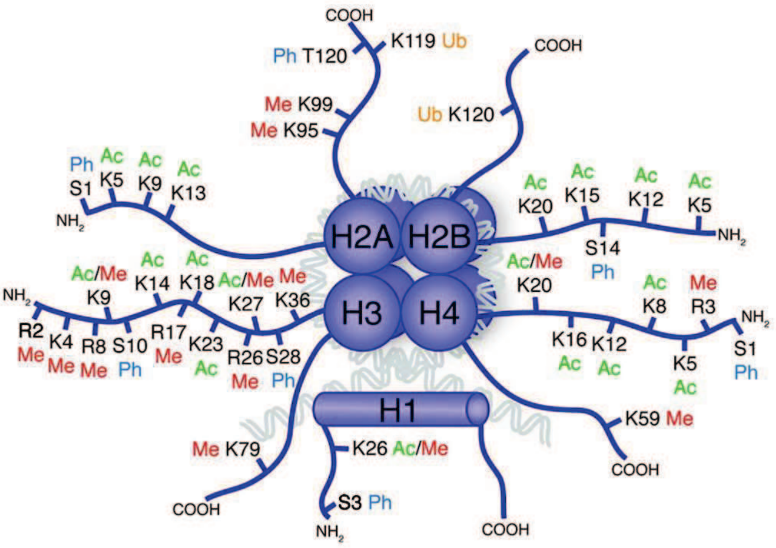

Fig. 11. Most common post-translational histone modifications.

Ac – acetylation, Me – methylation, Ph – phosphorylation, Ub – ubiquitination. From Tollervey and Lunyak, 2012.

1.3. Basic mechanisms of eukaryotic transcription.

Transcription is the first step of genome expression, that provides a flow of genetic information from DNA to RNA. It allows to selectively transcribe specific regions of the genome to different extents, producing from one to several thousands of RNA copies of a given DNA section. The key enzyme catalyzing RNA synthesis is RNA Polymerase (RNA Pol). Bacteria have only one such enzyme, whereas eukaryotic cells have 3 different RNA polymerases: RNA Pol I, II and III (Table 1). RNA Pol II is the most studied one among them, and is responsible for the transcription of all protein-coding genes. RNA Pol I and III are involved in transcribing specific RNA-encoding genes and synthesize major rRNAs, tRNA and other small RNAs, found in nucleus and cytosol. Nevertheless, all prokaryotic and eukaryotic polymerases share a core of subunits that is structurally and evolutionarily highly conserved. All these enzymes have a crab claw shape with pincers formed by the two largest subunits and a cleft between them, forming an active site (fig. 12). Various channels exist to provide easy enter and exit of DNA, RNA and ribonucleotides from active site center.

RNA Pols	**Prokaryotic**	**Eukaryotic**		
Type	Bacterial	RNA Pol I	RNA Pol II	RNA Pol III
Size, kDa (species)	375 (*T. aquaticus*)	589 (*S. cerevisiae*)	514 (*S. cerevisiae*)	693 (*S. cerevisiae*)
Number of subunits	5	14	12	17
Transcription product	mRNA	most of rRNAs	mRNA	tRNA, 5S rRNA

Table 1. Comparison of the common types of RNA polymerases.

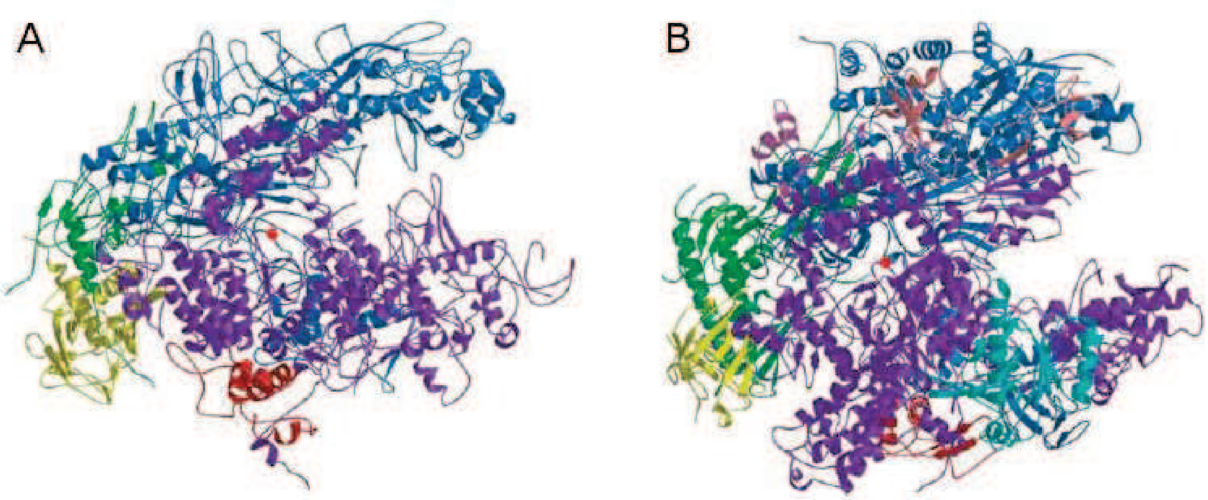

Fig. 12. Comparison of the crystal structures of prokaryotic and eukaryotic RNA polymerases. *(a) Structure of RNA Pol from Thermus aquaticus. The* Mg^{2+} *ion (red ball) marks the active site here and in part b. (b) Structure of RNA Pol II from yeast Saccharomyces cerevisiae. The subunits are colored to show their relatedness to those in the bacterial enzyme. From Watson, 2014.*

Gene transcription can be schematically described by the three steps: initiation, elongation and termination. Briefly, initiation starts with recruiting of RNA Pol (with the help of initiation factors) to the gene promoter – a specific DNA sequence, generally positioned upstream of the transcribed gene. Once the complex between RNA Pol and promoter is formed, the DNA around the active site will start to unwind, producing a “transcription bubble” of single-stranded DNA. One of the DNA strands will act as a template for synthesizing a complementary RNA strand by addition of ribonucleotides at the 3'-end. Once the polymerase has synthesized a short RNA (about 10 bases), it enters the elongation phase. During this phase, the polymerase performs a wide range of functions, besides the main catalytic one: it unwinds DNA in front of the enzyme and re-anneals behind it, dissociates growing RNA chain from template and also performs some proof-reading activity. It also interacts with several regulatory elongation factors, which facilitate transcription through naturally occurring pause sites and thus modulate the transcription elongation rate. After the enzyme reaches the end of the gene or specific termination sequence, it releases the newly synthesized RNA and detaches from the DNA.

1.3.1. Transcription initiation in eukaryotes: PIC assembly.

Transcription initiation of protein-coding genes in prokaryotes requires only one additional initiation factor, whereas in eukaryotic cells six different General Transcription Factors (GTFs) are required for efficient promoter-specific initiation by RNA Pol II on a naked DNA (Table 2). In reality, the situation is even more complicated, since GTFs alone are not sufficient to initiate transcription, and other protein complexes like Mediator complex, different remodellers and often chromatin-modifying enzymes are required.

A set of specific DNA sequences comprising the core promoter is required to accurately initiate transcription. A eukaryotic promoter is usually 40-60 bp long, extending upstream and downstream from the Transcription Start Site (TSS) and includes several sequence elements that are not phylogenetically conserved (fig. 13). The most common promoter elements include (reviewed in Smale and Kadonaga, 2003):

- TATA-box, which is a TA-rich DNA sequence located upstream of the TSS and is known to bind the TATA-box Binding Protein (TBP)
- Initiator element (Inr), located just at the TSS. As well as the TATA-box, Inr can alone initiate transcription in the presence of the GTFs
- BRE, which is a TFIIB recognition element located upstream of TSS
- DPE and DCE that are downstream promoter elements identified in *Drosophila*

Besides these core promoter elements, specific regulatory sequences are located further upstream or downstream (up to several hundreds base pairs from the promoter), that contain sequence-specific binding sites for transcriptional activators or repressors and influence transcription efficiency in response to external regulatory signals (reviewed in Spitz and Furlong, 2012; Pennacchio et al., 2013; Shlyueva et al., 2014).

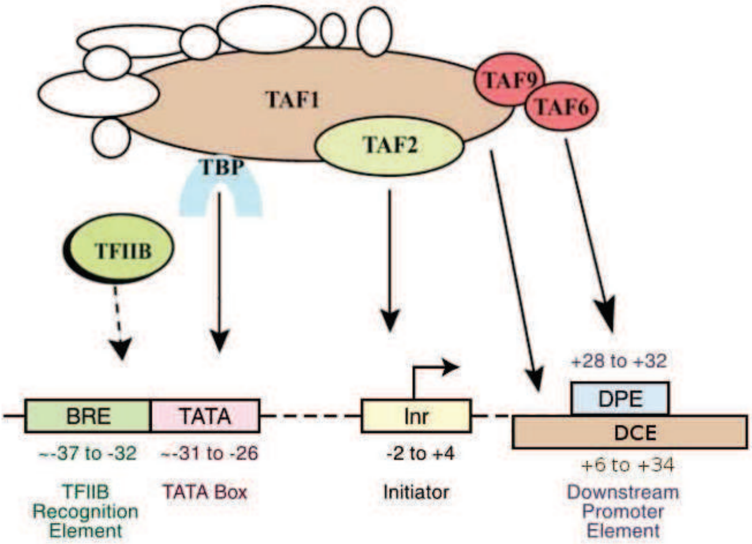

Fig. 13. Schematic organization of eukaryotic promoter.

Common promoter elements and their binding partners. Adapted from Thomas and Chiang, 2006.

The GTFs are required to recruit RNA Pol II to the promoter, activate the RNA Pol II, unwind the DNA (promoter melting), release RNA Pol II from the promoter and successfully switch to the elongation phase. Altogether, RNA Pol II and GTFs, bound to the promoter, form the Pre-Initiation Complex (PIC). The classical model of PIC assembly is a sequential pathway, when different GTFs are added in a stepwise manner (fig. 14; reviewed in Sainsbury et al., 2015). First, TFIID binds the TATA-box in the core promoter via its TBP subunit. Upon binding to the DNA, TBP bends it by an angle of ~ 90°. Although most promoters do not contain canonical TATA-boxes (TATA-less promoters), TBP and other GTFs are genome-wide located on many eukaryotic promoters (Rhee and Pugh, 2012). Several TAF (TBP-Associated Factors) subunits of TFIID were also found to bind promoter elements. For example, Taf1 and Taf2 were shown to recognize the Initiator element (Chalkley and Verrijzer, 1999), while Taf6 and Taf9 was shown to interact with downstream promoter elements in *Drosophila* (Burke and Kadonaga, 1997). In the next step of PIC assembly, TFIIA and TFIIB are recruited to stabilize the TBP-

DNA binary complex. Though TFIIA is not required for basal transcription, it increases TBP affinity to the DNA together with TFIIB (Imbalzano et al., 1994). TFIIB was shown to bridge TATA-bound TBP and RNA Pol II, since the N-terminal domain of TFIIB interacts with RNA Pol II close to the active site (Liu et al., 2010), while its C-terminus binds TBP and upstream DNA (Ha et al., 1993; Nikolov et al., 1995). TFIIF is recruited next in order to position RNA Pol II at the TSS (Eichner et al., 2010). TFIIE and TFIIH are the last GTFs recruited to form the PIC. Promoter melting is mediated by TFIIH (Holstege et al., 1996) and requires ATP hydrolysis. Once the PIC is assembled, RNA Pol II escapes the promoter in two steps: an ATP hydrolysis (Goodrich and Tjian, 1994), followed by the phosphorylation of the C-terminal tail of the largest RNA Pol II subunit by the kinase subunit of TFIIH (Lu et al., 1992).

Name	**Role in transcription initiation**
TFIIA	Interacts with TBP, facilitates TBP binding to the TATA-box
TFIIB	Recognizes BRE elements in promoters, stabilizes TBP-TATA-box complex, recruits RNA Pol II to the promoter and bridges TBP and RNA Pol II
TFIID	Recognizes TATA-box and other promoter elements, regulates DNA binding by TBP
TFIIE	Attracts and regulates TFIIH
TFIIF	Stabilizes interaction between RNA Pol II, TFIID and TFIIB; recruits TFIIE and TFIIH
TFIIH	Unwinds DNA at the TSS; phosphorylates Ser 5 on RNA Pol II, switching from initiation to elongation; releases RNA Pol II from the promoter

Table 2. List of GTFs required for transcription initiation by RNA Pol II.

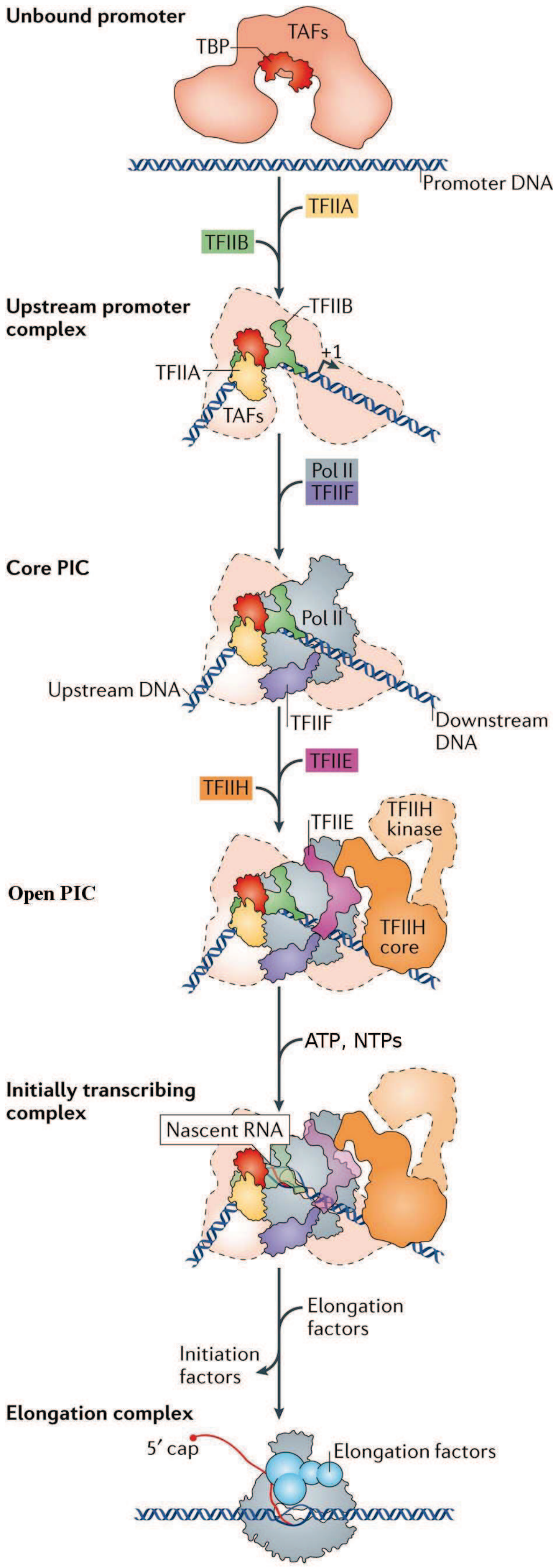

Fig. 14. Classical stepwise model of PIC assembly.

(Legend to the fig. 14 on the previous page).

TFIID via its TBP subunit binds to the promoter DNA, inducing a bend. The TBP-DNA complex is then stabilized by TFIIB and TFIIA, which flank TBP on both sides. The resulting complex is joined by the RNA Pol II-TFIIF complex, leading to the formation of the core PIC. Subsequent binding of TFIIE and TFIIH completes the PIC. In the presence of ATP, the DNA is opened and RNA synthesis starts. Finally, the dissociation of initiation factors enables the formation of the RNA Pol II elongation complex, which is associated with transcription elongation factors (blue). NTP, nucleoside triphosphate. From Sainsbury et al., 2015.

1.3.2. Transcription regulation at the initiation step: activators and coactivators.

The basal transcription machinery, composed of GTFs and RNA Pol II can initiate transcription *in vitro* on a naked DNA template. However, this is not sufficient to initiate transcription *in vivo*. One of the reasons is that DNA compaction prevents easy access to the promoters for transcriptional machinery. In that case, sequence-specific activator proteins bind far upstream from the promoter and recruit coactivator protein complexes, that in turn will recruit GTFs and RNA Pol II to the gene promoter. In general, most activators are comprised of two domains: a DNA-binding domain, which interacts with specific DNA sequences upstream of the promoter, and an activating domain, which binds other proteins of the transcription machinery. Activator-induced transcription can be illustrated on the example of Gal4 – a well-studied yeast transcriptional activator, controlling galactose-mediated induction of gene expression (Giniger et al., 1985). This activator is known to regulate numerous GAL genes, that are required for galactose metabolism in yeast. Gal4 is 881 amino acids long protein, containing a Zn-Cys DNA-binding domain, a linker domain, a dimerization domain and two acidic activation domains. It was demonstrated that Gal4 binds a consensus DNA sequence placed upstream of the TSS as a dimer (Marmorstein et al., 1992). In absence of galactose, Gal4 is inactive due to Gal80 repressor bound to the activation domains of Gal4 (fig. 15). Inhibition can be reversed by adding galactose to the growth medium: the Gal3 protein sequesters Gal80 thus releasing the Gal4 activation domains, that are now able to recruit coactivators and the general transcriptional machinery (Zenke et al., 1996). Gal4 was shown to interact *in vitro* with TBP (Melcher and Johnston, 1995), TFIIB (Wu et al., 1996), components of Mediator complex (Koh et al., 1998; Jeong et al., 2001; Ansari et al., 2002), SWI/SNF remodeling complex (Yudkovsky et al., 1999), SAGA and NuA4 histone acetyltransferase complexes (Brown et al., 2001). However, so far only SAGA was demonstrated to be directly targeted by Gal4 *in vivo* (Bhaumik et al., 2004). The proposed model for Gal4 activation is that SAGA is recruited by Gal4 to the promoter, that in turn recruits Mediator complex and other GTFs to form the PIC in order to initiate transcription.

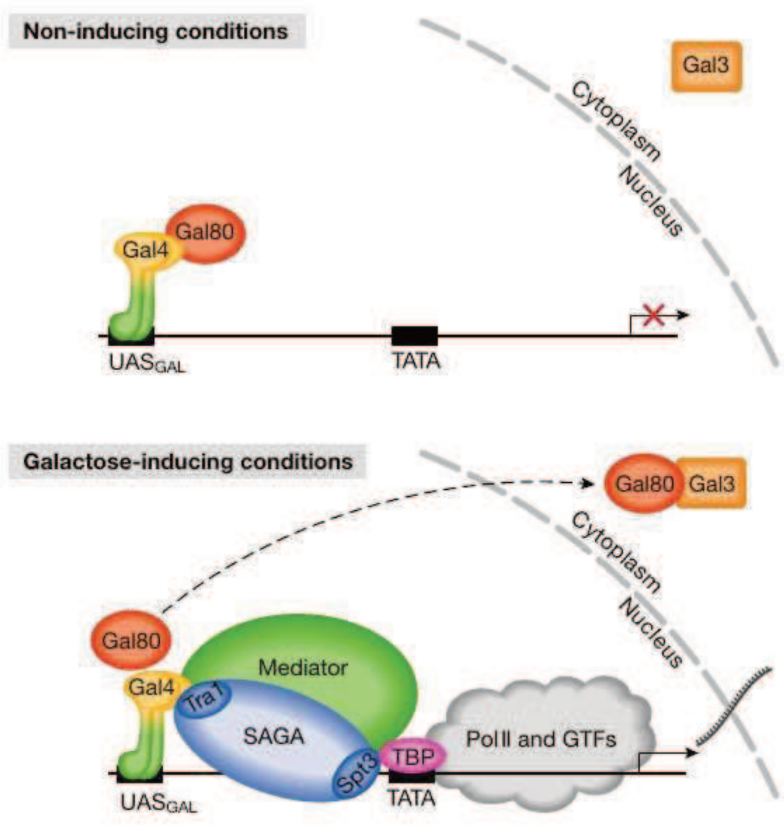

Fig. 15. Transcription activation by Gal4.

Under non-inducing conditions Gal4 activating domain is blocked by Gal80 repressor. Upon galactose addition Gal80 is sequestered by Gal3, releasing Gal4 and rendering it available for transcription machinery recruitment. UAS – upstream activation sequence. From Traven et al., 2006.

Gal4 was extensively used as a model system in the first experiments leading to the identification of yeast coactivators. A genetic screen was devised to isolate targets of the acidic activation domains present in transcriptional activators based on the toxicity of the chimeric activator, Gal4-VP16. This fusion protein contains the DNA-binding domain of Gal4 (Johnston and Dover, 1987) and a strong activation domain of the herpes simplex virus protein VP16

(Triezenberg et al., 1988). When added to yeast extract in increasing concentrations, this chimeric activator was shown to strongly stimulate transcription until all Gal4 DNA binding sites were occupied and then, for higher Gal4-VP16 concentrations, the transcription rate decreased up to the basal level. At the same time, transcription from other promoters, that did not contain the Gal4 binding sequence was dramatically inhibited by the addition of saturating amounts of Gal4-VP16. These experiments demonstrated that the mechanism underlying the strong transcriptional inhibition was the sequestering of the Gal4-VP16 target molecule – some adaptor (also called mediator or coactivator) protein, that links activators and the general transcription factors (Berger et al., 1992). Sequestration by overexpressed activator of the coactivator protein rendered it unavailable to interact with other activators present in the cell. Inhibition of transcription *in vivo* by Gal4-VP16 was exploited in genetic screens to select mutations in coactivator proteins that would overcome Gal4-VP16 toxicity and restore cell growth (fig. 16). Furthermore, other mutant hunts, originally performed in *S. cerevisiae*, resulted in the identification of numerous factors involved in chromatin-mediated transcription (reviewed in Rando and Winston, 2012). These studies have provided a basis for understanding that chromatin transcription regulation is controlled by a multitude of different factors, that regulate nucleosome assembly, dynamics and histone modifications.

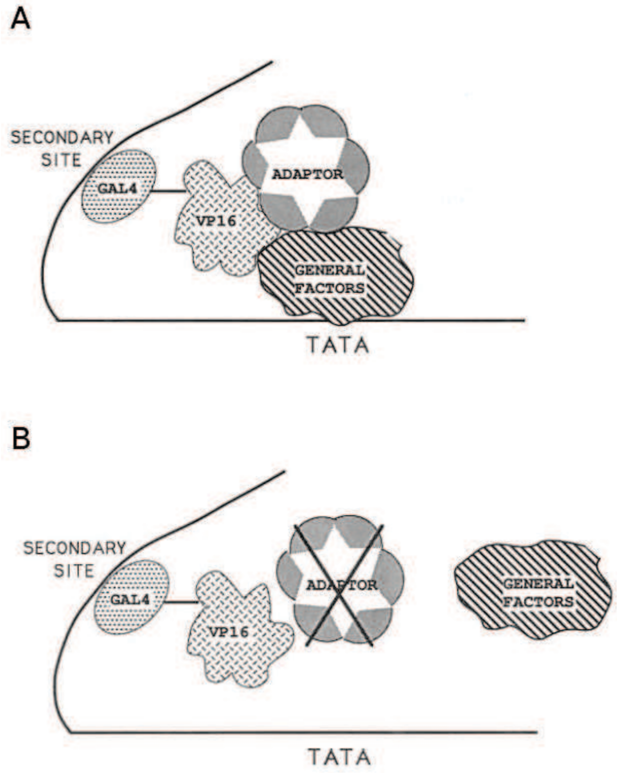

Fig. 16. Scheme of the genetic screen used to isolate mutations in transcriptional adaptors (coactivators).

(a) High level of Gal4-VP16 is toxic for the cell. VP16 recruits general transcription machinery via adaptor protein to many sites in yeast genome, sequestering general factors and therefore inhibiting the growth of yeast cells. Mutations in either Gal4 or VP16 domains prevent the assembly of these complexes and restore growth. (b) Mutations of the adaptor will also disrupt inhibition complexes and restore growth. From Berger et al., 1992.

1.3.3. Transcription regulation by chromatin structure.

Chromatin structure imposes a major obstacle on all aspects of protein transcription, mediated by the RNA Pol II. Numerous studies indicate that several types of chromatin-directed activities play an important role in facilitating the binding of activators and general transcription machinery to chromosomal DNA. Two main classes of activator-targeted coactivators can be distinguished that harbor nucleosome remodeling or modifying activities. The first class of protein complexes utilizes ATP hydrolysis to alter DNA-histone contacts. These remodeling complexes might act by sliding nucleosomes along the DNA, evicting histones off the DNA or promote the exchange of histone variants (reviewed in Narlikar et al., 2013). One of the first identified ATP-remodeling enzymes was yeast SWI/SNF complex (Côté et al., 1994) and today numerous SWI/SNF-related enzymes have been found in all eukaryotes. Recent studies suggested two possible mechanisms of ATP-dependent remodeling: the first one includes generation of positive supercoils that may lead to disruption of the histone octamer and/or displacement of the histones to different DNA locations (Owen-Hughes et al., 1996; Längst et al., 1999; Whitehouse et al., 1999), the second one is formation of altered dimeric form of a nucleosome that has altered accessibility to transcription factors (Schnitzler et al., 2001). However, the potential role of ATP hydrolysis in this process remains poorly understood.

The second class of coactivators are nucleosome modifiers that can deposit or remove post-translational modifications on histone tails of the nucleosomes. In the context of this work, we will focus here only on Histone AcetylTransferases (HATs). Over past fifty years after the discovery of histone acetylation, the study of HATs has significantly advanced as they become more amenable to biochemical and structural studies. It has been demonstrated that HATs are evolutionarily conserved, generally contain multiple subunits and that the function of their catalytic subunit is largely affected by other subunits present in those complexes. The diversity of HAT enzymes can be grouped into two large families on the basis of their catalytic domain (Table 3). Gcn5 N-AcetylTransferases (GNAT) family is built around its founding member, Gcn5, and contains over 10.000 members in all kingdoms of life including histone acetyltransferases as well as non-histone acetylatransferases (reviewed in Dyda et al., 2000). This family is characterized by four motifs comprising their HAT domain (named A, B, C, D) spanning over 160 residues, with different degree of conservation (Neuwald and Landsman, 1997). Motif A is most important for acetyl-CoA recognition and binding, has highest degree of conservation and is shared between all HAT families. In addition, GNAT members generally contain a C-terminal bromodomain, known to bind acetylated lysines.

MYST family takes origins from its members Morf, Ybf2, Sas2 and Tip60. The defining feature of this family is MYST homology domain, comprised of acetyl-CoA binding motif (a part of the motif A in GNAT family) and a Zn finger (reviewed in Avvakumov and Côté, 2007). Many MYST family members also have an N-terminal chromodomain, known to bind methylated lysine residues.

In addition, p300/CBP, Taf1 and nuclear receptor-related HATs SRC1 and SRC3 have been shown to possess intrinsic HAT activity. However, they do not have a consensus HAT domain and therefore represent an “orphan” class of HATs.

HAT family	Representative members	Complexes, containing HAT enzymes
GNAT	Hat1	HatB
	Gcn5	SAGA, SLIK, ADA, ATAC
	PCAF	SAGA, ATAC
	Elp3	Elongator
	Hpa2	
MYST	Esa1	NuA4
	MOF	MSL
	Sas2	SAS
	Sas3	NuA3
	MOZ/MORF	
	Tip60	TIP60
	Hbo1	ORC
Other	p300/CBP	
	Taf1	TFIID
	SRC1, SRC3	

Table 3. HAT classification based in their catalytic domain.

Just as HATs are a diverse set of enzymes, the multiprotein complexes in which they reside also vary. Besides regulating catalytic activity and substrate specificity of HATs, other protein subunits grouped around the catalytic enzyme can cooperate to HAT recruitment to the appropriate locations in the genome. Different chromatin-binding domains, that can recognize modified histone tails, have been characterized: bromodomains, chromodomains, WD40 repeats, PHD fingers and Tudor domains (see Lee and Workman, 2007). Intriguingly, these domains are often found in proteins that can modify or alter chromatin themselves.

Numerous transcription coactivator complexes that have been characterized was shown to play specialized roles, facilitating the action of sequence-specific transcription activators. However, a surprising structural and functional overlap exist between these complexes, often sharing individual or a whole set of subunits, indicating the modular and dynamic nature of coactivator complexes. Besides extended regulation possibilities, these properties might be essentially evolutionary constraints of the cell, expanding the versatility of apparent core complexes to adequately respond to a variety of cellular signals.

1.4. The SAGA transcriptional coactivator.

Coactivators are usually multiprotein complexes, that are recruited by activators to activate the transcription (Dynlacht et al., 1991). They include remodeling complexes, histone modifiers and complexes that bridge activators and GTFs or RNA Pol II. However, these categories often overlap due to the multifunctional nature of large coactivator assemblies.

As discussed above, the coactivator complexes can alter chromatin structure in order to provide access to the transcriptional machinery. One way of doing this is post-translational modifications of the histone tails in the nucleosomes. One of such modifications, histone acetylation, is well-known hallmark of active transcription. The first evidence showing the direct link between lysine acetylation and transcription activation was the sequencing of p55 protein from *Tetrahymena thermophila*, that had HAT activity (Brownell et al., 1996). This protein turned out to be homologous to Gcn5 – a well-studied yeast adaptor protein identified in genetic screen for mutants, that are unable to activate certain amino acid biosynthetic genes upon starvation (Penn et al., 1983; Georgakopoulos and Thireos, 1992). Gcn5 was also found required for full activity of several transcriptional activators (Gcn4, Hap, Hap3, Hap4). The biochemical characterization of Gcn5 has shown that it could acetylate free histones, but not its presumed *in vivo* substrate, the histones in the nucleosomes, suggesting that Gcn5 is included in larger complexes. The purification of HAT complexes that could acetylate nucleosomal histones revealed several Gcn5-containing complexes. One of them was shown to contain a set of Spt (Spt3, Spt7, Spt20) and Ada proteins along with Gcn5, and was termed SAGA (Spt-Ada-Gcn5 Acetyltransferase; Grant et al., 1997). In the next sections, we describe in details the components of SAGA complex and their associated functions.

1.4.1. The histone acetyltransferase (HAT) module of SAGA.

Since original proposal of Allfrey et al. in 1964, histone acetylation has long been considered correlated with activated transcription (Sealy and Chalkley, 1978; Vidali et al., 1978; Hebbes et al., 1988; Hong et al., 1993). The mechanistic link between acetylation and gene activation was directly shown by the discovery of Brownell et al., when the homology between HAT enzyme in *T. thermophila*, p55, and Gcn5, a transcriptional adaptor, was found (Brownell et al., 1996). Gcn5 and p55 were attributed to the A-type HATs, known to be located in the nucleus and acetylate chromatin-bound histones (Table 4). Recombinant yeast Gcn5 was demonstrated to acetylate free histones H3K14, H4K8 and K16 *in vitro*, but could not acetylate nucleosome core particles *in vitro* (Kuo et al., 1996). After the discovery of SAGA and ADA complexes in yeast, it was clear that Gcn5 ability to acetylate nucleosomal histones is modulated by other subunits of these complexes (Grant et al., 1997).

A-type HATs	**B-type HATs**
Transcription-related	Histone deposition-related
Nuclear	Cytoplasmic
Work on chromatin	Work on free histones
All core histones	H3 or H4 only
Distinct sites	H4K5 and K12
Post-synthetic acetylation	Synthetic acetylation

Table 4. Two general classes of HATs with respect to their intracellular location and substrate specificity.

Gcn5 is highly conserved during evolution (fig. 17) and comprised of a HAT catalytic domain and a bromodomain, that is known to recognize monoacetylated lysine residues (Dhalluin et al., 1999). Metazoan organisms possess an additional N-terminal domain PCAF-HD (PCAF homology domain), that increases total length of the protein almost twice. Interestingly, in human and mice two GCN5 isoforms exist due to alternative splicing: a short one, similar to yeast Gcn5, and a more predominantly expressed long one, that has a PCAF domain (Smith et al., 1998). The PCAF domain in mammalian GCN5 was shown to be required for acetylation of the nucleosomal substrates (Xu et al., 1998). In addition, in vertebrates *GCN5* gene was duplicated and two paralogs have been discovered: *Gcn5* and *PCAF*. PCAF (p300-CBP-associated factor) was found to be recruited to the promoters together with p300/CBP and acetylate histones H3 and H4 (Yang et al., 1996). In human, GCN5 and PCAF are 73% identical and most of the similarity comes from PCAF-HD domain. Both proteins have distinct but overlapping functions. For example, deletion of either of the two proteins can be compensated by overexpression of the other (Yamauchi et al., 2000). While PCAF is dispensable in mice, deletion of GCN5 leads to embryonic lethality; deletion of both paralogs shows even more severe defects (Xu et al., 2000).

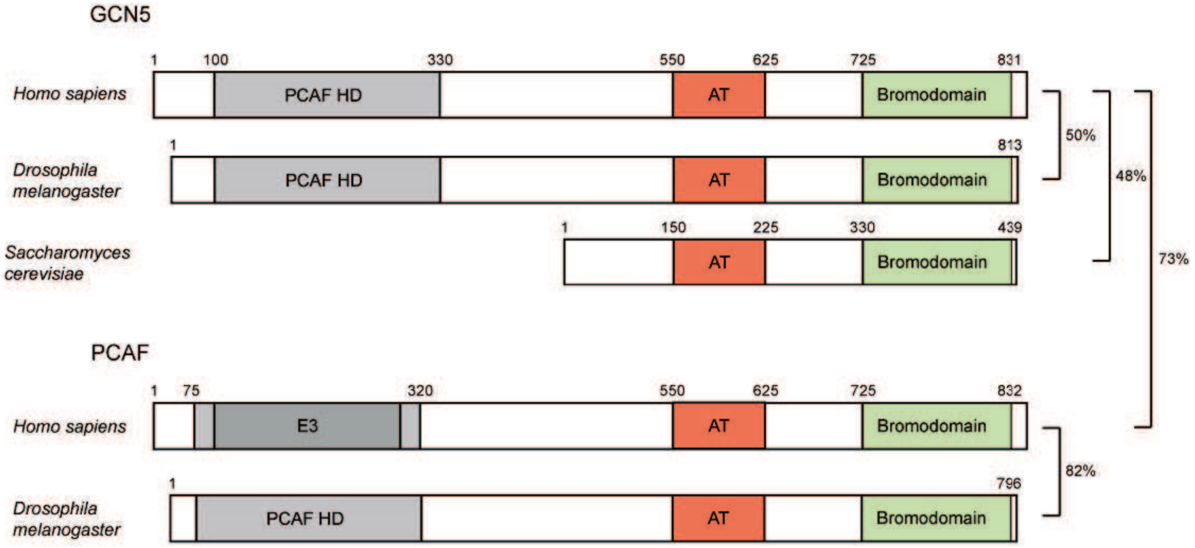

Fig. 17. Evolutionary conservation of Gcn5 and its homologs.

Different domains of Gcn5 are schematically represented: PCAF HD – PCAF homology domain specific to metazoans, AT – acetyltransferase domain, E3 – E3 ligase domain. Percents on the right side reflect protein identity. Adapted from Nagy and Tora, 2007.

Besides histones, several non-histone substrates have been identified as targets of acetylation by Gcn5 and PCAF, implicating these enzymes into different cellular processes, e.g. cell cycle and growth, regulation of DNA replication and repair, nuclear export etc. (reviewed in Glozak et al., 2005). However, it remains unclear how these activities correlate with the presence of Gcn5 / PCAF in multiprotein complexes.

Integration of Gcn5 into large complexes was shown to be necessary for acetylation of the nucleosomes. Extensive *in vitro* tests have revealed that integration of Gcn5 into coactivator complexes not only improves its ability to work on nucleosomal histones, but also changes and expands its histone specificity (Balasubramanian et al., 2002). In yeast, SAGA acetylates nucleosomal histones at H3K14, K18, K9, K23 with decreasing preference (Grant and Berger, 1999). Numerous genetic and biochemical evidence of physical interaction between Gcn5, Ada2 and Ada3 have been observed, that proposed an existence of a HAT module in SAGA (Marcus et al., 1994; Georgakopoulos et al., 1995; Horiuchi et al., 1995; Candau and Berger, 1996). The C-terminus of Ada3 was shown to interact with Ada2 *in vitro*, while Ada2 could bridge Ada3 (via C-Ada2) and Gcn5 (via N-Ada2) (Horiuchi et al., 1995; Candau et al., 1996; Candau and Berger, 1996). In addition to Gcn5, Ada2 and Ada3, a new subunit, Sgf29, was confirmed as a component of the HAT module (Bian et al., 2011; Lee et al., 2011). These inter-subunit interactions were for a long time speculated to extend Gcn5 specificity towards nucleosomal histones. Besides bromodomain of Gcn5, known to bind acetylated histones, several domains implicated in chromatin interaction have been discovered in the HAT module subunits. The SANT domain of Ada2 was shown to be indispensable for SAGA HAT activity mediated by Gcn5 and was later extensively characterized as histone tail binding module (Boyer et al., 2002; Sterner et al., 2002a). The SANT domain has a very acidic surface that explains its interactions with basic amino-terminal tails of histones. In addition, Ada2 was shown to contain a C-terminal SWIRM domain, that can bind DNA and potentiate chromatin remodeling (Qian et al., 2005). Finally, the TUDOR domain of Sgf29 binds trimethylated histones H3K4, serving as an additional chromatin anchor besides the bromodomain of Gcn5 (Vermeulen et al., 2010; Bian et al., 2011).

Up to date, a plethora of *in vitro* and *in vivo* studies have demonstrated the implication of SAGA HAT nucleosomal activity in transcription activation. However, many questions, regarding the exact mechanism of how SAGA executes its activity when recruited to the promoters and how this is coupled with other transcription factors and complexes, remains unclear. One of the questions to address is a possible functional relationship between SAGA and SWI/SNF remodeling complex in yeast. Several observations have been made in support of their interaction: (i) SAGA recruitment by several activators has been shown to stimulate chromatin remodeling (Gregory et al., 1999; Wallberg et al., 2000); (ii) the bromodomain of Gcn5 was found dispensable *in vivo* for SAGA HAT activity, however, it was necessary for subsequent recruitment of chromatin remodellers (Syntichaki et al., 2000; Hassan et al., 2001, 2002); (iii) mutations that disrupted both complexes had more severe phenotypes than single mutations (Pollard and Peterson, 1997; Roberts and Winston, 1997). However, the current data cannot distinguish whether SAGA and SWI/SNF are recruited sequentially or simultaneously to the promoters.

1.4.2. The deubiquitination module of SAGA.

Initially, the only enzymatic activity assigned to the SAGA complex was histone acetylation. Another function was unveiled by discovery of a deUbiquitinating (dUb) enzyme as a component of SAGA. Ubp8 was first found in yeast mutant screen for deleted dUb genes (Amerik et al., 2000) and later was attributed to the SAGA complex, together with the newly discovered proteins Sgf73 and Sgf29 (Gavin et al., 2002; Ho et al., 2002; Sanders et al., 2002).

In the context of SAGA, Ubp8 was shown to efficiently deubiquitinate histone H2B *in vitro* (Henry et al., 2003; Daniel et al., 2004). Initial studies of Ubp8 activity have discovered that a dynamic balance of H2B ubiquitination/deubiquitination is necessary for optimal transcription of several yeast genes, including *Suc2*, *Gal10* and *Gal1* (Daniel et al., 2004; Wyce et al., 2004). Furthermore, such balance influences the methylation status of certain H3 residues, exposing a new cross-talk between different post-translational modifications of histone tails (reviewed in Shilatifard, 2006). In the meantime, new SAGA subunits have been discovered: Sgf11 (SAGA-associated Factor 11 kDa, Powell et al., 2004), that was shown to be required for both Ubp8 association with SAGA and H2B deubiquitination (Ingvarsdottir et al., 2005; Lee et al., 2005); Sus1 – an mRNA export factor that binds to both Ubp8 and Sgf11, forming a trimeric dUb submodule of SAGA (Köhler et al., 2006). Another SAGA protein, Sgf73 and its human homolog ATXN7, have been identified (Sanders et al., 2002; Helmlinger et al., 2004) as a key components of the dUb module, linking it to the SAGA complex (Köhler et al., 2008).

The role of histone ubiquitination has been extensively studied (reviewed in Weake and Workman, 2008) and it has become clear that ubiquitination is implicated in the regulation of many cellular processes, including transcription initiation and elongation, silencing and DNA repair. Usually, the protein substrates can be mono- or polyubiquitinated. While polyubiquitination targets proteins for degradation by proteasome, monoubiquitin marks the substrate protein to signal for a particular function. Consistent with this signaling role, the monoubiquitin tag is reversible and can be removed by specific proteases: Ubp in yeast and USP in human. H2B ubiquitination has been implicated into cross-talk with H3 methylation and a model has been proposed to reflect this observation (Weake and Workman, 2008). In this model, the H2B ubiquitinating enzymes, Rad6 and Bre1 are recruited to promoters by activators and then associate with RNA Pol II to start the transcription. However, this recruitment is not sufficient for H2B ubiquitination, and additional complexes are required as well as the elongating form of RNA Pol II. Monoubiquitination of H2B at K123 is prerequisite for H3K4 and H3K79 di- and trimethylation in yeast and higher eukaryotes, and is mediated by the COMPASS complex in yeast (MLL complex in human). In the next step, the SAGA complex is recruited and Ubp8 deubiquitinates H2B, triggering a phosphorylation of the C-terminal domain of the RNA Pol II by Ctk1 kinase (Wyce et al., 2007), that in turn will trigger subsequent steps of transcription elongation. Thus, the SAGA dUb activity is required for optimal transcription activation and might link transcription initiation and elongation.

More recently, the deubiquitinating enzyme of SAGA was found to work on the substrates other than histones. Ubp8 in *S. cerevisiae* was shown to deubiquitinate Snf1, a highly conserved AMP-activated serine/threonine protein kinase, that serves as an energy sensor in a cell (Wilson et al., 2011). Snf1 is known to phosphorylate different substrates, including histones and transcription factors. Thus, SAGA tightly regulates Snf1 activity by deubiquitination via Ubp8. Another target of USP22, a human homolog of Ubp8, is TRF1 – a component of shelterin complex, important for telomere maintenance (Atanassov et al., 2009). Depletion of USP22 was shown to increase ubiquitination and turnover of TRF1, indicating SAGA importance for post-transcriptional maintenance of TRF1. In addition, Ubp8 was demonstrated to antagonize Anaphase-Promoting Complex/Cyclosome (APC/C) function indirectly by modulating H2B ubiquitination status in *S. pombe* (Elmore et al., 2014). Altogether, these results suggest that SAGA utilizes multiple mechanisms to fine-tune gene expression in addition to altering chromatin structure via HAT activity.

Sgf11 subunit of the dUb module in *Drosophila* was shown to have other activities besides deubiquitination. In particularly, it is involved in mRNA export through interaction with the CBC and AMEX complexes (Gurskiy et al., 2012). That study has provided a first connection between SAGA-dependent transcription and nuclear export of mRNA in yeast via the dUb module of SAGA. Interestingly, another dUb module protein was found to provide a link between SAGA and nuclear export machinery – Sus1 (Rodríguez-Navarro et al., 2004). Sus1 was found associated with both SAGA and TREX2 complex, involved in mRNA export and transcription elongation (Pascual-García and Rodríguez-Navarro, 2009). In addition, other SAGA subunits were found important for mRNA export: Sgf11 removal was demonstrated to enhance mRNA export defects imposed by *sus1*-knockout (Köhler et al., 2006), while loss of Sgf73 partially disrupted Sus1-TREX2 interaction (Köhler et al., 2008). Recently, a specific interaction between proteasomal ATPase and Sgf73 was described, that led to dissociation of whole dUb module from SAGA complex *in vitro* and *in vivo* (Lim et al., 2013). The dissociation did not affect dUb activity and was shown to be important for targeting the TREX2 complex to the Nuclear Pore Complex (NPC) by dUb, promoting mRNA export.

Among other complexes found to interact with SAGA, the connection with the 19S proteasome Regulatory Particle (19S RP) looks intriguing. The ATPase components of RP were shown to stimulate SAGA-activator interactions and SAGA recruitment to the promoters (Lee et al., 2005). Another evidence points to the recruitment of the 19S RP to chromatin, that depends on H2B ubiquitination (Ezhkova and Tansey, 2004). This indicates that both complexes function subsequent to ubiquitination. Recently, the 19S RP was also shown to facilitate SAGA loading onto chromatin (Lee et al., 2005). More recently, Garcia-Oliver et al. have proposed a model showing SAGA implication in both transcription and mRNA export (García-Oliver et al., 2012). In this model, SAGA is recruited to chromatin upon transcription activation, while 19S RP stimulates SAGA interaction with activators. SAGA HAT activity relaxes the chromatin, allowing the recruitment of general transcriptional machinery that starts transcription. SAGA and messenger RiboNucleoprotein Particle (mRNP) interaction with NPC-bound TREX-2 complex might promote gene looping at the NPC, used for more rapid transcription initiation, and facilitate smooth transition from transcription elongation to mRNA export.

In addition, possible SAGA involvement in elongation was demonstrated on several SAGA mutants, that were shown to genetically interact with elongation complexes, such as TFIIS, Elongator and RNA Pol II subunit Rbp9 (Van Mullem et al., 2002; Wery et al., 2004).

SAGA complex was also found to play a role in co-transcriptional pre-mRNA splicing. Genetic analysis has demonstrated that the SAGA complex has functional interactions with the splicing machinery, since SAGA-dependent acetylation was found necessary for recruitment of U2 snRNP (Gunderson and Johnson, 2009; Merkhofer and Johnson, 2014). In addition, human SAGA was found to associate *in vivo* in HeLa cells with Spliceosome-Associated Protein 130 (SAP130) and with UV-damaged-DNA-binding factors, suggesting the possibility of additional functions for SAGA in transcription-coupled pre-mRNA splicing and DNA damage repair *in vivo* (Martinez et al., 2001).

In summary, several gene expression stages from transcription activation to mRNA processing and export may be influenced by SAGA at least for some genes, although it remains to be determined how universal this connectivity is.

Polyglutamine (polyQ) expansions within N-terminus of ATXN7 (mammalian homolog of yeast Sgf73) have been shown to cause SpinoCerebellar Ataxia type 7 (SCA7), a disease associated with progressive neurodegeneration. Normally, ATXN7 contains from 4 to 35 glutamine residues, but in SCA7 patients, this number can rise from 36 up to few hundreds (Todd and Lim, 2013). Recent studies showed no effect of polyQ expansions in ATXN7 on dUb integrity or activity *in vitro*, however, *in vivo* these expansions promoted ATXN7 aggregation, leading to sequestering of USP22 activity and subsequent increased global levels of H2Bub, most likely contributing to SCA7 development and progression (Lan et al., 2015). The effects of polyQ expansions on SAGA HAT activity or on SAGA-mediated gene activation are not clear, since contradicting results were obtained on this topic: while several groups observed disruption of HAT activity upon integration of PolyQ-expanded ATXN7 in SAGA (McMahon et al., 2005;

Palhan et al., 2005; Burke et al., 2013), other studies argue against this (Helmlinger et al., 2006).

In addition, SAGA has been shown implicated in the preservation of genomic integrity beside its role in genome expression. For example, UV damage has been shown to provoke H3K9 and K14 acetylation, facilitating Nucleotide Exchange Repair (NER) (Brand et al., 2001; Yu et al., 2005; Ferreiro et al., 2006). Weake et al. have shown that Nonstop, a homolog of Ubp8 in SAGA complex in *Drosophila*, is required for neural development, particularly in the accuracy of axon guidance in the optic lobe (Weake et al., 2008). USP22, a human homolog of Ubp8, was identified as a putative cancer stem cell marker (Zhang et al., 2008). Other studies have identified a possible role of SAGA in cancer development: Sgf29 subunit of the HAT module was found overexpressed in hepatocellular carcinoma (Kurabe et al., 2007); human TRRAP and GCN5 were found to be recruited by c-MYC, a known oncoprotein, whose deregulation is implicated in many types of cancer (Liu et al., 2003); cytoplasmic localization of ADA3, that was found to directly interact with estrogen receptors, has predicted poor prognosis for breast cancer patients (Mirza et al., 2013). At the same time, mammalian ADA3 was found essential for embryonic development and cell cycle progression (Mohibi et al., 2012).

1.4.3. TAF proteins are required for SAGA structural integrity.

Quickly after discovery of yeast SAGA complex, a set of TAF subunits have been characterized as a part of SAGA (Grant et al., 1998a). Originally, TAF subunits were found to be associated with TBP within TFIID general transcription factor (Dynlacht et al., 1991). Sequence analysis of TAF subunits has shown that they contain a Histone Fold Domain (HFD) which is present in core histones and is responsible for histone octamer formation (Xie et al., 1996). TAF proteins have displayed strong sequence conservation between the species ranging from yeast to human (Sanders and Weil, 2000; Walker et al., 2001). Subsequently, TAF subunits were shown to be important for assembly and structural integrity of TFIID as well as interaction with DNA, activators and histone modifications (reviewed in Cler et al., 2009). While TFIID contains 13-14 different TAFs, SAGA has only five: Taf5, Taf6, Taf9, Taf10, Taf12. Beside structural integrity (Michel et al., 1998; Durso et al., 2001; Kirschner et al., 2002), some of the TAFs seems to play an important role in ability of SAGA to stimulate transcription. For example, Taf12 was found to be required for both SAGA-dependent nucleosomal HAT activity (Grant et al., 1998a) and transcriptional activation from chromatin templates *in vitro* (Grant et al., 1998a; Natarajan et al., 1998). In TFIID, Taf6/Taf9 and Taf4/Taf12 form a stable octameric complex composed of two heterodimers (Selleck et al., 2001). SAGA does not have Taf4 and it was demonstrated that Taf12 forms a dimer with Ada1 instead (Gangloff et al., 2000). The same authors have shown that Spt7 and Taf10 dimerize through their HFD (Gangloff et al., 2001). The formation of Taf10/Spt7 and Taf12/Ada1 heterodimers in SAGA is also supported by cross-linking mass spectrometry studies (Han et al., 2014). In summary, all current results suggest that TAF subunits, similarly to TFIID, compose a structural core of SAGA, that provides integrity of the whole complex. The molecular weight of SAGA complex was estimated by gel filtration at 1.8-2 MDa, while the total weight of all subunits is only 1.4 MDa. This fact allows to speculate that several subunits might be indeed present in more than one copy in SAGA. However, more investigation is required to determine exact stoichiometry of the SAGA complex.

Besides sharing a common set of TAFs, both TFIID and SAGA possess a TBP binding activity and play major roles in transcription regulation. Deletion of shared TAF subunits is known to result in loss of expression of 70% genes in yeast (Lee et al., 2000). SAGA was found to predominate the transcription of 10% yeast genes, most of which are stress-induced genes (Huisinga and Pugh, 2004). While TFIID was shown to bind to core promoters, SAGA bound Upstream Activation Sequences (UAS) and recruited basal transcription machinery (Larschan and Winston, 2001; Bhaumik and Green, 2001, 2002). However, more recent genome-wide studies revealed that SAGA has only few hundred binding sites in different organisms (Krebs et al., 2011; Venters et al., 2011; Weake et al., 2011). Furthermore, besides specific gene regulation, it has been proposed that SAGA acts globally, acetylating histones throughout the genome in an untargeted manner (Vogelauer et al., 2000). Recently, a genome-wide analysis of H3K9 acetylation and H2B deubiquitination, revealed that SAGA acts on the whole transcribed genome in yeast and human cells, acetylating promoters and deubiquitinating transcribed regions of all expressed genes, and is involved in RNA Pol II transcription (Bonnet et al., 2014).

1.4.4. Spt proteins are necessary for SAGA structural integrity and interaction with TBP.

Numerous transcription factors were identified by selecting for suppressors of promoter mutations, caused by Ty retrotransposons in yeast. Upon insertion, these elements can exert strong opposite stimulatory or inhibitory effects on the transcription of adjacent genes, depending the position and orientation of insertion, as well as on the sequence of specific Ty element. Genes that could suppress these mutations have been identified and named *Spt*

(SuPpressors of Ty). Spt mutants were able to reverse transcription defects caused by Ty insertion (Winston et al., 1984). Interestingly, one of the *Spt* genes (*Spt15*) was shown to encode TBP (Eisenmann et al., 1989) and at the same time, another group of *Sp*t genes encodes histones (Clark-Adams et al., 1988).

Yeast SAGA is known to contain 4 different proteins of the Spt family: Spt3, Spt7, Spt8 and Spt20. Spt3 contains two histone fold domains connected by a 120-residue linker domain: the domain at the N-terminus is homologous to Taf13 (Mengus et al., 1995), while C-terminal domain is homologous to Taf11 (Birck et al., 1998). Based on this observation, it was proposed that Spt3 is likely to adopt histone fold structure that can interact with TBP similarly to the Taf11/Taf13 heterodimer in TFIID.

Genetic and biochemical analyses have indicated that Spt3, Spt7, Spt8 and Spt20 help SAGA-TBP interactions and might assist TBP recruitment to the promoters. Identification of Spt8 subunit as a SAGA component evidenced that SAGA might be important for TBP function at certain promoters, since Spt8Δ SAGA mutants failed to bind TBP *in vitro*. Certain mutations in Spt3, Spt7, Spt8, Spt15 (TBP) and Spt20 (Ada5) all caused common mutant phenotypes displaying defective transcription from certain promoters (Winston et al., 1984, 1987; Hirschhorn and Winston, 1988; Eisenmann et al., 1989, 1994; Gansheroff et al., 1995; Roberts and Winston, 1996). Other group of studies has demonstrated direct interaction between Spt3, Spt8 and TBP *in vitro* and *in vivo* (Eisenmann et al., 1992, 1994; Warfield et al., 2004; Sermwittayawong and Tan, 2006; Laprade et al., 2007; Mohibullah and Hahn, 2008). Binding sites of Spt3 and Spt8 on TBP were mapped in different parts of TBP: while Spt3 binds residues 176-196, Spt8 binds to residue 79 (Mohibullah and Hahn, 2008). Spt8 was shown to compete with TFIIA for TBP binding *in vitro* (Warfield et al., 2004) and at the same time to compete with DNA for TBP binding (Sermwittayawong and Tan, 2006). Spt3 is unable to bind TBP alone, and cross-linking experiments have shown that Spt7 and Spt20 might be involved in these interactions: Spt7 cross-links both TBP and Spt8, while Spt20 cross-links TBP in a nearby area (Han et al., 2014). Altogether these results indicate a complex network of interaction of Spt3/Spt7/Spt8/Spt20 subunits and TBP (fig. 18).

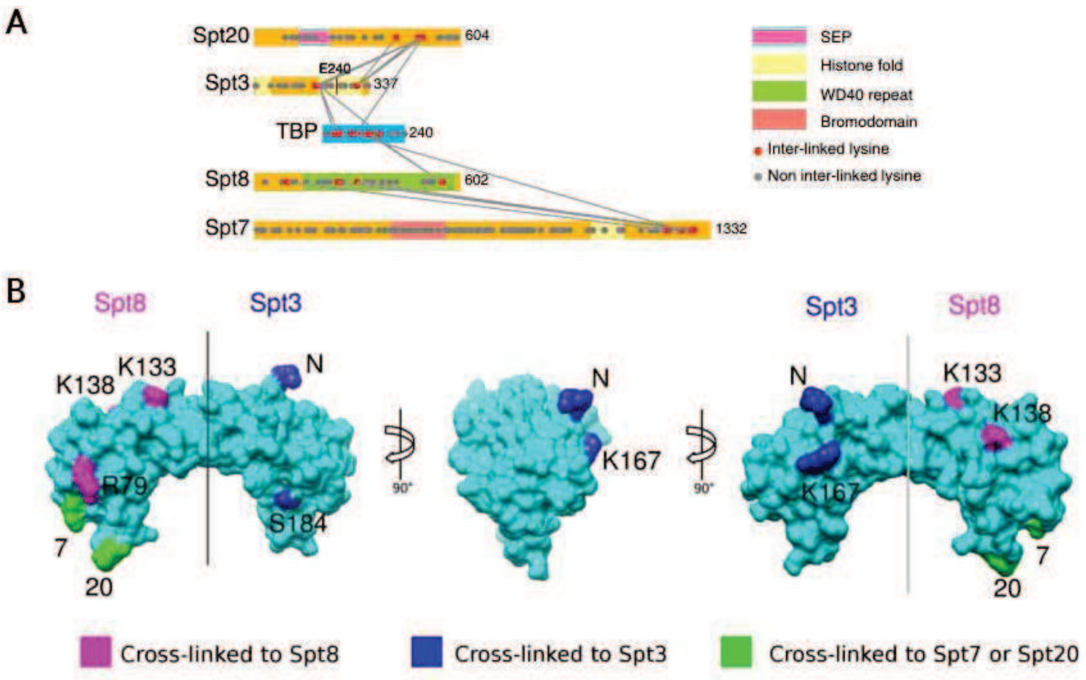

Fig. 18. Two distinct surfaces on TBP bind Spt3 and Spt8.

(a) A cross-linking map between TBP and Spt SAGA subunits, based on MS analysis. (b) Surface representation of TBP (PDB 1TBP). The residues that cross-link to Spt3 and Spt8 are colored in blue and purple, respectively. Residues cross-linked to Spt7 and Spt20 are colored in green. Adapted from Han et al., 2014.

Spt7, Spt20 and Ada1 subunits of SAGA were found essential for complex integrity (Grant et al., 1997; Sterner et al., 1999; Wu and Winston, 2002; Nagy et al., 2009). Both Spt7 and Ada1 contain HFD and can form heterodimers *in vitro* with Taf10 and Taf12, respectively (see before). In addition, the C-terminus of Spt7 was shown to bind Spt8 and to control the protein levels of Spt20 and Ada1 (in a *spt7*Δ mutant, both Spt20 and Ada1 proteins are barely detectable), thereby suggesting Spt7 structural role in formation of the SAGA complex (Wu and Winston, 2002; Han et al., 2014). The Spt7 protein also contains bromodomain that is known to bind acetylated histone tails, however, it fails to do so in the context of SAGA complex (Hassan et al., 2002). Deletion of this domain did not affect binding, targeting, or the retention of the complex on nucleosome arrays in either the presence or absence of acetylation (Gansheroff et al., 1995; Hassan et al., 2002).

1.4.5. Tra1 subunit is important for SAGA assembly and binding of transcriptional activators.

Ada (from "Adaptor") protein family was identified in the genetic screens for selection of the mutants (Berger et al., 1992), resistant to Gal4-VP16-mediated toxicity (see part 1.3.2). These proteins have been later shown to play a role of coactivators, bridging activators with the transcriptional machinery. After realizing that several Ada proteins (Ada1, Ada2, Ada3, Ada4=Gcn5, Ada5=Spt20) are components of large multiprotein complexes like SAGA, it was not clear how exactly SAGA interacts with activators. Shortly after the discovery of the SAGA complex, it was shown to bind VP16 activation domain and acetylate Gal4-VP16-bound nucleosomes (Utley et al., 1998; Ikeda et al., 1999; Vignali et al., 2000). Later, the interaction interface between SAGA and activators was localized on the Tra1 subunit (Brown et al., 2001). The Tra1 protein was further demonstrated to bind the activation domains of VP16, Gal4, Hap4, Gcn4 (Brown et al., 2001) *in vitro* and Gal4 *in vivo* (Bhaumik et al., 2004).

Tra1 is a large 433 kDa protein that was found to be a stable component of SAGA complex (Grant et al., 1998b). Tra1 was identified as a homolog of human TRRAP protein, a co-factor required for c-MYC and E1A/E2F-mediated oncogenic transformation in mammalian cells (McMahon et al., 1998; Kulesza et al., 2002). It was shown that Tra1 is also present into NuA4 HAT coactivator complex (Allard et al., 1999). Among the 13 subunits of NuA4, only Tra1 is shared with SAGA. The common Tra1 subunit can directly recruit either of these complexes to the promoters (Nourani et al., 2004; Govind et al., 2007; Pascual-García et al., 2008; Ginsburg et al., 2009). Recently, deletion analysis of Tra1 has identified the regions important for cell viability, integrity of SAGA and NuA4 complexes, transcription activation, activator interaction, coactivator recruitment, and *in vivo* HAT activity (Knutson and Hahn, 2011). Most of the Tra1 protein sequence contains short helical repeats organized into HEAT and FAT domains, comprising almost 85% of total protein length. The major part of Tra1 forms a scaffold for binding other SAGA / NuA4 subunits and is important for complex integrity. Gcn4 and Rap1 activators were shown to target mostly the N-terminal region of HEAT repeat domain, but it cannot be excluded that multiple distinct activator binding sites exist on Tra1, since none of analyzed mutations was fully defective for activator binding.

In vivo studies of Tra1 and TRRAP are challenging since in *S. cerevisiae* Tra1Δ mutants are lethal and in mice TRRAPΔ mutants also die early during embryonic development. However, in *S. pombe* Tra1 deletion was found nonessential for viability since Tra exists in two isoforms, Tra1 and Tra2, interacting specifically with SAGA and NuA4, respectively (Calonge et al., 2010). A recent study of Tra1 deletion mutants in this yeast has proposed that Tra1 plays a regulatory role in SAGA, rather than structural (Helmlinger et al., 2011). The same authors also demonstrated, that Tra1 is not absolutely required for the recruitment of SAGA to promoters *in vivo* indicating that SAGA can associate with promoters in a Tra1-independent manner. Previous studies have identified other SAGA subunits as potential targets contributing to the recruitment of SAGA at promoters, including Spt3, Taf6, Taf12 or Ada1 (Klein et al., 2003; Fishburn et al., 2005; Reeves and Hahn, 2005). Further investigations are necessary to determine which SAGA subunits are required for recruitment to activators and whether these interactions are gene-specific.

1.4.6. SAGA evolutionary conservation.

Shortly after the discovery of the HAT activity of Gcn5 numerous genetic and biochemical studies were carried out and concluded that Gcn5 is a part of larger multiprotein complexes. Homologs of Gcn5 HAT enzyme were found in ciliates (Brownell et al., 1996), flies (Smith et al., 1998), human (Candau et al., 1996), and they turned out to be highly conserved throughout evolution. In the following years a large number of different HAT complexes has been described in different species and, at least partially, characterized (reviewed in Brown et al., 2000; Carrozza et al., 2003; Lee and Workman, 2007; Nagy and Tora, 2007). The SAGA complex, first discovered in yeast, is not an exception. Recent evidence indicates that the SAGA complex has diverged into several distinct complexes in metazoan (Spedale et al., 2012). Several SAGA-like complexes have been described, including human SPT3-TAF9L-GCN5-L Acetyltransferase (STAGA; Martinez et al., 1998), TBP-Free TAF-containing Complex (TFTC; Wieczorek et al., 1998), PCAF (Ogryzko et al., 1998), as well as ATAC (Ada2A-Containing) complexes in *Drosophila* (Guelman et al., 2006) and human (Wang et al., 2008). Eventually, when all 19 subunits of yeast SAGA complex were characterized, their homologs were found also in vertebrates (Helmlinger et al., 2004; Palhan et al., 2005; Kurabe et al., 2007; Zhang et al., 2008; Zhao et al., 2008; Kurshakova et al., 2009; Nagy et al., 2009). Finally, it was realized that human TFTC/STAGA/PCAF all refer to the same complex, and small differences in subunit composition can be explained by differences in purification protocol (Pijnappel and Timmers, 2008; Nagy et al., 2009). Thus, it was proposed to name them all as human SAGA (hSAGA). ATAC complexes, on the contrary, seem to be diverged from SAGA during evolution, since they share only HAT core with SAGA, while the other subunits are different.

***S. cerevisiae* (ySAGA)**	***D. melanogaster* (dSAGA)**	***H. sapiens* (hSAGA)**	**Functions**
yGcn5	dGcn5	hGCN5 (KAT2A) / hPCAF (KAT2B)	Histone acetyltransferase module
yAda2	dAda2b	hTADA2B	
yAda3	dAda3 (CG7098)	hTADA3	
ySgf29	dSgf29 (CG30390)	hSGF29 (CCDC101)	
ySpt7	dSpt7 (CG6506)	hSPT7 (SUPT7L)	Structural integrity
ySpt20 (yAda5)	dSpt20 (CG17689)	hSPT20 (p38IP)	
yAda1	dAda1 (CG31866)	hTADA1	
yTaf5	dWda	hTAF5L	
yTaf6	dSaf6	hTAF6L	
yTaf9	dTaf9	hTAF9	
yTaf10	dTaf10b	hTAF10	
yTaf12	dTaf12	hTAF12	
yTra1	dTra1	hTRRAP	Activator binding
ySpt3	dSpt3 (CG3169)	hSPT3 (SUPT3H)	TBP interaction
ySpt8	-	-	
yUbp8	dNonstop	hUSP22	Deubiquitination module
ySgf11	dSgf11	hATXNL3	
ySgf73	dCG9866	hATXN7	
ySus1	dE(y)2	hENY2	

Table 5. List of known SAGA subunits and their homologs in eukaryotes.

In yeast, all 19 subunits of SAGA complex can be divided in several groups: Spt proteins, TAF proteins, HAT and dUb modules and Tra1 subunit, responsible for activator interaction (Table 5). Most domains of SAGA subunits are highly conserved throughout evolution from yeast to mammals (Spedale et al., 2012). However, there are several exceptions. One of them is absence of human ortholog of Spt8 gene. In addition, the C-terminus of Spt7 was found truncated in all human SAGA variants. Similarly, a version of yeast SAGA, named SLIK or SALSA, was found (Belotserkovskaya et al., 2000; Pray-Grant et al., 2002; Sterner et al., 2002b). The SLIK (SAGA-like) complex lacks Spt8 and has a truncated Spt7 protein

(Mischerikow et al., 2009; Spedale et al., 2010), thus resembling homologous complexes in higher eukaryotes. The absence of Spt8 most likely affects TBP binding, since in the context of the complete SAGA complex Spt8 is responsible for TBP interaction together with Spt3 (Sermwittayawong and Tan, 2006; Laprade et al., 2007).

During evolution, many yeast genes have undergone duplications, as in the case of *Gcn5*, *Ada2*, and several TAFs (*Taf5*, *Taf6*, *Taf9*, *Taf10*). In yeast all these TAFs are unique and are incorporated into TFIID or SAGA. In human, *TAF5* and *TAF6* genes have been duplicated and diverged into TAF5L and TAF6L. While Taf5 and Taf6 are part of TFIID, TAF5L and TAF6L are found only in hSAGA. Similarly, Taf9 has diverged into Taf9 and Taf9B, but both of them can incorporate into either TFIID or SAGA. Taf10 has been duplicated in flies into Taf10 and Taf10B, specific to TFIID and SAGA, respectively. Gcn5 exists only in one form in yeast, while in higher eukaryotes an additional larger homolog (GCN5) has been found, containing N-terminal PCAF homology domain. In vertebrates, the divergence went further and a separate *PCAF* gene appeared, that gave rise to the PCAF protein. Both PCAF and GCN5 are highly similar (73% identity; Yang et al., 1996) and have the PCAF homology domain at the N-terminus. Interestingly, both PCAF and GCN5 are present in SAGA complexes in mutually exclusive manner (Krebs et al., 2010). In human, PCAF and GCN5 can be incorporated into either SAGA or ATAC complexes. *ADA2* gene has been duplicated in flies and mammals, giving two proteins ADA2A and ADA2B, that are included into ATAC (Ada2a-containing) or SAGA complex, exclusively.

In summary, while yeast have one SAGA complex, in flies SAGA and ATAC exist and their HAT module differs only by Ada2 subunit (Ada2a or Ada2b); in mammals, the number of complexes multiplied and reached four: two types of ATAC (with GCN5 or PCAF) and two types of SAGA (with GCN5 or PCAF). Besides different subunit composition of SAGA and ATAC complexes (they share only GCN5/PCAF, ADA3 and SGF29 subunits), they seem to display functional differences. For example, *in vitro* studies have shown that while SAGA specifically acetylates H3 histones on nucleosomal substrates, ATAC is more specific towards H4 ((Suganuma et al., 2008). Both complexes have shown to be involved in different stress-induced signaling pathways (reviewed in Spedale et al., 2012), as well as regulation of housekeeping and tissue-specific genes (Krebs et al., 2011).

1.4.7. Structure of SAGA complex.

Shortly after the discovery of SAGA complex, it was realized that SAGA contains not only subunits related to HAT activity, but also many other proteins that make the complex a truly multifunctional coactivator. One of the first attempts to understand the molecular structure of the complete SAGA was the electron microscopy study of yeast complex, performed in our group (Wu et al., 2004). The first images revealed the elongated shape of the complex (fig. 19), comprised of several domains, linked together in an almost linear arrangement and forming two pronounced clefts. The overall shape of the complex was found similar to the structure of the human TFTC complex (now considered as hSAGA), also determined by EM (Brand et al., 1999). In the same study, the authors have performed immunolabeling experiments to localize 9 out of 19 subunits of the complex. Taf5, Taf6 and Taf10 subunits were localized in the central part of the complex (domains II, III and IV), as well as Spt7, Spt20 and Ada1, altogether supporting the idea of their architectural role in the complex. Domains I and V have been shown to comprise Tra1 and Spt3, respectively, displaying that activator-binding and TBP-binding activities are spatially separated in the complex. Gcn5, the key HAT subunit, was located in domain III. At that time, the presence of the dUb activity in SAGA was not yet discovered. Altogether, the first structure of SAGA complex suggested that its components are organized into distinct structural modules, reflecting their distinct functional roles.

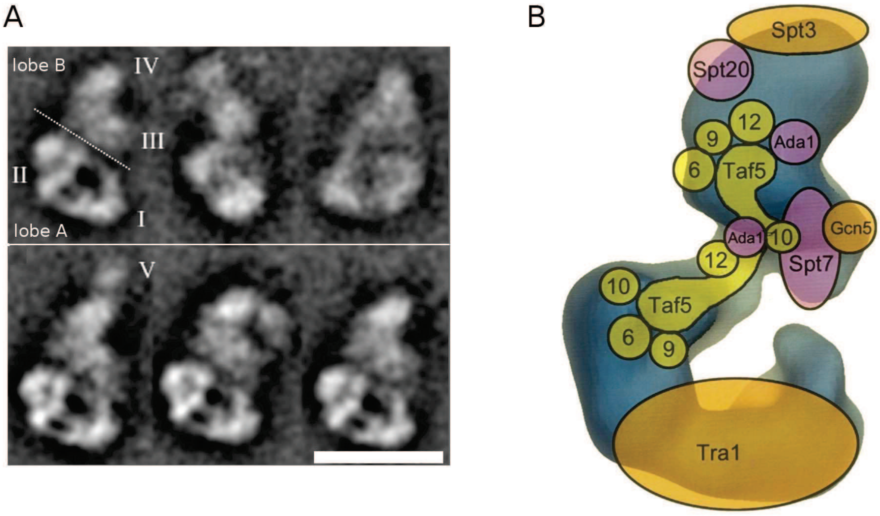

Fig. 19. First electron microscopy structure of S. cerevisiae SAGA complex.

(a) Representative SAGA views, showing elongated structure composed of two lobes. Upper panel shows different views, lower panel reveals a domain V, that can adopt different positions. Scale bar is 20 nm. (b) Immunolabeling of different subunits on a schematic SAGA model. Adapted from Wu et al., 2004.

Later, when the dUb module has been found and characterized in SAGA, it was crystallized simultaneously by two groups (Köhler et al., 2010; Samara et al., 2010). The structure comprising Ubp8, Sus1, Sgf11 and N-terminal (1-104) of Sgf73 revealed two distinct functional lobes of the module, each of them organized by Ubp8 (fig. 20). N-terminal domain of Ubp8 forms an "assembly lobe", nucleating other three subunits, while C-terminal "catalytic lobe" comprises active site of Ubp8 and C-terminus of Sgf11. The two functional lobes of the module are joined together by the Sgf73 N-terminal fragment, consistent with Sgf73 role in anchoring the dUb module to SAGA.

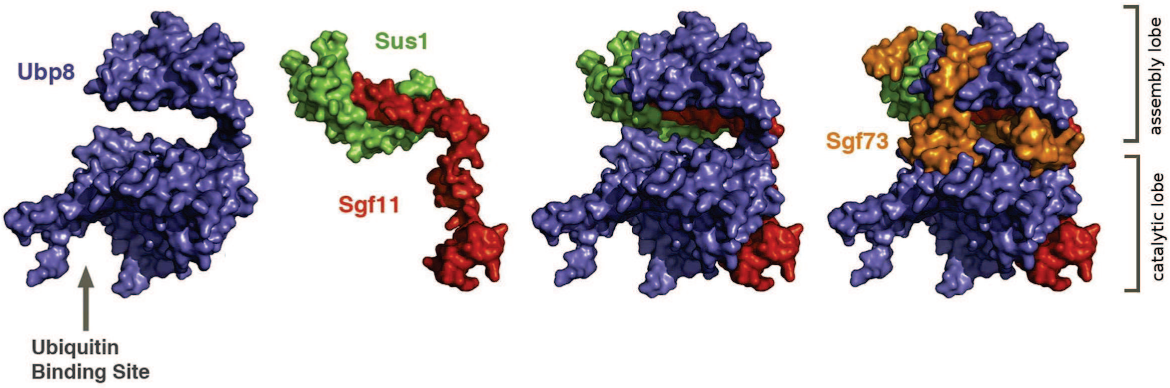

Fig. 20. Surface representation of different dUb subunits showing their spatial relationship. *From Kohler et al., 2010.*

More recent proteomic study, based on combinatorial depletion analysis has confirmed the modular organization of the SAGA complex (fig. 21; Lee et al., 2011). In this study, four distinct modules have been revealed: the TAF module (Taf5, Taf6, Taf9, Taf10, Taf12), Spt module (Spt3, Spt7, Spt8, Spt20 together with Tra1 and Ada1), dUb module (Ubp8, Sgf11, Sgf73, Sus1) and HAT module (Gcn5, Ada2, Ada3, Sgf29). The authors performed extensive deletion mutant analysis, that led to several conclusions: (i) Ada2 is critical for HAT module formation and links the module to the SAGA complex; (ii) all TAFs are grouped together, indicating that they are less likely dimerized with Spt or Ada proteins via HFD; (iii) Sgf73 is indeed the anchor between the dUb module and SAGA and interacts with Spt20; (iv) deletion of Spt20 leads to disintegration of the full complex, however HAT and dUb modules seem to remain intact, suggesting the existence of stable SAGA subcomplexes. The proposed interaction map, based on this analysis, shows significant differences from previous EM-based study, the most striking differences are: (i) position of Spt20 with respect to Tra1 – EM study has located these subunits in completely different parts of the complex, while proteomic approach suggested interaction between these two proteins; (ii) the EM structure has shown proximal positions of Spt7 and Gcn5, while it was not the case in the proteomic map. The TAF subunits, however, in both studies were grouped in a single module.

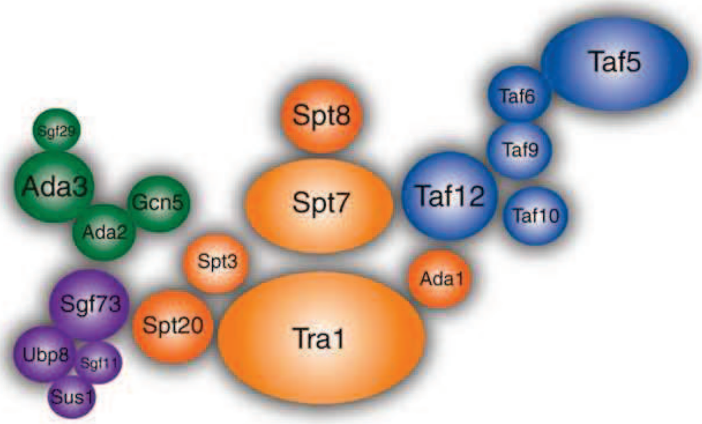

Fig. 21. Interaction network of the yeast SAGA subunits based on deletion purifications, existing data from yeast two-hybrid assays and genetic interaction experiments.
From Lee et al., 2011.

Our group recently used EM approach to position the dUb module within SAGA complex structure (Durand et al., 2014). The dUb module was localized in the domain III, close to Gcn5, as determined in previous EM study (Wu et al., 2004; fig. 22). Interestingly, upon removal of dUb domain V was shown to disappear, and it was proposed that all three domains III-V undergo a large conformational rearrangement upon dUb deletion. Approximately in the same position (domain III) the authors have localized Spt8 subunit by immunolabeling. This study has also revealed the flexibility of one of the SAGA lobes (lobe B), comprising domains III, IV and V: domains III and V formed a molecular clamp adopting different conformations. Based on immunolocalization of Spt3 (Wu et al., 2004) and Spt8 (Durand et al., 2014), the authors have proposed that TBP is bound by the molecular clamp, interacting at the same time with both Spt subunits.

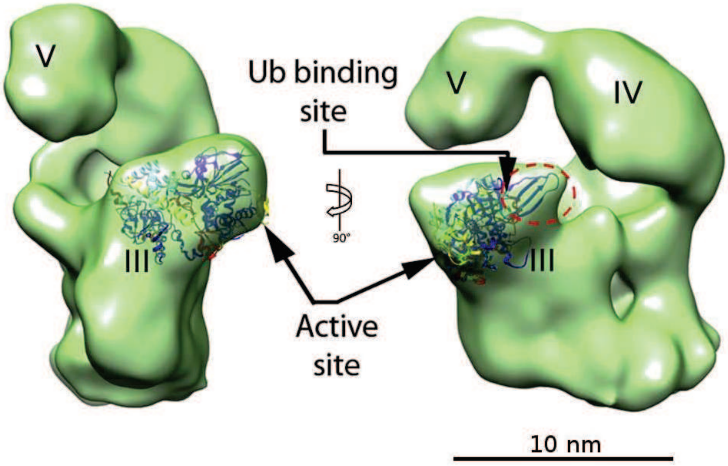

Fig. 22. Fitting of the atomic structure of the dUb module into lobe B of SAGA complex. *Domain numbers are indicated. From Durand et al., 2014.*

The group of S. Hahn have used another approach to address the architecture of the SAGA complex. In their study, they applied a combination of chemical cross-linking and mass spectrometry to determine a proximity map between SAGA subunits and also investigated the SAGA-TBP interactions (Han et al., 2014). The authors have observed cross-links between all TAF subunits (except Taf6 and Taf9), including Taf10-Spt7 and Taf12-Ada1 cross-links close to HFD domains of these proteins. These results are in agreement with previous studies and confirm the presence of a TFIID-like structural core in the center of the SAGA complex. This core complex was shown to cross-link all other subunits of SAGA. The absence of Taf6-Taf9 cross-links can be explained by the inaccessibility of these subunits to the cross-linker molecule, suggesting that they may be buried deep inside the complex or the tight interface is not accessible to the cross-linking agent. Another group of SAGA subunits, known to play structural role: Spt7, Spt20 and Ada1 – was found to occupy similar positions to the ones described in the EM study (Wu et al., 2004). The Spt7 and Spt20 subunits cross-link a different set of subunits, as compared to the proteomic map from Lee et al., 2011. Also in agreement with the EM study, Spt7 cross-links subunits of the HAT module, showing their close proximity. The cross-linking network of the HAT and dUb modules suggested that they are positioned close to each other and may functionally interact. The authors also analyzed several deletion mutants and have shown that specific deletions in Sgf73 and Ada3 completely dissociate the dUb and HAT modules, respectively, from the SAGA complex. In addition, disruption of the dUb module led to a significant decrease in the abundance of the HAT subunits, suggesting that the dUb module might stabilize the association of the HAT module with SAGA.

Cross-linking of the SAGA-TBP complex has revealed 4 subunits positioned closed to TBP: Spt3, Spt7, Spt8 and Spt20. While many biochemical and genetic studies implicated mostly Spt3 and Spt8 in TBP binding, Spt7 and Spt20 subunits might facilitate these interactions. The network of interactions between TBP and these Spt subunits is shown in fig. 18A.

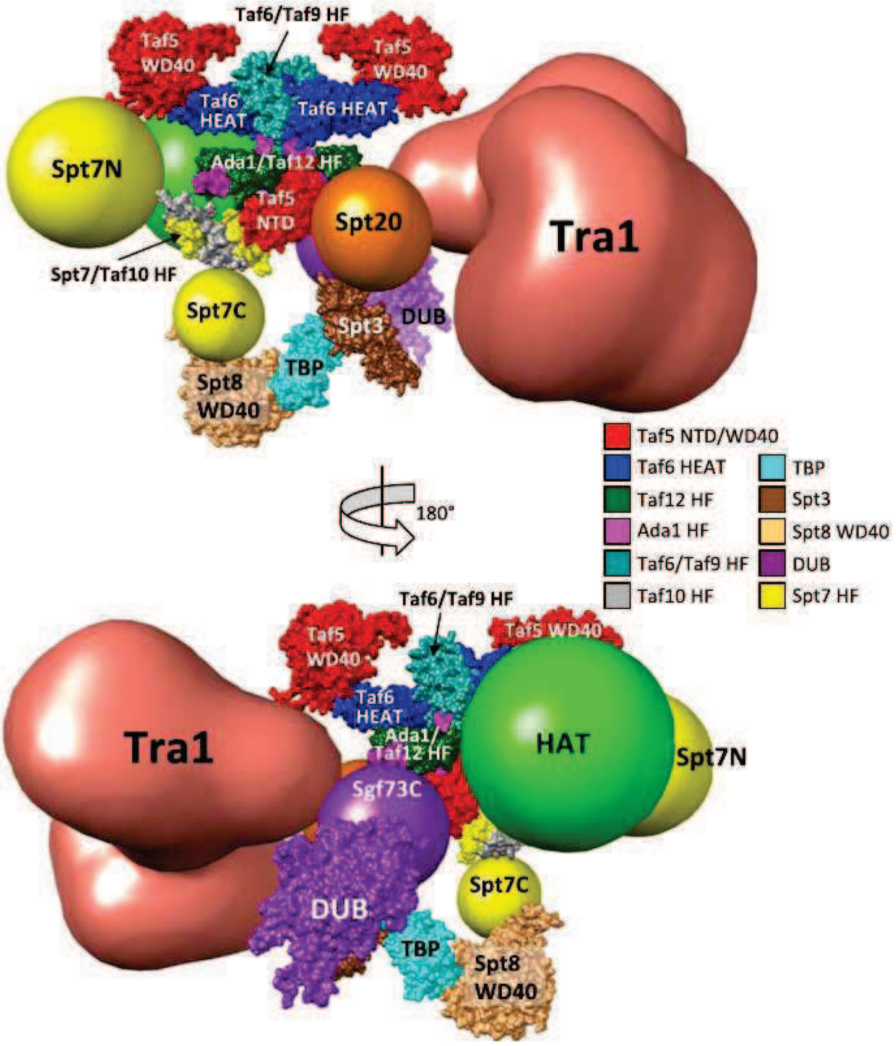

Fig. 23. Model of the molecular architecture of yeast SAGA complex.

SAGA subunits were positioned around the TFIID core EM structure. Spt7/Taf10 HF, Spt3, Spt8 WD40 are homology models. Crystal structures used: TBP, 1TBP; DUB, 3M99; Taf5 NTD (N-terminal domain), 2J49. HF – histone fold domain. From Han et al., 2014.

Based on the cross-linking experiments, the authors have build a model of SAGA (fig. 23), that displays several differences to the previously proposed models. Most of the mappings are in agreement with the EM study of Wu et al., except the close proximity between Spt20 and

Tra1, reported by Han et al. The new model is also distinct from the proteomic-based study of Lee et al., that placed Spt proteins in a single module.

Another recent study of SAGA structure was performed by group of C. Yip at the same time when the current work was in progress (Setiaputra et al., 2015). They used EM to analyze the flexibility of SAGA complex and readdress the subunit localization by immunolabeling, similar to the first study of Wu et al. The authors have shown that SAGA complex displays high flexibility, involving massive rearrangements of domains III-V, and can adopt different conformations. In general, the structure of SAGA looks similar to the one reported by Durand et al., however, domains III and V were not well resolved in the study of Setiaputra et al., probably due to limited size of the EM data set.

Deletion of key subunits, linking HAT and dUb modules to the rest of the complex (Ada2 and Sgf73, respectively), has allowed to approximately position these two modules within the SAGA structure. The HAT module was localized within domain V, in the most flexible region of SAGA, on the contrary to the previous EM study by Wu et al., that positioned Gcn5 in the domain III. In addition, the authors observed a significant loss of the HAT module (domain V) when dUb was removed, confirming that the removal of dUb destabilizes the HAT module, as proposed by Han et al. and demonstrated by Durand et al. Yip's group has also readdressed the positions of Spt3, Spt8 and Spt20. While Spt8 position in the domain III correlate with study of Durand et al., other two subunits have also been localized in domain III, conflicting with previous results. Furthermore, the dUb module was localized by antibody labeling in the domain IV, in contradiction with the previous study of Durand et al. Finally, the authors have addressed TBP binding to SAGA and localized its binding site near domain III, close to the Spt subunits.

1.5. Aims of the work.

Transcription of protein coding genes by RNA Pol II is tightly controlled and requires the coordinated action of a large number of protein complexes that trigger the remodeling of the chromatin, followed by specific assembly of the PIC at the promoter of the transcribed genes, leading to the correct positioning of RNA Pol II at the transcription start site. Sequence-specific transcriptional activators and post-translational modifications of nucleosomal histones contribute to the recruitment of multisubunit coactivator complexes acting as bridging factors between the transcriptional activators and the PIC. Coactivators function to modify the chromatin structure around the promoter region and coordinate PIC assembly with epigenetic chromatin modifications and with activator-mediated signaling events.

The multiprotein SAGA coactivator complex is a paramount example of this class of transcriptional regulators. SAGA interacts with many transcriptional activators such as Gal4 or Gcn4 in yeast, and the c-MYC or E2F oncogenes in human. It contains several protein domains that recognize modified histone tails and is a major histone H3 acetylating enzyme. SAGA regulates the expression of many genes in yeast, flies and mammals, and has been shown to contribute to many fundamental processes, including cell proliferation, metazoan development and response to stress signals. Recently, a genome-wide analysis of post-translational modifications, mediated by SAGA complex, has revealed that SAGA acts on a whole transcribed genome in yeast and human cells, acetylating promoters and deubiquitinating transcribed regions of all expressed genes, and is involved in transcription mediated by RNA Pol II. That indicates a role for SAGA as a general cofactor required for RNA Pol II recruitment and transcription.

In yeast, the 1.8 MDa SAGA complex is composed of 19 distinct subunits, most of which have highly conserved homologs in higher eukaryotes. Important research efforts aim at understanding the molecular organization of SAGA, but up to now no high resolution structure allowing the 3D positioning of each subunit is available. The SAGA complex was shown to have a modular organization by genetic complementation studies, electron microscopy structures, quantitative proteomics of systematic deletion mutants and protein cross-linking experiments coupled to mass spectrometry. However, depending on the method used, the inter-subunit interactions and the relative subunit positions differ.

SAGA contains two enzymatic activities to acetylate or deubiquinate nucleosomal histone tails. Impairment of either of SAGA’s enzymes leads to embryonic lethality, and mutations have been described in human genes, encoding these subunits, that cause neurodegeneration or contribute to tumorigenesis. HAT activity is carried out by the Gcn5 subunit and modulated by the Ada2 and Ada3 subunits. Together with Sgf29, these four subunits form the HAT module, a major regulator of histone H3 acetylation in yeast cells. The spatial location of the HAT module within SAGA complex is under debate.

At SAGA-dependent promoters, SAGA is required for the recruitment of the transcription machinery. The Spt3 and Spt8 subunits were both shown to directly interact with TBP indicating that SAGA plays a role in regulating the promoter-TBP interaction. TBP can bind to the entire SAGA complex, but in contrast to interaction studies performed on isolated SAGA subunits, this interaction requires Spt8 and Ada3 and was found to be independent of Spt3. The positions of the Spt3 and Spt8 subunits and the interaction site of TBP have been roughly mapped in 2D views obtained by electron microscopy and conflicting results have been reported.

Thus, to the date, controversial results were obtained in the field for the position of the HAT module and for several key subunits such as Gcn5, Spt3 and Spt20. The objective of this study was to understand the structural organization of the transcriptional coactivator SAGA and provide a consensual structural framework to understand the functions of the complex. During the course of this work, the following issues were examined:

- **What is the architecture of the HAT module and its localization in yeast SAGA complex?**

To address the architecture of the SAGA HAT module, chemical cross-linking mass spectrometry was used in combination with homology modeling in order to build an initial low resolution structure of the SAGA HAT. We have recombinantly expressed its human homolog in insect cells and addressed the topological organization of SAGA HAT by cross-linking mass spectrometry. A linkage map clearly highlights several discrete inter-linked areas: in particular, extensive cross-linking was observed between ADA3 and ADA2B, suggesting a strong network of connections between these two subunits, while no cross-links were detected between SGF29 and ADA2B. On the other hand, no cross-links could be detected for several domains of HAT proteins. Interestingly, these non-cross-linked domains are often involved in interactions with modified histone tails and are likely to be exposed to bind nucleosomes thus being less prone to interactions with subunits of the HAT module. The obtained cross-linking map was used to build a model of the SAGA HAT module based on available atomic structures or homology models of HAT protein domains. We compared our results with the recently published linkage map of the full yeast SAGA and noticed that in full complex Ada3 makes fewer links with other subunits of the HAT module, at the same time revealing numerous cross-links with other proteins like Taf5, Taf6, Taf12, Spt7, Sgf73 and Ada1. We showed that in human HAT module ADA3 is extensively cross-linked to the other HAT subunits: ADA2B, SGF29 and GCN5. Altogether these results suggest that the HAT module is reorganized upon integration in the SAGA complex.

In addition to cross-linking/MS, we used EM analysis of negatively stained HAT module to dissect its structure. However, the resulting 3D structure displayed a very high heterogeneity, varying in size and shape, thus we could not obtain a good structure of HAT by this method.

To position the HAT module inside full SAGA, we have purified the *S. cerevisiae* SAGA complex from deletion mutant yeast strain, where the whole HAT module was removed. The mutant complex was negatively stained and analyzed by EM. The structure was compared to complete *S. cerevisiae* SAGA structure, obtained previously in the lab, and revealed that the HAT is located at the most flexible part of the complex – the tail of the upper lobe. The mutant complex displayed even higher heterogeneity, since the upper lobe was more flexible, precluding high-resolution analysis.

- **What is the molecular organization of yeast SAGA complex?**

To improve the existing knowledge of SAGA complex, we have determined its structure using cryo electron microscopy in near-native conditions. The single particle EM image analysis of fully hydrated SAGA complex allowed us to reach a resolution of 18 Å which is the first cryo-EM structure of this coactivator complex at this resolution. The structure reveals that the complex is composed of two large lobes: the lower lobe (lobe A), containing Tra1 and several TAF subunits, and the upper lobe (lobe B), containing Ada, Gcn5, Spt and dUb subunits. In comparison with previous studies, we have analyzed the flexibility of the upper lobe and shown two types of movement: opening and closing of a molecular clamp formed by the upper lobe and a complex rotation of the upper lobe with respect to the lower one. Such flexibility represents a key obstacle to attaining higher resolution of SAGA structure.

- **How does SAGA interact with TBP?**

To assess the TBP binding site within SAGA, the SAGA complex was incubated with purified TBP and analyzed by negative staining EM. We were able to identify the open and closed conformations of the lobe B, indicating that TBP seems not to be involved in the opening and closing of the molecular clamp formed by the lobe B of SAGA as proposed before. At the resolution of about 25 Å we were able to identify the positions of TBP in the EM maps. Interestingly, we have observed that in the closed conformation, lobe B undergoes a large conformational change upon TBP binding.

Chapter 2. Methods.

2.1. Purification of recombinant human SAGA HAT subcomplex.

Solving the structure of multiprotein macromolecular complexes imposes a challenging task of producing them in sufficient quantity and purity. Many of such complexes are present in low amounts in the endogenous host, hampering their purification from the native source. Multiple recombinant techniques for heterologous protein expression have been developed to overcome these limitations. Most popular expression system remains *Escherichia coli*, and today numerous plasmids and strains exist for this system. However, expression of eukaryotic proteins in *E. coli* sets several restrictions on a target protein, such as limited molecular size, absence of eukaryotic chaperones for proper folding or specific post-translational modifications. Thus, eukaryotic expression systems using yeast, insect or mammalian cells have been developed. In this work, recombinant human SAGA HAT subcomplex was produced in insect cells using Multibac – a baculovirus expression system.

2.1.1. Baculovirus expression in insect cells.

Baculovirus expression systems are based on a recombinant baculovirus carrying the gene of interest, that is used to infect insect cells, leading to high-level production of the single protein or protein complex of interest. Multibac system was originally designed to produce protein complexes consisting of many subunits (Berger et al., 2004). It includes a set of synthetic DNA plasmids for recombination with genetically engineered baculovirus genome (bacmid), originally derived from *Autographa californica* nuclear polyhedrosis virus (AcNPV). This bacmid is then used for infection of *Spodoptera frugiperda* insect cells. The general scheme of Multibac system is outlined in fig. 24 and is described further.

Expression of a multiprotein complex implies integration of several genes of interest into a baculovirus genome, that requires laborious and technically difficult serial insertions of each gene into the bacmid. Multibac system facilitates multigene assembly by utilizing small (< 3 kb) synthetic DNA plasmids, called acceptors and donors (fig. 24A).

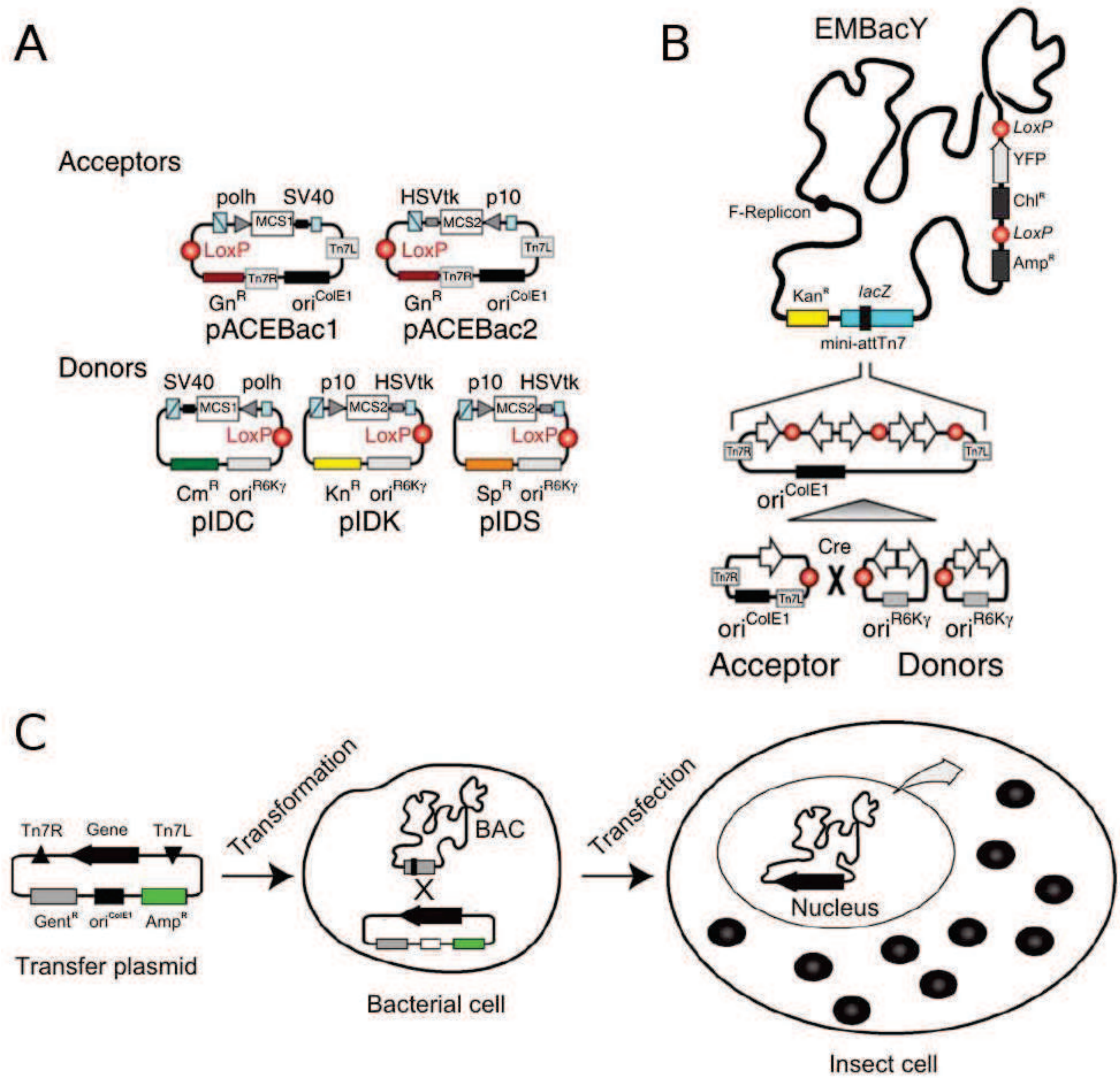

Fig. 24. The Multibac system.

(a) Multibac includes an array of small synthetic DNA plasmids: donors and acceptors. The acceptors can replicate normally, while donors have a conditional origin of replication. (b) Donors and acceptors with loaded gene constructs are recombined by Cre-Lox recombination, producing multigene cassettes, that are integrated into bacmid by Tn7 transposition. YFP gene in the bacmid serves as an additional expression marker. (c) Principle of protein production by Multibac system. Multigene constructs are integrated into the Multibac bacmids maintained in special E. coli cells. Then composite bacmid is isolated and used for infection of the insect cells. See details in the text. Adapted from Trowitzsch et al., 2010 and Bieniossek et al., 2012.

Acceptors and donors can carry one or several genes and are easily recombined into a single multigene construct. Acceptors contain a standard origin of replication (ori^{ColE1}), antibiotic resistance marker, LoxP recombination site, Tn7 transposition site and expression cassette that includes: strong baculoviral promoter (p10 or polh), DNA sequence for inserting multiple genes of interest and eukaryotic polyadenylation signal (SV40 or HSVtk). Donors are similar to acceptors, except the absence of transposition site and having a different resistance marker as well as different origin of replication ($ori^{R6K\gamma}$). The latter one is a conditional origin of replication, that requires presence of specific protein π for proper replication in *E. coli.* As a result, donors can survive in absence of π protein only if they are fused with acceptor, that harbors a normal origin of replication. This allows flexible and efficient assembly of multigene constructs from donors and acceptors. The conjoining of one or several donors with a single acceptor is achieved by Cre-Lox recombination via LoxP sites (fig. 24B), resulting in a mixture of plasmid dimers, trimers and/or tetramers. Successful recombination can be selected by screening in the background lacking π protein. The target plasmid containing multiple genes can be selected by precise combination of antibiotics, due to presence of different resistance markers in different donors and acceptors. In addition, Cre-Lox recombination can be reversed, opening numerous possibilities for combinatorial assembly of different multigene expression plasmids.

After the desired construct is assembled and selected by multiple antibiotic screen, it is integrated into genetically modified bacmid by Tn7 transposition (fig. 24C). The Multibac bacmid, expressed in a special *E. coli* strain, represents an artificial chromosome in which several genes, responsible for protease and apoptotic activity of the virus, have been deleted to enhance protein production. Successful transposition disrupts *LacZα* gene present in the bacmid, enabling standard blue/white screening of colonies with integrated multigene cassette. The bacmid also contains an additional LoxP site, that can be used for integrating other constructs. Finally, the composite bacmid is purified and used for infection of insect cells.

In this work, human SAGA HAT subcomplex (hGCN5, hADA2B, mADA3, hSGF29) was expressed as follows. The constructs for the expression were generated with Multibac system. GCN5 with TEV-cleavable 6xHis tag on N-terminus was cloned into pFL acceptor plasmid, the other three genes encoding ADA2B, ADA3 and SGF29 were cloned into donor plasmids (pIDS, pIDC and pIDK respectively), which were subsequently combined via Cre-Lox recombination. The total construct was integrated via Tn7 transposition into a bacmid containing YFP as an independent expression level marker. The bacmid was isolated and used for infection of 1 liter of Sf21 insect cells (160 ml of virus for 1 liter of $0.8*10^6$ cells/ml SF21 insect cells in

Sf900II serum-free media, MOI 15-20). Infected cells were incubated in Erlenmeyer flasks on the shakers in a 27°C room. YFP fluorescence was used to monitor overall expression levels after the day of infection. After 72 h cells were harvested by centrifugation (400 g, 20 min, 4°C). The pellet was washed in ice cold PBS with 10% glycerol, frozen in liquid nitrogen and kept at -80°C until the purification.

2.1.2. Purification of SAGA HAT subcomplex.

Here we will briefly summarize the purification protocol, the complete description can be found in Annex 7.1. All purification steps were performed at 4°C. Cell pellet from 1 liter of culture was thawed and dissolved in 40 ml of lysis buffer (20 mM HEPES-KOH pH 8.0, 500 mM NaCl, 10% glycerol, Protease Inhibitor Cocktail (Roche)). Cells were mildly sonicated on ice for 3 min. 40 units of bovine DNAse I were added, and the extract was incubated for 30 min at 4°C. To remove the debris, cell extract was cleared by ultracentrifugation for 1 h at 50.000 g.

SAGA HAT subcomplexes were purified by immobilized metal affinity chromatography (IMAC) using His tag on the N-terminus of GCN5 subunit, followed by size-exclusion chromatography. Two different IMAC resins were tested for efficient SAGA HAT binding: cobalt (TALON resin) and nickel (Ni-IDA). In the final protocol SAGA HAT was purified using HiTrap Chelating HP 5 ml column (GE Healthcare) loaded with nickel. After washing steps with increasing concentration of imidazole, HAT was eluted at 250 mM imidazole, as confirmed by 10% SDS-PAGE. Peak elutions were pooled and further purified on Superdex S200 16/60 GL (GE Healthcare) gel filtration column. Subcomplexes were eluted in a similar buffer (20 mM HEPES-KOH pH 8.0, 500 mM NaCl, 10% glycerol, 1 mM DTT, 0.1 mM EDTA) and analyzed by 10% SDS-PAGE. Peak fractions were pooled, concentrated on Amicon 30 kDa column, then flash frozen in liquid nitrogen and stored at -80°C.

2.1.3. Pull-down assays with HAT subunits.

Constructs for expression of different subunit combinations of the HAT module were generated using Multibac system and expressed in insect cells as described before. Small scale purifications using His tag on GCN5 were performed in batch with cobalt resin and eluates were analyzed on 10% SDS-PAGE.

2.2. Purification of yeast SAGA complex.

Electron microscopy requires a protein sample purified in sufficient quantity (Table 7) and quality. For negative staining (see Chapter 2.3.4.1), the concentration that normally gives good particle distribution on the grid is about 25 µg/ml. Cryo-EM (see Chapter 2.3.4.2) typically needs 2 to 4 times more than that. This is due to the fact that after deposition on the grid, most of the sample is adsorbed to the carbon, and much smaller amounts go into the holes. Purified sample has to be as homogeneous as possible, free of contaminants, stable during the grid preparation and (for cryo-EM) dissolved in a compatible buffer. High concentrations of glycerol (> 1%), sucrose (> 1%), salt (approx. > 250 mM) and detergents should be avoided in sample buffers since they will interfere with the high-contrast image formation in vitrified ice.

Yeast SAGA contains many different subunits and represents a large fragile protein complex, present in very low amounts (about 100-200 copies per yeast cell) in the cell at the endogenous level. Due to its complexity, the task of producing such a large complex in recombinant systems (e.g., by Multibac) is presently considered not possible. Endogenous SAGA complex was purified from yeast using high affinity tags. Original protocol of the purification of SAGA for negative stain study (used for only Ada2Δ dataset) was developed in the lab by A. Durand (Durand et al., 2014) and based on TAP purification method (Rigaut et al., 1999; Puig et al., 2001). The improved purification protocol was developed by A. Ben-Shem (unpublished) and based on SBP-tag purification. The latter one was used for SAGA purification followed by both negative stain and cryo-EM studies. Both protocols will be briefly summarized here.

2.2.1. Yeast strains generation.

In this work, *S. cerevisiae* yeast strains were the gifts from our collaborators: the strain with TAP tag, integrated at the C-terminus of Spt20 was provided by the lab of Marc Timmers (University Medical Center Utrecht, Netherlands); the strain containing Ada1 with TAP tag on C-terminus and deleted Ada2 was generated in BY4741 in Jerry Workman's lab (Stowers Institute, USA). *Pichia pastoris* (*Komagataella pastoris*) strain with SBP tag at the C-terminus of Sgf73 was generated by A. Ben-Shem in our lab. Summary of different strains used in this work is provided in Table 6.

Strain name	Yeast species	Genotype
BY4742	*S. cerevisiae*	MATα his3Δ1 leu2Δ0 lys2Δ0 ura3Δ0
Spt20-TAP	*S. cerevisiae*	Isogenic to BY4742, except SPT20-TAP::URA3
BY4741	*S. cerevisiae*	MATa his3Δ1 leu2Δ0 met15Δ0 ura3Δ0
Ada1-TAPΔAda2	*S. cerevisiae*	Isogenic to BY4741, except Ada1-TAP::HIS, ada2Δ::Kan
Sgf73-SBP	*P. pastoris*	X-33 wild-type

Table 6. List of yeast strains used in this work.

2.2.2. TAP tag *S. cerevisiae* purification.

Tandem affinity purification tag (TAP) is a fusion cassette including calmodulin-binding peptide (CBP), TEV cleavage site and protein A of *Staphylococcus aureus* with two IgG binding domains. The tag can be fused with either termini of a target protein. In the case when TAP tag is placed on a C-terminus, a natural promoter can be used, providing normal expression level. The purification procedure consists of two steps: first, the protein extract is incubated with IgG beads that bind protein A. After several washes, the target protein complex is recovered by incubation with TEV protease. In the second step, the eluate is incubated with calmodulin-coated beads in the presence of calcium. After washing, the purified complex is released with EGTA.

In this work, *S. cerevisiae* cells were grown in flasks to OD of 2-2.5 in YPD at 30°C. 6 liters of culture allowed to produce about 15 g of yeast cells. Cell pellets were incubated in lysis buffer (50 mM HEPES-KOH pH 8.0, 300 mM NaCl, 2 mM $MgCl_2$, 10% glycerol, 0.1% Tween-20) with protease inhibitor (Roche) and the lysis was performed by vortexing the suspension with 15 g of glass beads in 4°C room five times for 1 minute with 1 minute breaks to let the cells cool down. Lysate was cleared by centrifugation at 40.000 g for 1 h at 4°C. Next, the supernatant was incubated with 200 µl of IgG sepharose beads (GE Healthcare, ref. 17-0969-01) for 5-6 h at 4°C. Beads were washed four times with 3ml of buffer (30 mM HEPES-KOH pH 8.0, 300 mM NaCl, 0.1% Tween-20) and TEV cleavage was performed overnight with 3-5 µg of TEV in 400ul of buffer. Eluates were incubated with 50 µl of calmodulin affinity resin (Agilent Technologies, ref. 214303-52) for 1-2 h at 4°C. Beads were washed four times with 30 mM HEPES-KOH pH 8.0, 300 mM NaCl, 0.1% Tween-20, 2mM $CaCl_2$ and SAGA was eluted in 30 mM HEPES-KOH pH 8.0, 300 mM NaCl, 2 mM EGTA. The estimated amount of purified SAGA complex by this method was 2-3 µg in 100 µl.

2.2.3. SBP tag *P. pastoris* purification.

Streptavidin binding peptide (SBP) is a 38 amino acid long peptide (Keefe et al., 2001) that is able to bind streptavidin with very high affinity (Kd = 2.5 nM). Elution with biotin permits recovery of target protein complexes in mild conditions.

In this work, *P. pastoris* cells were grown in flasks to OD of 10-20 in YPD at 30°C. Due to higher growth rate than *S. cerevisiae*, 1-2 liters of culture allowed to produce 20-30 g of yeast cells. Cell wall was lysed by addition of lyticase, followed by mechanical dispersion (T25 digital ULTRA-TURRAX, IKA) in a phosphate buffer (pH 6.5, 18% Ficol400), that allowed to release intact nuclei. After several washes, nuclear extract was obtained by mild salt extraction (20 mM HEPES-KOH pH 8, 300 mM NaCl). Next, 1% PEG was used to pellet and remove the debris, followed by 6% PEG precipitation that allowed to pellet protein complexes. The pellet was solubilized in a minimal volume and avidin was added to block biotinylated proteins. The mixture was incubated with streptavidin beads for 4 h in 20 mM HEPES-KOH pH 8, 250 mM NaCl, 10% glycerol and eluted with biotin. Eluate was concentrated on Vivaspin 50 kDa concentrators and loaded onto 10-30% sucrose gradient. SAGA was fractionated at about 25% sucrose, centrifuged and buffer exchanged in Vivaspin concentrators into 10% sucrose solution.

2.2.4. Cross-linking of purified *P. pastoris* SAGA for electron microscopy.

Full SAGA complex was shown to be very fragile, highly sensitive to the temperature changes and to easily undergo subunit dissociation during sample preparation for electron microscopy (unpublished data; Setiaputra et al., 2015). To stabilize the complex we used cross-linking with glutaraldehyde. Briefly, 0.5% glutaraldehyde was prepared from 25% stock in the same buffer as SAGA (20 mM HEPES-KOH pH 8, 150 mM KOAc, 5 mM MgAc, 10% sucrose, 0.0045% DDM, 2 mM TCEP). 2 µl of this solution were added to 8 µl of 0.5 mg/ml SAGA and incubated on ice for 2 hours at 4°C, so that the final cross-linker concentration was 0.1%. Cross-linking time and glutaraldehyde concentration was optimized by running a 10% SDS-PAGE: the complex was fully cross-linked, when only a single band was observed between the stacking and separating gels.

For the supported cryo-EM (see part 2.3.4.3.4), the cross-linked sample was diluted 10 times in the sample buffer (without the sucrose) to approximately 0.05 mg/ml. For the unsupported cryo-EM, in order to completely eliminate the sucrose, 10 µl of cross-linked sample were dialyzed against 1 ml of sample buffer (without the sucrose) for 4 h at 4°C using 10 kDa Slide-a-lyzer MINI dialysis kit (Thermo Scientific, ref. 69572). The final sample volume was about 14-16 µl. Sucrose concentration decrease after dialysis was followed by absorbance at 319 nm after addition of 95% sulfuric acid (Houle et al., 1967).

2.2.5. Reconstitution of *P. pastoris* SAGA-TBP complex.

To reconstitute the complex formed between SAGA and TBP, five molar excess of TBP was used. Full length His-tagged *S. cerevisiae* TBP was overexpressed and purified from *E. coli*. Few drops of MES pH 5.5 was added to SAGA to reduce the pH to the values used in *in vitro* transcription (7.6-7.8). TBP was first diluted to roughly the same concentration as SAGA (0.5 mg/ml) in the same buffer with 10% sucrose. After SAGA addition, the mixture was incubated at 4°C for at least 4 hours. Next, the cross-linking with glutaraldehyde was performed as described before.

2.3. Electron microscopy analysis of yeast SAGA complexes.

2.3.1. Overview of the methods in structural biology.

Over the last century structural biology has developed into an indispensable framework for explanation and interpretation of many ideas coming from molecular biology. Today it represents an essential tool, combining many different methods in order to untangle the intricate and complex processes of life. Proteins, being responsible for nearly every task in cellular life, usually carry out their functions within large macromolecular complexes inside the cell. Thereby, simply resolving the structure of single proteins *in vitro* is not sufficient, as long as they are taken out of the cellular context, lacking interaction partners and proper environment. Thus, the major challenge of structural biology today is revealing the architecture of dynamic protein assemblies, paving the way towards the ultimate goal of building a 3D model of a living cell at most detailed, atomic level.

Traditional structural biology methods include X-ray crystallography (XC), nuclear magnetic resonance (NMR) and electron microscopy (EM). Over the years many other methods have evolved such as native and cross-linking mass spectrometry (MS), small angle X-ray and neutron scattering (SAXS and SANS), super-resolution light microscopy, atomic force microscopy, various biophysical and bioinformatical methods (Perrakis et al., 2011). Since each of the technique has its unique advantages and disadvantages (Table 7), it has become apparent that only through integrating structural data derived by different methods we will be able to understand how proteins and their complexes dynamically interact in cell environment in order to provide their function (Cowieson et al., 2008).

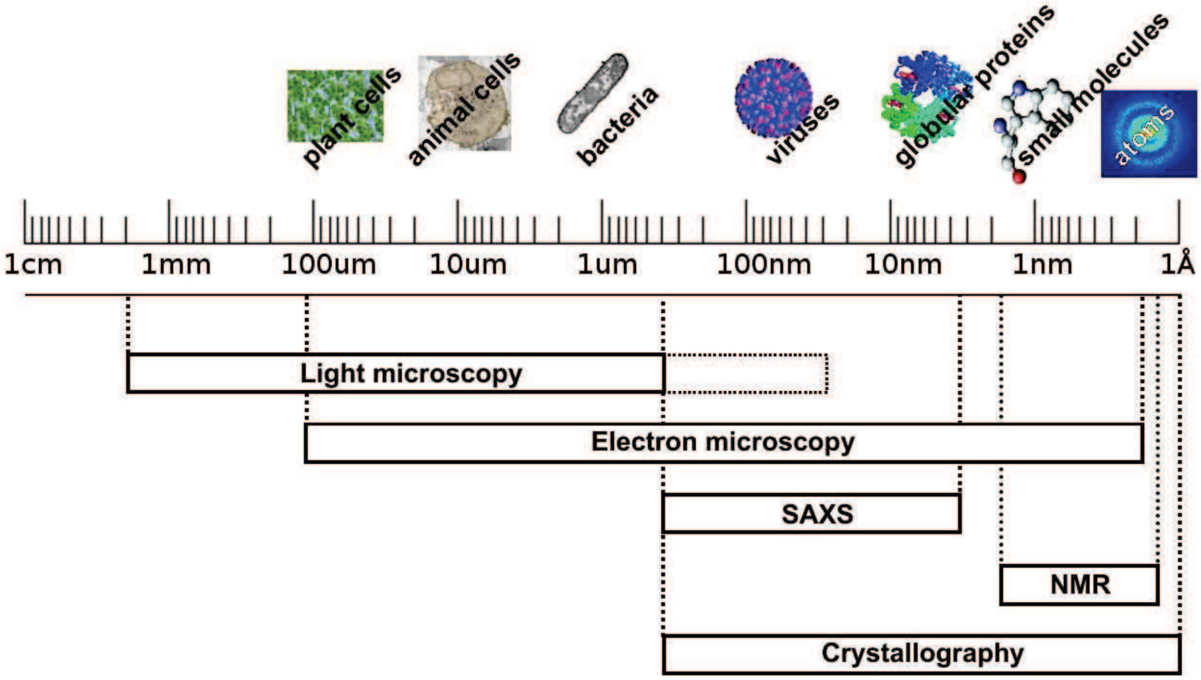

Fig. 25. Illustration of the dimensions of some life-related systems on a logarithmic scale (powers of 10).

The range, visible by some of the different structural biology methods, is indicated.

Proper functioning of the cell demands temporal and spatial coordination of complex biological processes over a tremendous range of length and time scales (fig. 25). Hence, filling the gaps between methods with different resolution capabilities by multiscale data integration deems necessary to be able to understand protein functions at each level, from atoms through molecules, organelles and cells. Macromolecular X-ray crystallography (XC) has evolved as a major technique for resolving protein structures at atomic level, owing to very short wavelength of X-ray radiation. Most of the structures presented in the PDB database were solved by this method. However, the major challenge remains obtaining protein crystals and large sample quantities. Besides, XC does not provide temporal information, since a crystal represents a static snapshot of a certain macromolecule. NMR does not depend on protein crystals and can give comparable level of resolution of protein structures, it also allows to observe conformational changes in solution, but this technique is limited to small proteins (up to 30-50 kDa) and requires very high sample amounts (10-30 mg). Mass spectrometry occupies a special niche in structural biology applications and allows to study protein complexes ranging in mass, size, solubility, oligomeric composition, stoichiometry etc. Recently it has been applied to determination of three-dimensional shape information and topology of protein complexes by combination with cross-linking and molecular modeling, producing the information at amino acid level. Electron microscopy has proved to be an extremely useful technique for studying macromolecular complexes that are difficult to crystallize or produce in large quantities. Covering the range from single proteins to the size of whole cells (few microns), having moderate restrictions of protein size (only the lower limit of 200-250 kDa) and exploring protein structure in near native conditions, it often becomes a method of choice for gaining structural information on macromolecular assemblies. Heterogeneous samples can be also analyzed, allowing dynamic and unstable complexes to be studied. However, the main problem in cryo-EM remains radiation damage, limiting attainable resolution to the levels above the ones, usually achievable with XC or NMR. Nevertheless, in last few years advances in the development of electron detectors and image processing software made it possible to significantly push the resolution towards 3-5 Å (Kuhlbrandt, 2014). Outstanding examples include numerous ribosome structures (Bai et al., 2013; Voorhees et al., 2014; Wong et al., 2014; Khatter et al., 2015), ion channels (Liao et al., 2013; Paulsen et al., 2015), enzymes (Allegretti et al., 2014; Bartesaghi et al., 2014; Lu et al., 2014), receptors (Yan et al., 2015) etc. Altogether these results suggest that cryo electron microscopy enters a “new era”, when structure determination of large and/or flexible macromolecular complexes at almost atomic level will be easier and quicker than before.

Method	Advantages	Disadvantages	Sample requirement*
X-ray	Resolution 1-5 Å	Need for crystals High amounts of sample	500 pmol
NMR	Resolution 1-10 Å Dynamic studies in solution	High sample concentration 50 kDa size limit Isotope marks needed	0.2-0.4 µmol
Cryo-EM	Resolution 3-20 Å Low sample concentration Heterogeneous complexes can be studied	Radiation damage Low contrast images	0.25 pmol

Table 7. Comparison of standard structural biology methods.
** values taken from Frank, 1996.*

SAGA is low abundant protein in the cell at endogenous level (100-200 copies per yeast cell), thus difficult to produce in large quantities, at the same time it is a 1.8 MDa complex with modular architecture. Therefore, cryo-EM appears to be a suitable technique for structural study of this complex.

2.3.2. Transmission electron microscopy: introduction.

Transmission electron microscopy uses high energy electron beam that interacts with the specimen when passing through it and creates an image, reflecting the specimen composition and structure. To understand how electron interacts with the sample, one first need to consider why we use electrons as a source of radiation. Electromagnetic spectrum is a standard way to represent most types of electromagnetic radiation. It extends from several picometers (gamma rays) up to few kilometers (radio waves), covering a huge range of wavelengths, frequencies and energies. Most parts of the spectrum are used in study and characterization of matter. Use of electromagnetic radiation to visualize objects requires use of wavelength comparable with the smallest features one wishes to resolve.

The most obvious way to study a structure of an object is to look at it. Visible light contributes only a small portion in the full spectrum (400-700 nm), however it's indispensable for most living organisms. In application to structural studies, visible light is considered not very damaging to the sample and easy to focus with glass lenses, also a human eye is a perfect detector; however the long wavelength of light limits the minimum size of the objects of the study. Indeed, one can use a light microscope to study cell structure or localization of specific proteins in the cell, but it is impossible to see the details smaller than half of the wavelength (hence, diffraction limit). In the last decade development of super-resolution fluorescence microscopy techniques, such as STED, GSD, STORM, PALM etc. has allowed to overcome the diffraction limit and reach resolutions about 20 nm in lateral dimensions and 50 nm in the axial dimension (Huang et al., 2008). However, most of these techniques are applied at the cellular level and are not yet able to resolve individual protein structures.

X-rays have a small wavelength (few Angstroms), comparable to the atom size, and high sample penetration, however they are damaging to the sample and difficult to focus. During interaction with the specimen, X-rays are scattered by sample electrons, however scattering from a single molecule is incredibly weak and difficult to detect, that's why crystals are needed. Composed of the same type of molecules, arranged in a regular way, crystal acts as an amplifier of weak scattered signal. However, in a crystallographic experiment we cannot directly obtain an image of atoms inside the molecule, but rather a diffraction pattern reflecting electron positions in the specimen. It is a result of so-called phase problem in crystallography. Scattered in the crystal, X-ray waves with different phase sum up and produce a pattern of diffraction spots, where they have added up constructively. We can measure only the intensity of the spots, which gives essentially the number of X-ray photons, that is proportional to the squared amplitude of diffracted waves. Phases are lost during this experiment and can be recovered only with other methods such as isomorphous replacement (IR), multiple-wavelength anomalous dispersion (MAD) or molecular replacement (MR). These methods are based on producing a similar crystal, where few atoms have been replaced with heavy ones. This will perturb the diffraction pattern and will allow to deduce positions of few heavy atoms, giving a good guess about the phases. Phases together with amplitudes of the scattered waves are required to build an electron density map, since the information about the positions of electrons in the specimen is encoded in the phases and not in the amplitudes.

Neutrons and electrons are the types of particles that can be accelerated to bring their wavelength to Angstrom range. Neutrons have a wavelength of a few picometers and cause low sample damage, however they are very difficult to produce and to focus. On the other hand, accelerated electrons have also very short wavelength (about 2 pm at 300 keV), they can be focused by aid of electromagnetic lenses. In the sample, electrons are scattered much stronger than X-rays by both nuclei and sample's electrons, therefore allowing to image individual protein complexes and eliminating the need for crystals. However, electrons have poor sample penetration and produce high specimen damage, that represents a major obstacle for obtaining high-resolution information. Nevertheless, several techniques have been developed to overcome these limitations, and today electron microscopy fills the special niche in the field of structural biology, between X-ray crystallography and light microscopy.

2.3.3. Basic components of an electron microscope.

The first prototypes of electron microscope were built in 1932 by Knoll and Ruska (Knoll and Ruska, 1932), as soon as the idea of electromagnetic lenses that can focus electrons has developed into practical reality. At that time the microscope consisted of just two lenses, producing a total magnification of only few hundred times. Nowadays, as a result of gradual development in instrumentation, a modern electron microscope allows to reach a resolution of 50 pm (Erni et al., 2009) and magnification of few million times.

In the TEM, an electron source (gun) produces a beam of electrons that is made parallel by condenser lenses; this passes through a specimen and is then focused as an enlarged image onto the detector by the objective lens (fig. 26). The whole trajectory from source to detector is under vacuum and the specimen has to be very thin to allow the electrons to pass through it. In a more detailed view, the TEM includes electron gun, several lens systems, sample chamber, vacuum system and electron detectors (fig. 27). In order to understand how electron microscope works, its composition will be described, along with a brief image formation theory.

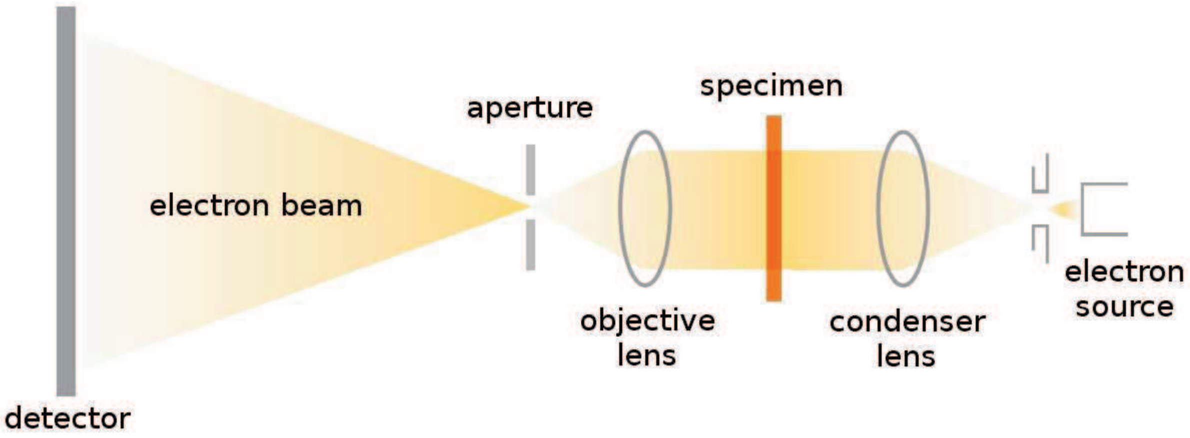

Fig. 26. The principle of transmission electron microscope (TEM).

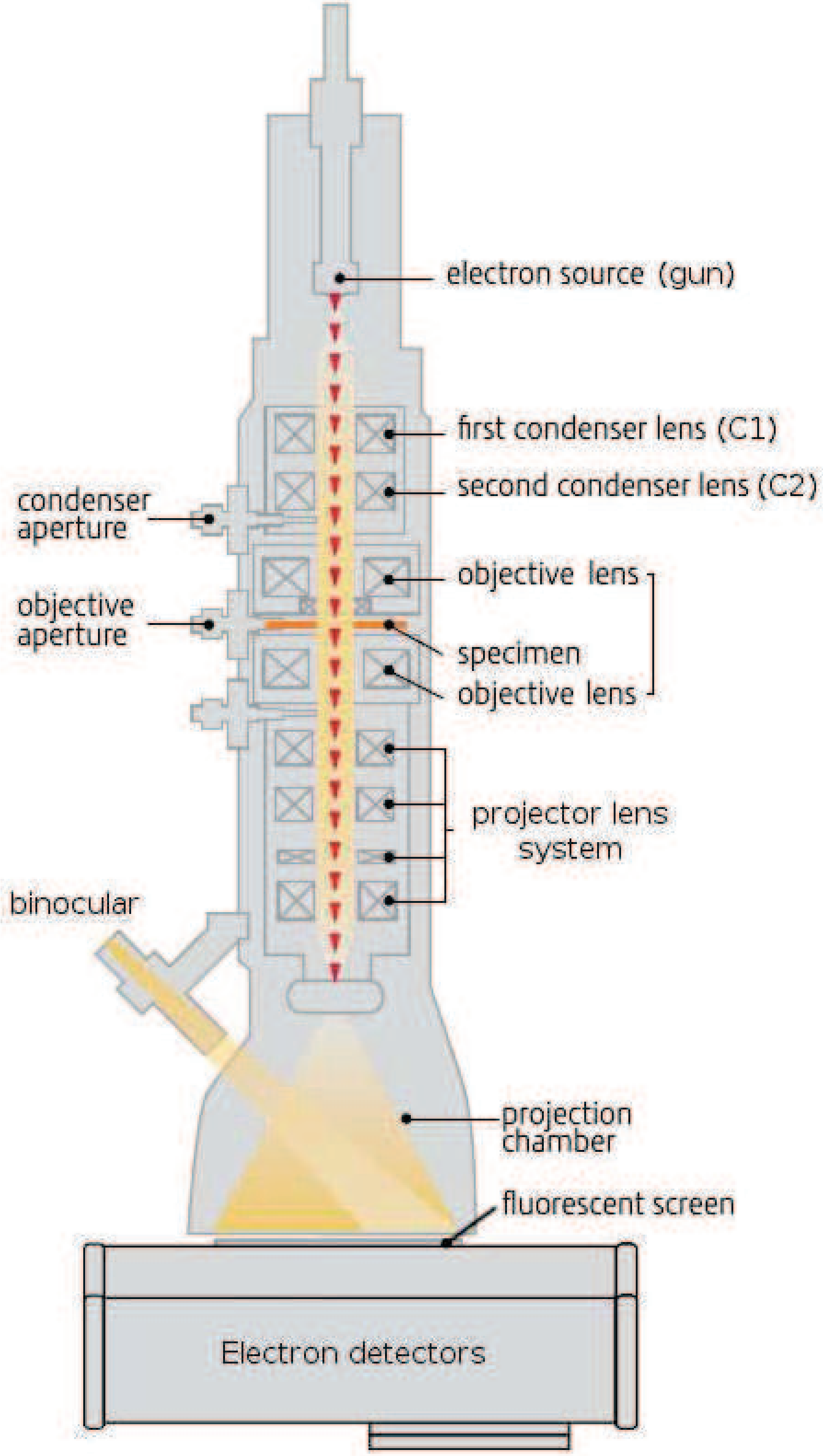

Fig. 27. A general scheme of TEM.

Vacuum systems are not shown for simplicity. Adapted from http://fei.com

2.3.3.1. *Electron gun.*

At the top of the electron microscope there is a source – a gun, generating highly accelerated electrons, and several condenser lenses, that focus the electron beam, produced by the gun, onto the sample. Parameters of the gun and condenser system determine the beam size that hit the specimen and the intensity of the final TEM image. The main component of the electron gun is a filament, that plays a role of cathode, to which a high negative voltage is applied, e.g. -300 kV. This causes electrons to escape from the tip of the filament and produce an electron spray. Afterward, electrons are focused into a single point called crossover and accelerated to very high speed. For this purpose, below the tip a set of anode electrodes is located. The first one meant to focus the electron spray is a Wehnelt cylinder, which is essentially a cup, surrounding the tip, with a hole inside, kept at slightly lower negative voltage than the tip. After focusing by the Wehnelt, electrons enter an accelerator stack – a set of discs, connected by resistors. To each disc slightly more positive voltage is applied, ranging from very negative to 0 kV. This allows electrons to accelerate and reach about 76% of the speed of light and wavelength of 2 pm (for 300 keV electrons). The filament, Wehnelt cylinder and accelerator stack comprise the basic components of an electron gun (fig. 28).

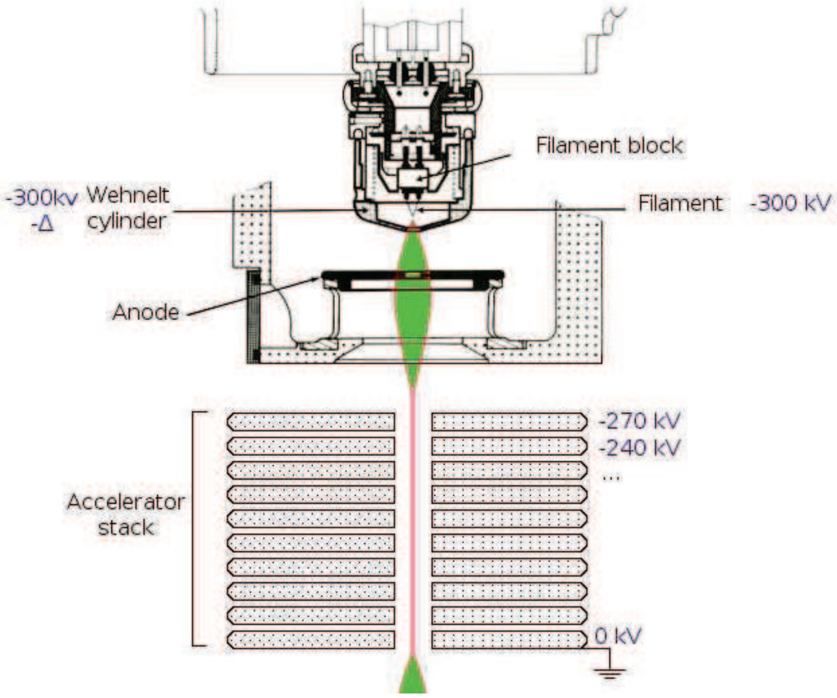

Fig. 28. Basic scheme of electron gun, running at 300 kV.

Filament plays a role of cathode, supplied with negative voltage of 300 kV. Wehnelt cylinder has slightly lower voltage (-300 kV - Δ) in order to focus outcoming electron spray. After passing through anode plate, electrons go into accelerator stack.

There are two types of electron guns used in TEM: thermionic and field emission gun (FEG). Thermionic gun comprises a wire or a crystal, that when heated strong enough produces a spray of electrons escaping from its surface. To avoid melting of the metal at high temperatures, usually tungsten or lanthanum hexaboride (LaB_6) are used. Since electrons are strongly scattered by the matter, the thermionic gun is operated at medium vacuum levels (10^{-2} – 10^{-4} Pa). FEG exploits a different principle to generate electrons: a high voltage, applied to a sharp fine needle of tungsten (called a tip), creates an electric field, stimulating electrons to tunnel out of the material, the process called field emission. FEG can operate at room temperature but requires much higher vacuum (10^{-6} – 10^{-9} Pa) to avoid surface contamination of the tip. Some modern FEGs (Schottky FEG) are treated with ZrO_2 and continuously heated to assist field emission and improve the gun stability.

When comparing thermionic guns and FEGs the important parameters to consider are brightness, energy spread, crossover size and lifetime. The brighter the beam, the more electrons will be produced by the gun and interact with the sample, hence more information can be obtained about the specimen. Energy spread and crossover size are correlated with temporal and spatial coherence, respectively. Temporal coherence implies that all electrons produced by the gun will have the same energy and subsequently will be focused at the same conjugate planes inside the microscope, reaching the detector at a certain plane. Electrons with slightly higher or lower energy will be focused less or more strongly and the final image on the detector will be blurred. Spatial coherence is achieved when all electrons are emanated from the same point in the source. If electrons go through the same point on the specimen at different angle, they will reach the detector in different places, thus again producing a blurred image. Thermionic guns are cheap and easy to replace, but they have high energy spread and crossover size. FEGs have higher brightness and longer lifetime, and at the same time lower energy spread and crossover size, effectively making this type of source perfect for high resolution applications.

2.3.3.2. Electromagnetic lenses and their aberrations.

Electromagnetic lens is comprised of many coils of copper wire, through which one passes the current, creating a magnetic field that is symmetric around the optical axis of the lens. When electron enters the lens in the direction diverging from the optical axis, it experiences a force that bends its trajectory into a concentric spiral, bringing the electron back onto the optical axis. Thus, the behavior of electromagnetic lens can be approximated to the action of convex glass lens in light optics.

Unfortunately, we still cannot produce perfect lenses for focusing the electrons, and every lens has a number of aberrations, that deteriorate the overall performance of the microscope. Also, one cannot perfectly align mechanically the lenses along the optical axis inside the microscope column. Therefore, almost every lens system in electron microscope has additional components: deflectors, stigmators and one or several apertures (fig. 29). Small adjustments of electron trajectory, required to account for lens misalignments, are provided by set of additional coils called deflectors. Stigmators are needed for the correction of astigmatism, a lens aberration caused by the imperfections in electromagnetic lens that create a slightly asymmetric magnetic field. Apertures are thin metal diaphragms with a few microns hole in the middle. They are used to block the electrons passing far away from the optical axis, that are usually prone to different aberrations. The most common lens aberrations are spherical and chromatic aberrations. Spherical aberration (C_s) appears when off-axis rays are focused stronger than the rays close to optical axis. In biological TEM it limits the achievable resolution to about 3 Å. Chromatic aberration (C_c) occurs when electrons with slightly lower energy are focused stronger than the others. This happens more often in electron tomography, when thick sample of a cell section is imaged. In both aberration cases, each point in the object will be focused by the lens into a disc of certain radius, depending on the extent of the aberration. However, it is possible to correct for these aberration with set of additional lenses (spherical aberration correction) or energy filter (chromatic aberration correction).

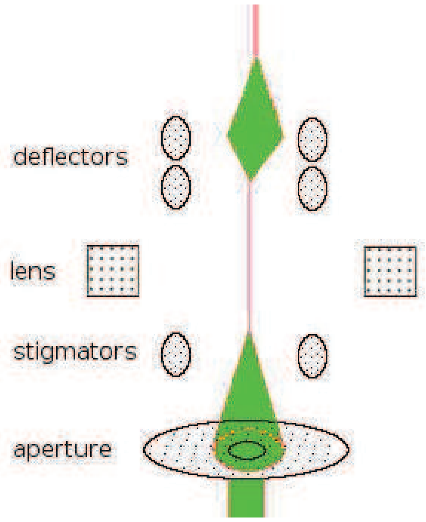

Fig. 29. Basic elements of an electron lens.

Deflectors are set of coils that bring the electron beam back onto optical axis, stigmators correct for possible astigmatism, aperture blocks far off-axis electrons.

2.3.3.3. Condenser, objective and projector lens systems.

In the electron microscope, below the gun there is a set of lenses comprising the condenser system (fig. 27). There are usually two or three lenses, that provide a uniform illumination of the sample. The first C1 lens takes the electrons emerged from accelerator stack of the gun and focus them into a spot of certain size. The small spot sizes (in diameter) are more coherent, since the most divergent electrons are not focused, however they have low intensity. The large spot sizes have exactly opposite properties. Next, the C2 lens defocus the spot produced earlier and spread the beam that allows to adjust the beam size and intensity of the specimen illumination. The most advanced microscopes has an additional C3 (mini-condenser) lens, that allows to vary the current in C2 lens and simultaneously have a parallel beam, even with a very small, high intensity beam. This gives better spatial resolution and stronger signal.

At the heart of the electron microscope lies the objective lens. It is the strongest lens in the microscope, that gives a primary magnification of 50 to 100 times and determines the final resolution and contrast. Right below the objective lens (also true for C1 and C2 lenses) there is an aperture, that is used to increase image contrast by blocking the part of scattered beam. The main function of the objective lens is to gather the electrons, scattered by the sample and produce a first focused image. This image is further magnified by projector system with a set of lenses, providing a range of possible final magnifications (10^3 – 10^6).

2.3.3.4. Vacuum system and sample chamber.

The microscope column is always maintained under high vacuum to avoid unnecessary scattering of the imaging electrons. In fact, if the microscope would be operated in air, it would be difficult to generate accelerated electrons, since air is a very bad insulator. The electron beam would be lost already 1 meter below the gun, due to very low penetration of the electrons through the air. Also, the sample would be immediately contaminated with hydrocarbon molecules. Thus, in order to maintain the vacuum inside the microscope column, several type of pumps are used each working in a specific pressure range. The ultra-high vacuum (10^{-9} Pa) has to be preserved in the gun area, the rest of the column is kept at about 10^{-5} Pa. The only place in the column where the vacuum can be temporarily switched off is a sample chamber. Many microscopes have a manual side-entry holders, the most modern ones have also an automatic sample loading system. In both cases, the sample, deposited onto electron microscopy grid, is inserted into the column from the side, right in the middle of the objective lens. In fact, the objective lens is split into two parts, called pole pieces, that are essentially the copper coils enclosed in soft iron shell. The sample holder is positioned exactly in the gap between the two pole pieces (fig. 27). Small gap allows more precise control of the electromagnetic field of the objective lens, thus increasing the attainable resolution. However, the gap cannot be too small, because it would limit the sample holder rotation. Also, in cryo-EM the sample has to always stay at approximately 80 K, thus to avoid residual gas molecules in the column depositing on the cold surface of the sample and contaminating it, there is a so-called cryo box, mounted between two pole pieces of the objective lens. The cryo box protects the sample and works as a cold trap for contaminants.

2.3.3.5. Electron detectors.

Electron detectors are usually located at the bottom of the microscope column. The most simple ones are viewing screen and TV-camera, that can be used to observe the images in live fashion. Viewing screen is generally coated with ZnS, that emits a green light when excited by electrons hitting the screen. To be able to record images and process them in a computer, films, charge-coupled device (CCD) or direct detection device (DDD) cameras can be used. Films typically have a big size (9 by 6 cm), allowing to record large areas of interest with high resolution. They are made of small silver halide crystals embedded in gelatin, that are able to absorb electron radiation. Films have to be developed to transform silver halide into visible silver grains. For further computer analysis, films are scanned and digitized in special scanners, that use a bright focused beam of light and convert optical densities of the film into a digital image. Due to tedious and time-consuming handling of the films, they are practically not used anymore. Digital cameras, such as CCD and DDD, allow to immediately register a digital image, thus enabling automatic data collection strategies. Because high energy electrons can damage a sensitive CCD chip, CCD cameras include a scintillator layer that converts the electrons into photons, transferred further by fiber optics to the CCD chip, that convert photons back into electron counts. Main disadvantages of these cameras are backscattering (when electron can exit the scintillator layer from the wrong side, back into the column) and shot noise (when one electron due to its stochastic nature of interaction with materials can produce different number of counts depending on how much it was scattered). Both processes contribute to the noise in CCD image. However, these cameras have a high dynamic range, that is important for imaging diffraction patterns, allowing to detect weak spots and very intense spots in the same image.

DDD cameras, originally designed a decade ago, have undergone dramatic development over last few years. Their key feature is a direct detection of electrons by a camera chip, that allows electron conversion to counts in more precise, localized manner. If the direct detector is backthinned, the chances of backscattering are minimized, reducing the noise. Another strong point of these detectors is an ability to record movies instead of a single image: DDD allows to acquire hundreds of frames during short exposure time, which provides two benefits. First, each electron hitting the detector can be interpreted as a single detection event, irrespectively of number of counts produced, thus drastically reducing the shot noise. Second, during exposure the sample experiences beam-induced motion, that can be partially compensated by frame alignment, therefore producing much sharper final images.

Another important parameter of any detector is detective quantum efficiency (DQE), which is the ratio of the number of electron detected to the number of incident electrons, that effectively measures how much noise a detector adds to a recorded image. For the perfect detector the DQE is 1. The higher the DQE of a camera, the lower the noise it contributes to the image and the higher the signal to noise ratio. In reality, DQE decreases with spatial frequency (fig. 30), reaching very low levels at highest possible resolution (Nyquist frequency). Spatial frequency here refers to the level of brightness modulation that can be distinguished within an image: at low frequency only very coarse details of the image can be discerned, while at high frequency much more finer details can be recognized. Advantages and disadvantages of each detector type are summarized in Table 8.

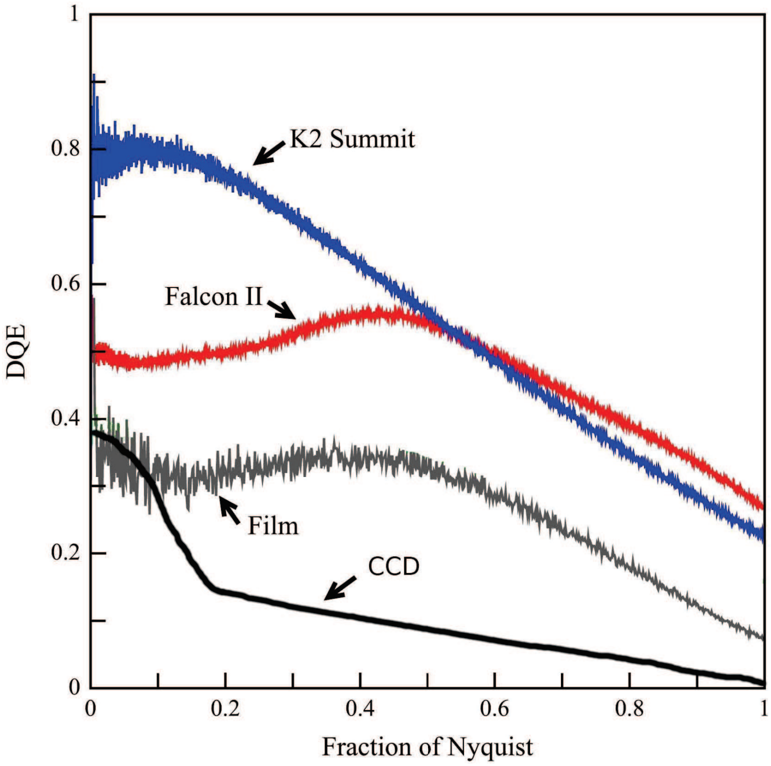

Fig. 30. Measured DQE at 300 kV as a function of spatial frequency.

Maximal possible frequency is called Nyquist. K2 Summit (Gatan) and Falcon II (FEI) are DDD detectors. Adapted from McMullan et al., 2014.

Parameter	**Film**	**CCD**	**DDD**
Image size in pixels	approx. 20.000 by 15.000	from 1000 by 1000 to 8000 by 8000	4096 by 4096 or more
Pixel size	Ag crystals are smaller than CCD pixel, but you need few crystals for 1 gray pixel	15-25 μm	5 μm (Gatan), 14 μm (FEI)
Dynamic range	10^2	$>10^4$	$>>10^4$
DQE at 1/2 Nyquist frequency	35%	10%	up to 50%

Table 8. Comparison of different types of electron detectors.

In this work, four different microscopes were used: (i) FEI/Philips CM120, equipped with LaB_6 filament and operated at 100 kV, was used for quick sample control of negatively stained specimens; (ii) FEI Tecnai F20 with a FEG running at 200 kV and a 2k x 2k CCD camera was used for acquiring the data sets of negative stain images; (iii) FEI F30 Polara equipped with a

FEG operated at either 100 or 300 kV, multi-specimen loading system and a 4k x 4k Falcon I DDD camera, was used mainly for fast screening of cryo samples; (iv) FEI Titan Krios equipped with a FEG operated at 300 kV, C3 condenser lens for constant parallel illumination, automatic sample loading system, C_s corrector and 4k x 4k Falcon II DDD camera, was used for acquisition of large high-resolution "movie" data sets of cryo-EM samples.

2.3.3.6. *Basics of image formation in an electron microscope.*

Resolution of an electron microscope images is generally limited by the sample contrast, not by the wavelength or lack of resolving power, that extends up to tens of picometers (1 pm = 10^{-12} m). For most biological specimens the resolution is 1-3 nm (1 nm = 10^{-9} m). Contrast is defined as relative intensity difference between the object and the background, that depends on nature and extent of interactions of electrons with the specimen. Most common types of interaction are summarized in fig. 31.

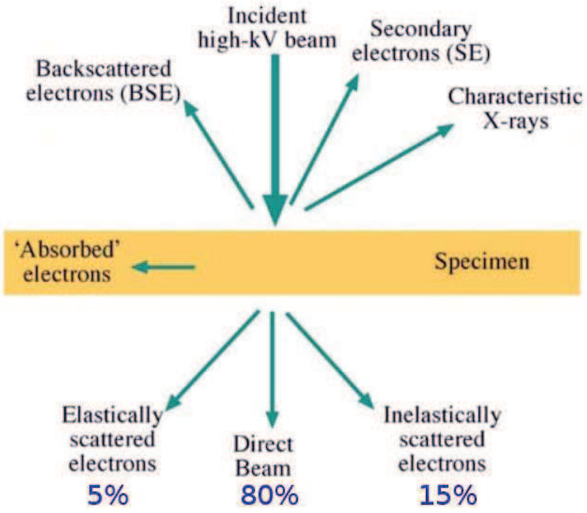

Fig. 31. Most common types of electron-specimen interaction.

In biological TEM, most of electrons go unscattered through soft matter sample (80%), while 20% are scattered. However, many electrons lose part of their energy to the sample, hence inelastic scattering. The most important signal one wants to detect in biological TEM is elastically scattered electrons, that carry information about the sample. The percents reflect only the transmitted part of the signal. From Williams and Carter, 2009.

2.3.3.6.1. Electron-specimen interaction.

In light microscopy, where the specimens are commonly stained, the main contrast mechanism is differential absorption of the photons, that produces intensity variations in the images. On the contrary, in TEM, unstained biological samples absorb a minimal portion of electron beam due to fact that they mainly consist of carbon, hydrogen, nitrogen and oxygen, that all have low mass density. Therefore, a weak contrast is a main problem in the imaging of biological samples. To be absorbed, an electron has to lose all its energy and transfer it to the sample. If an electron loses only part of its energy, the inelastic scattering occurs. In cases when electron do not lose any energy, when passing through the sample, elastic scattering takes place (or it can be simply unscattered). The nature of the two types of scattering can be explained by interaction of electron with either nucleus or electrons in the sample. During elastic scattering the incident electron passes close to the nucleus, becomes attracted to its positive charge and gets deflected by a certain angle, up to 10^{-2} rad. The heavier the nucleus (higher atomic number), the higher the scattering angle. In case of inelastic scattering, electron undergoes electrostatic interaction with sample's electrons, but it is more likely to experience a loss of energy. In latter case the electron gets deflected to smaller angle (10^{-4} rad) and most certainly pass through the objective aperture down to the detector. Inelastic scattering can lead to production of X-rays or secondary electrons, ultimately causing sample’s atom ionization, creation of free radicals and chemical bond rearrangement. Most of these processes result in radiation damage of the sample, thus decreasing the important high-resolution information. In cryo-EM of relatively thin samples (100-200 nm), for each elastic event there are 3 inelastic scattering events (Henderson, 1995).

2.3.3.6.2. Amplitude and phase contrast.

Contrast in TEM images includes two major components: amplitude and phase contrast. Amplitude contrast is more appropriately termed scattering contrast, since it arises when an electron, considered as a particle, gets scattered at high angle by the sample's heavy atoms and further removed by objective aperture or loses some energy to the sample and further being removed by an energy filter. Altogether, the number of electrons arriving at detector is decreased compared to the number of incident electrons passing through the sample. Thereby, amplitude contrast can be varied to some extent by reducing the objective aperture size or lowering the voltage (the latter leads to larger scattering angles). Amplitude contrast plays important role in negative staining technique where it rises up to 30% of total contrast (Erickson and Klug, 1971), but in cryo-EM it contributes only about 3% at 300 kV (Yonekura et al., 2006). Thus, in biological TEM the phase contrast is predominant. To understand phase contrast, or more exactly, interference contrast, one must consider electrons as waves with certain amplitude and phase. In the first approximation, electron can be viewed as a plane wave with constant amplitude and continuously oscillating phase, that is advancing downwards in the microscope column from the gun to the sample. When the wave reaches the sample, it experiences a delay in phases due to the interaction with sample's atoms and electrons. At the same time, since the matter is mainly empty space, the part of the same wave passes through the sample unscattered. At the detector level the scattered and unscattered components interfere, producing the phase contrast. However, detectors are unable to register the phase information, they can only record intensities, which are essentially squares of the amplitudes of the waves. Since the elastically scattered wave contributes only 5% of total transmitted wave (fig. 31), the amplitude of the total wave at the detector level is only slightly different from the amplitude of unscattered wave, thus significantly decreasing the probability of being detected (fig. 32b). Indeed, if one would try to image a living bacteria cell in water, it would appear as a homogeneously gray image. In order to restore the phases, containing the information about the object, we need to convert phase changes into amplitude changes to be able to detect them. This is achieved by defocusing the objective lens, as explained below.

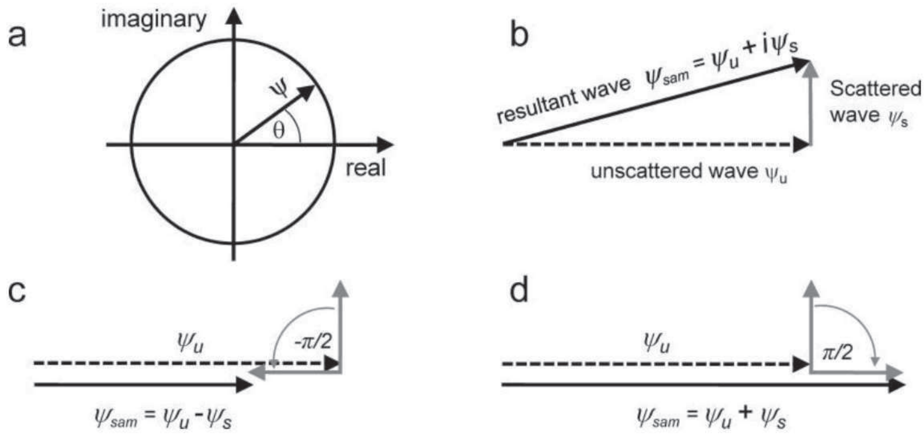

Fig. 32. Nature of phase contrast.

(a) Wave representation with amplitude ψ and phase θ on a complex plane. (b) Resultant wave ψ_{sam} that reaches the detector is a sum of unscattered wave ψ_u and scattered wave ψ_s with π/2 phase shift. Since the scattered wave represents only 5% of transmitted signal, the total amplitude of the resultant wave is almost the same as for the unscattered one, hence almost no phase contrast. (c) Additional -π/2 phase shift of the scattered wave leads to decrease in resultant wave amplitude, noticed by the detector, and produces a negative phase contrast. (d) Equivalent case as in (c), except the phase shift is π/2, producing a positive phase contrast. From Orlova and Saibil, 2011.

2.3.3.6.3. Contrast transfer function and defocusing effects.

The phase shift experienced by a scattered wave includes two components: phase shift of 90° due to the scattering process in the sample (so-called "weak phase object approximation", fig. 32b) and additional phase shift caused by the lenses of the microscope (fig. 32c, d), due to the path length differences between unscattered and scattered waves. While the first component is constant, the second one depends on the scattering angle (and defocus – see further). According to Bragg's law, the smaller the distance between scattering centers in the sample (higher resolution), the higher the scattering angle and thus longer would be the path length traveled by the scattered electron in comparison with unscattered one. Therefore, with increasing resolution, the total phase shift will "oscillate" from 0 to 360°, and consequently the amplitude of the resulting wave (unscattered + scattered) will vary from minimum to maximum value through zero. This effect is described by contrast transfer function (CTF), that effectively oscillates with increasing spatial frequency (fig. 33). CTF leads to contrast inversions from positive to negative and disappearances of certain frequencies from the image (when CTF = 0).

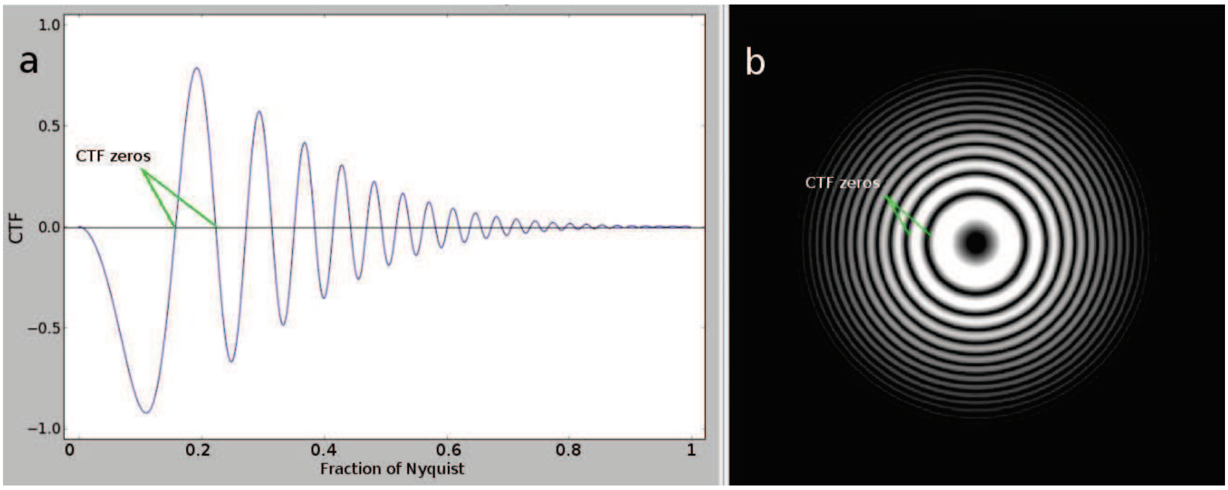

Fig. 33. Contrast transfer function (CTF).

(a) 1D graph of oscillating CTF function. Amplitude of oscillations decrease with spatial frequency due to the dampening of high resolution information. (b) 2D plot of the same function. In both images CTF zeros are shown.

When the image is in focus, we don't see anything, since the low-resolution information is lost due to the CTF (CTF starts from zero, fig. 34, red curve). However, if we defocus the objective lens, i.e. decrease its strength, its focal point will move downwards in the microscope, path length of the scattered waves will change, CTF function will shift (fig. 34, blue curve) and at the detector level we will see a defocused image. Now some frequencies in the image become visible, others would be lost. CTF is described by the formula, comprising two components: phase contrast (sin term) and amplitude contrast (cos term), which proportion is described by amplitude contrast ratio (Q).

$$CTF=[\sin\chi(\alpha)+Q(\alpha)\cos\chi(\alpha)], \quad \chi(\alpha)=\frac{2\pi}{\lambda}[-Cs\frac{\alpha^4}{4}+\Delta f\frac{\alpha^2}{2}],$$

$$\textit{where}\ \alpha-\textit{scattering angle},\chi-\textit{phase shift},Q-\textit{amplitude contrast ratio},\lambda-\textit{wavelength}$$

The CTF depends on spherical aberration of the objective lens, C_s, and defocus, Δf. Additionally, at high resolutions, CTF is dampened due to the effects of partial spatial and temporal coherence (see part 2.3.3.1), specimen drift and vibration, detector noise etc. In ideal case CTF would be always equal 1 (both Δf and C_s equal zero) – e.g., all information at all spatial frequencies would be transferred at equal level, but this is never the case in a real microscope. Therefore, to restore the information about lost spatial frequencies, one usually collect data with different defocus levels and also apply CTF correction methods to compensate for contrast oscillations in the image (see part 2.3.6.1).

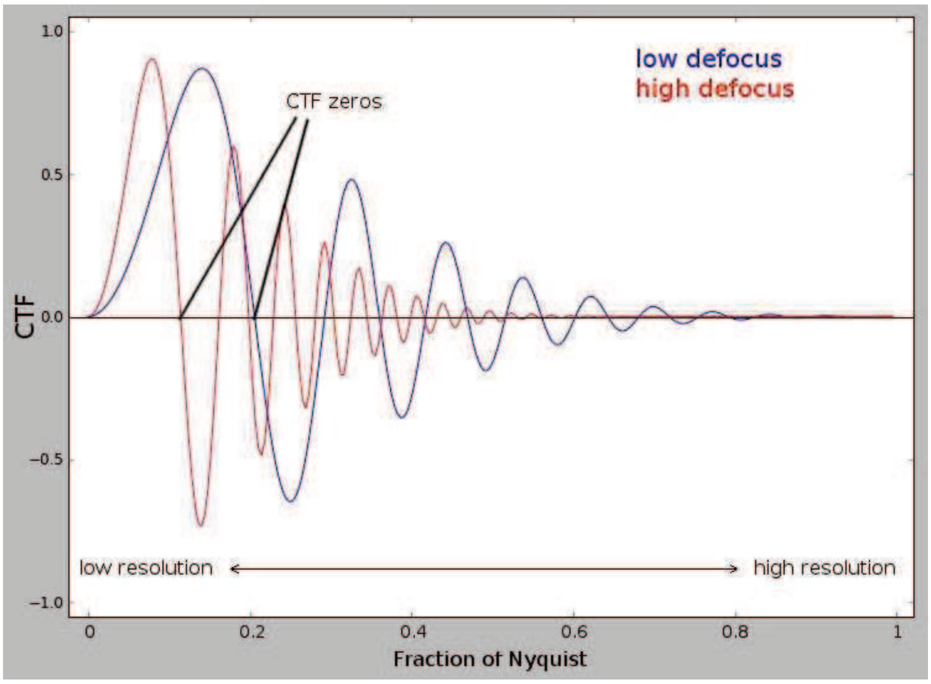

Fig. 34. Effect of defocusing on the CTF.

At low defocus (red curve), oscillations are less fast, but low resolution information is more dampened. At high defocus (blue curve) the low frequencies are stronger, hence contrast is higher, but high resolution information is more dampened.

2.3.4. Sample preparation for electron microscopy.

In most biological applications of TEM the sample is applied onto 3 mm grids, that can be made of different material. In this work, copper mesh grids with 300 grid squares per inch were used. Copper grids are commonly used because they are conductive, stable under the electron beam and inexpensive. Very often, carbon is deposited on top of the grid as an additional support for the sample. It can be a continuous carbon layer or holey carbon with holes of certain size. Continuously coated grids are commonly used in negative staining technique and the holey carbon-coated grids (“unsupported”) are normally used in cryo-EM. Sometimes, another thin layer of carbon is deposited onto holey carbon grids (“supported”), providing an additional support for the sample.

Carbon is deposited onto the grids by a process of evaporation. This is commonly done in an evacuated chamber under medium vacuum, when a sharpened carbon rod is heated by the current passing through it. At some point small patches of carbon start to evaporate from the rod and eventually deposit onto the grids. Freshly evaporated carbon is usually hydrophilic and the sample solution can be applied immediately. However, if the grids are stored for a long period, the carbon surface changes and becomes hydrophobic. In that case it becomes necessary to render the surface hydrophilic by glow discharging. Glow discharging is performed in a partially evacuated chamber, where the grid with carbon is placed between two metal plates. High voltage difference is applied to the plates, so that few molecules of air left in the chamber become ionized and these ions accelerate towards the plates. In the end, the surface of the carbon-coated grids becomes ionized, that renders it hydrophilic.

2.3.4.1. Negative staining.

As discussed previously, the biological objects produce very little contrast by themselves, since they are made of light atoms, that do not scatter electrons strong enough. An oldest contrast enhancement method used in biological TEM is negative staining (Brenner and Horne, 1959). Stains of heavy metal salts, such as uranyl acetate, uranyl formate, phosphotungstic acid and others, are commonly used, since they readily interact with the beam, strongly scatter the electrons and produce high amplitude contrast. A small drop (2-5 µl) of the sample is deposited on a carbon-coated grid, allowed to adsorb for a minute, blotted dry if necessary, and then a drop of stain is applied. After a while, the stain drop is also blotted dry. During the blotting, the stain will precipitate and surround the sample protein particles (fig. 35a). During the imaging, electrons will primarily get deflected by stain grains to high angles, most of them will be removed by the objective aperture. Thus, the areas with stain will appear dark in the image (hence, “negative” staining), while the protein will stay bright. However, it's important to understand, that negative staining reveals only the surface and the overall shape of protein molecules, it does not provide any information about the inner structure. Benefits and drawbacks of this method are summarized in Table 9.

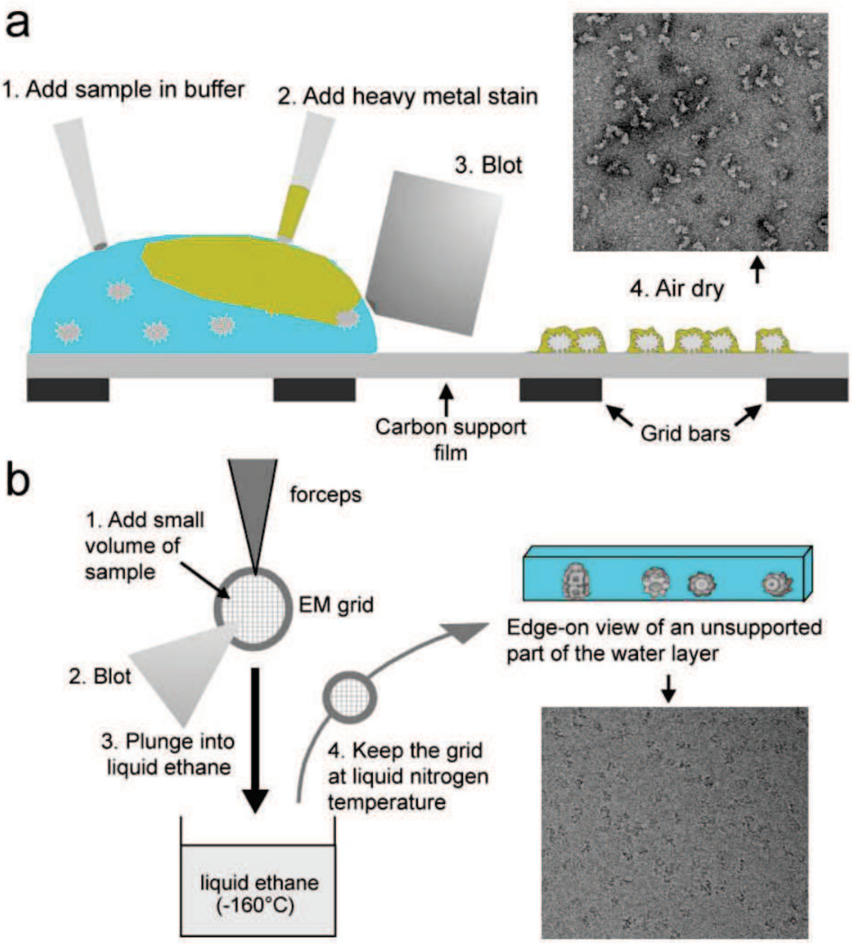

Fig. 35. Negative staining and cryo-EM sample preparation.

(a) Schematic process of negative staining and (b) sample vitrification by plunge freezing. Adapted from Orlova and Saibil, 2011.

Advantages	Disadvantages
High contrast	Particle distortions due to drying process
Radiation damage is mostly irrelevant, since it doesn't affect stain salts at the same rate as protein	Artifacts such as incomplete or uneven staining
Quick and easy to prepare	Resolution is limited to 15-20 Å, that is the size of metal salt grains

Table 9. Advantages and disadvantages of negative staining method.

2.3.4.2. Plunge freezing.

The main disadvantage of negative staining is that the sample is dehydrated during the process, so there is no guarantee that the structure of the molecules is preserved. Cryo-EM has been developed as a technique allowing to observe the molecules of the sample in near native conditions (Dubochet et al., 1988). It is based on a plunge freezing method: (i) small amount of sample (2-5 µl) is applied to a grid hold by tweezers inside a chamber with computer-controlled temperature and humidity; (ii) after a certain period of time the grid is blotted and immediately plunged into a reservoir with liquid ethane, that is cooled down by surrounding liquid nitrogen to approximately 80K; (iii) the frozen grid is transferred into a grid box kept in liquid nitrogen and then stored until the moment it is loaded into the microscope. During the plunge freezing a thin layer of protein solution, applied to the grid, is cooled down at very high rate (10^6 °C/s), so water does not crystallize but forms a vitreous ice, leaving the sample structure intact. Beside this, cryo-EM has other advantages: (i) keeping the sample always at liquid nitrogen temperature reduces the evaporation, when it is inserted into the vacuum of microscope column; (ii) molecular rearrangements caused by radiation damage are less pronounced at low temperatures, preserving the high resolution information (Hayward and Glaeser, 1979; Henderson, 1990; Stark et al., 1996); (iii) at such low temperatures the thermal stability of the sample holder is increased, thus reducing specimen drift, and providing the conditions to acquire less blurred images. Cryo-EM is usually done with two types of holey carbon grids: unsupported (fig. 35b; the protein particles, frozen inside a thin layer of ice, are imaged inside the carbon holes of the grid) and supported (protein particles are adsorbed on a thin layer of carbon, lying on the top of the holey carbon grid). The advantages and disadvantages of cryo-EM are summarized in Table 10.

Advantages	Disadvantages
The sample is best preserved, always staying in solution	Low contrast due to absence of any stains and "low dose" imaging
Preferential orientation of the particles is mostly avoided in case of unsupported grids	Less quick and more complicated than negative staining

Table 10. Advantages and disadvantages of cryo-EM method.

2.3.4.3. Protocols for preparation of electron microscopy grids.

In this work, the electron microscopy grids were prepared as follows.

2.3.4.3.1. Continuous carbon preparation.

For continuous carbon preparation, freshly cleaved mica (Euromedex, ref. 71851) was placed into the chamber of the vacuum coater (Edwards Auto 306), evacuated up to $5*10^{-6}$ mbar, and 3 A current at 10 V was passed through a sharpened carbon rod (Euromedex, ref. 70300). The duration of evaporation ranged from few seconds to few minutes, depending on the thickness of a carbon required. The next day, empty Cu/Rh 300 mesh grids (Euromedex, ref. M300-CR) or Quantifoil R2/2 grids with previously evaporated thick holey carbon were placed onto a filter paper inside a glass Petri dish filled with distilled water. Carbon was gently floated from the mica onto the water surface, then the water was slowly aspirated, leaving the carbon on the top of the grids. The grids were left to dry overnight, and kept dust-free until further use.

2.3.4.3.2. Holey carbon preparation.

Thick layer of carbon was evaporated, as illustrated before, on top of the Quantifoil Cu/Rh 300mesh R2/2 grids (ref. N1-P16NCR30-01) with manufactured holey plastic film. Next day, the plastic was dissolved during 24-48 hours by submerging the grids into ethyl acetate. Subsequently, the grids were air-dried and used for cryo-EM. Otherwise (in case of supported cryo-EM), an additional thin layer of continuous carbon was evaporated on mica and floated onto the grids, as described before.

2.3.4.3.3. Negative staining.

Cu/Rh 300 or 400 mesh grids with previously evaporated or manufactured (Ted Pella, ref. 01822-F) continuous carbon were glow-discharged in ELMO apparatus (Cordouan

Technologies). The parameters were: 70% voltage, 2 mA, 25 s in air atmosphere ($2*10^{-1}$ mbar), rendering the surface hydrophilic with predominantly negative charge. Straight away 5ul of the sample at concentration of approximately 25 µg/ml were applied on the grid and left for 1 min adsorption. Then, 50 µl of filtered (0.2 µm filter) 2% uranyl acetate (Eloise SARL, ref. 19481) were applied and left for staining for 1 min. The excess of stain was removed by blotting paper and the grids were air-dried.

2.3.4.3.4. Cryo electron microscopy.

For the unsupported cryo-EM data collection, Quantifoil grids with evaporated holey carbon and removed plastic were left under vacuum for 5-10 min and then glow-discharged for 25 s (same parameters as before). 3 µl of the sample (at approx. 0.4 µg/µl) was applied to the grid inside the chamber at 100% humidity and 4°C (Vitrobot Mark IV, FEI). The grid was automatically blotted with 595 filter paper (Ted Pella, ref. 47000-100), using blot force 8, blot time 1 s, and plunged into liquid ethane. The frozen grids were stored in liquid nitrogen until data collection.

For the supported cryo-EM data collection, Quantifoil grids with evaporated holey carbon and additional thin layer (5-10 nm) of continuous carbon were glow-discharged for 40 s (other parameters as before). Plunge freezing of 2.5 µl of sample (at approx. 0.06 µg/µl) was performed similarly as before, but with different blotting conditions: 30 s waiting before the blot, blot force 5, blot time 3 s. Temperature was set at 15°C, humidity at 100%.

2.3.5. Data acquisition.

Inherently low contrast of biological samples leads to low signal to noise ratio (SNR) of acquired images. Though it is possible to get atomic resolution information, the total dose applied to the sample would destroy it completely before an image could be acquired. Besides radiation damage, two other factors affect the attainable resolution: DQE of the detector used and beam-induced movement and charging of the specimen.

In order to minimize radiation damage, a so-called “low dose” imaging procedure has been developed (Unwin and Henderson, 1975). In that case, focusing and adjustment of illumination conditions are done with high intensity beam, few microns away from target area, and only then an exposure of the target area is taken with low electron fluence (“electron dose”), about 20 ē/Å^2. This electron fluence represents a compromise, used by the cryo-EM community, between obtaining the images with clearly visible particles and limiting the radiation damage.

Development of the direct detectors (DDD) with DQE much higher than of a film or a CCD made possible to significantly improve SNR of the acquired images. Additionally, these detectors allow to acquire “movies” with high frame rate (17 or 400 frames per second, for the current products), providing a possibility of beam-induced motion correction at later steps of image processing. Recent studies have shown (Brilot et al., 2012; Li et al., 2013) that during the exposure different types of sample movement are observed: besides the unidirectional drift caused by instabilities in the sample holder, there are complicated patterns of both global and local beam-induced particle movements. During first 1-2 frames (total fluence 3-5 ē/Å^2) large shifts and rotations of the particles have been detected (>10 Å). Later, the speed and the scale of movements decrease and the later frames are increasingly affected by the radiation damage. Thus, one of the strategies for data collection on DDD detectors is to acquire a movie with high total electron fluence (50-60 ē/Å^2) and then, at the very end of image processing, discard both first frames (with large particle movements) and the very last frames (with high accumulated damage, hence decreased resolution). In addition, the frames can be aligned right after data acquisition to correct for global specimen drift and summed up, producing a high-contrast average image that can be used for image processing.

In this work, data acquisition of negatively stained samples was done on FEI Tecnai F20 microscope at 200 kV, using CCD camera at magnification 40.000 (corresponding to a pixel size of 2.6 Å/px) in “low dose” mode. Exposure time was set to 1 s, the electron fluence – to 25 ē/Å^2. Images were taken with the defocus ranging from -1.7 μm to -2 μm. For cryo-EM data acquisition on FEI Titan Krios, operated at 300 kV, Falcon II DDD camera was used for acquiring movies with 17 frames/s, at calibrated magnification of 47.000 (pixel size 1.41 Å/px). Defocus values ranged from -1.4 μm to -3.5 μm. Movies were 1 second long, with total electron fluence 45-60 ē/Å^2. Out of 17 frames, only the frames from 2 to 8 (accumulated 20-30 ē/Å^2) and the total exposure (accumulated 45-60 ē/Å^2) were saved for further image processing.

2.3.6. Image processing.

To overcome low SNR of the images, in cases where purification of multiple copies of the same object of interest (such as protein complexes) is possible, single particle analysis (SPA) has been developed. This analysis is based on averaging thousands of low dose images, that increases final signal to noise ratio and allows to extract high-resolution information. We should not forget that transmission electron microscope image represents a 2D projection of a 3D object viewed from a certain direction. In fact, in a typical cryo-EM single particle experiment, protein molecules that are frozen in vitreous ice, exhibit different orientations, so the resolution of the final structure is limited by the precision with which we can determine these orientations. Many different workflows exist for the SPA, but the overall goal is to determine the projection direction of each 2D image with respect to 3D object and to reconstruct a 3D map of the object by combining multiple 2D images. The different steps of the workflow used in this work are summarized in fig. 36.

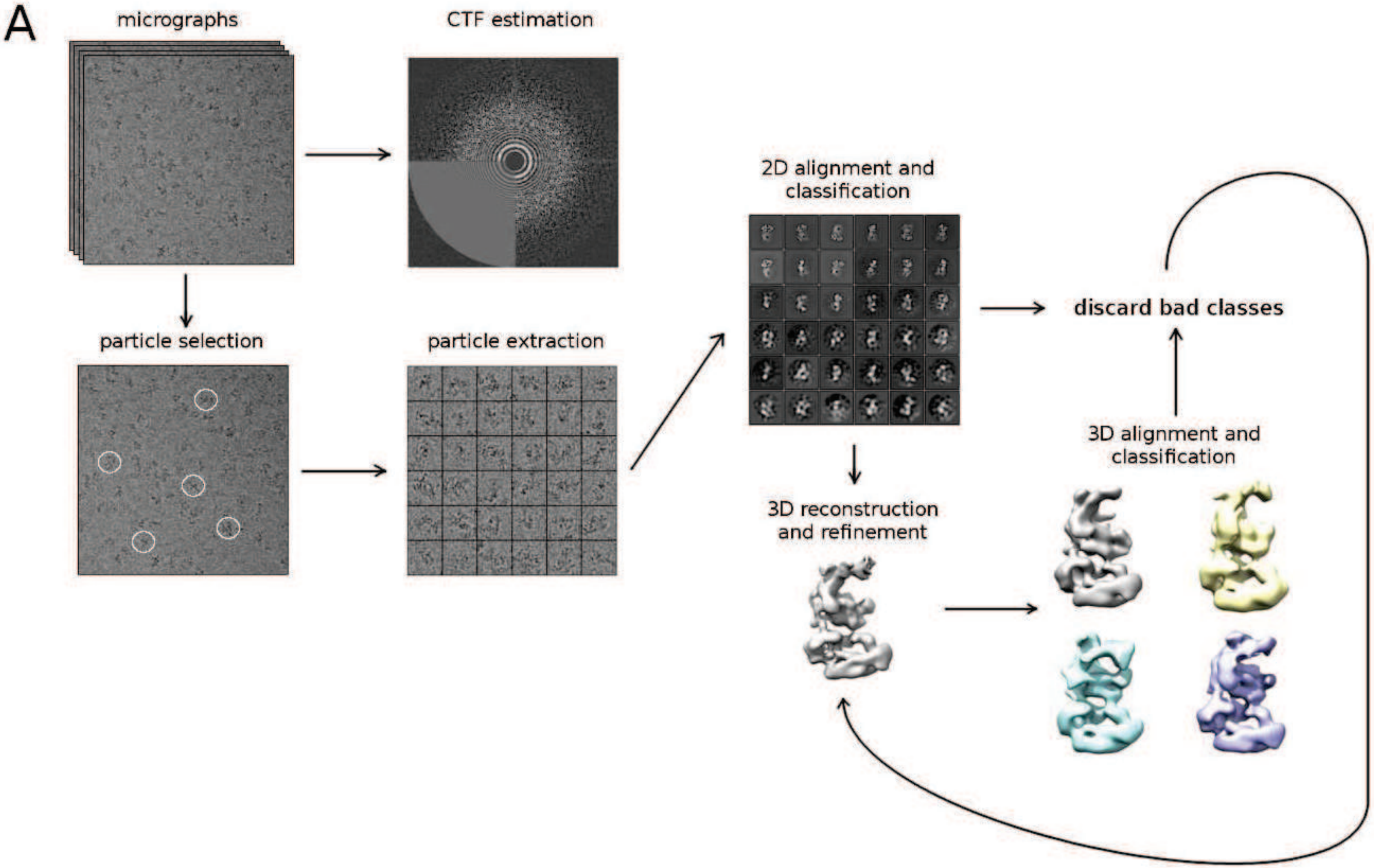

Fig. 36. Single particle image processing workflow. *(see the legend on the next page)*

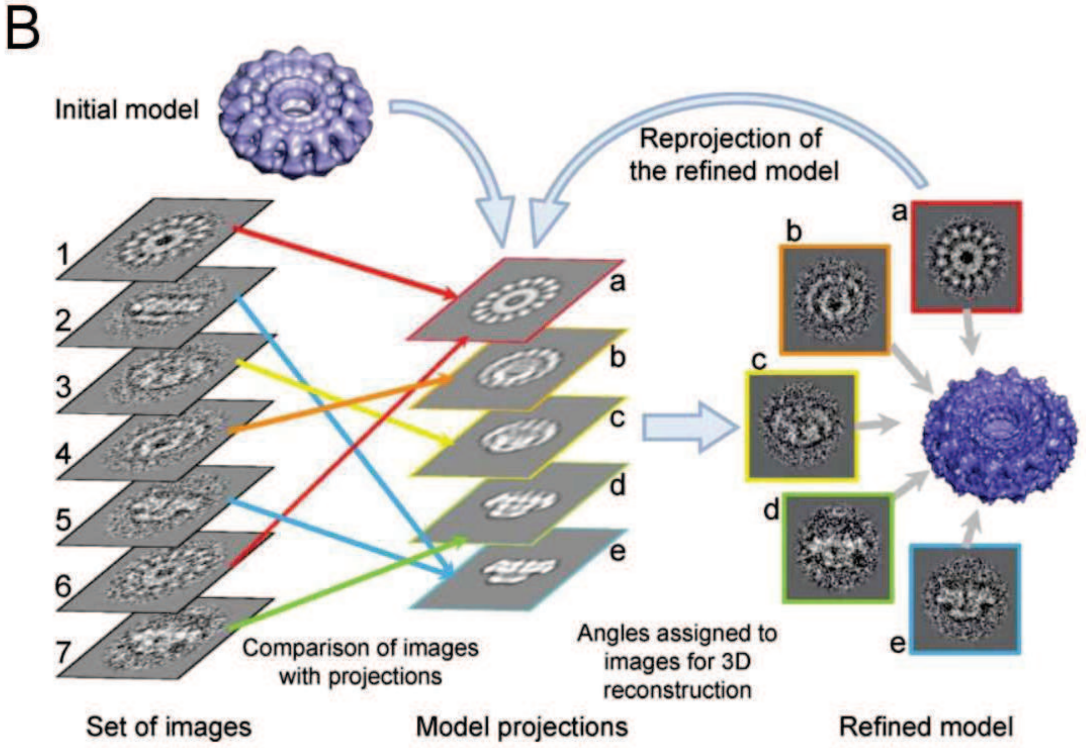

Fig. 36. Single particle image processing workflow.

(a) Different steps of SPA protocol. First, the CTF is estimated from the micrographs. Single particle images are selected and extracted, then subjected to 2D alignment and classification to sort out bad images. Afterward, good particles are used in 3D reconstruction and refinement. Additionally, 3D alignment and classification are used to sort out heterogeneity in the data: bad/unwanted structures are discarded, the best ones are further refined in 3D. (b) 3D refinement protocol in details, based on projection matching. For simplicity, only one 3D reference model is considered. Panel (b) is taken from Orlova and Saibil, 2011.

2.3.6.1. CTF determination and correction.

After the data collection one usually ends up with a set of several hundred or thousand of TEM images, or micrographs. Micrographs are then routinely screened for signs of high astigmatism, drift, vibration, sample contamination etc. and bad images are discarded. The next step is the determination of the CTF parameters, such as exact defocus and astigmatism of every micrograph. Both screening for drift or astigmatism and CTF estimation are performed by computing a power spectrum of each individual micrograph. Power spectrum represents a squared Fourier transform of an image. CTF effect can be immediately recognized by a pattern of rings (fig. 33b), called Thon rings (Thon, 1966). Alternating white and black rings represent the CTF, oscillating from positive to negative values. As shown in fig. 37, power spectrum can be used to estimate the micrograph quality. Correction of the CTF is performed at later steps, on the individual single particle images. During that procedure one should use correction of both phases (flip the phases where CTF is negative) and amplitudes (properly weight signal at all frequencies) of the CTF function.

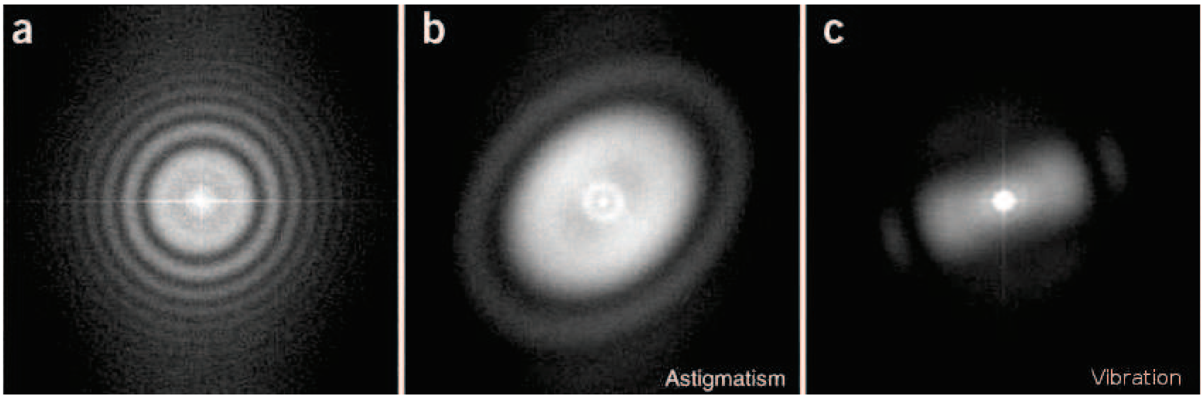

Fig. 37. Example power spectra from CTF-affected micrographs.

(a) – a normal micrograph affected by CTF, (b) – a micrograph with high astigmatism or (c) vibration. From Scheres et al., 2008.

2.3.6.2. Particle selection and extraction.

After obtaining a set of good micrographs, each single particle has to be detected and extracted from every image. Since manual particle picking of hundreds of thousands individual particles is very tedious and time-consuming, many automated algorithms have been developed. In this work, a reference-based picking was used (Hoang et al., 2013):

- (i) several strongly filtered template 2D images were in-plane rotated with 5° step;
- (ii) rotated templates were cross-correlated (CC) with all micrographs;
- (iii) CC peaks were detected;
- (iv) threshold range for “good” CC values was chosen empirically by user;
- (v) images of certain size, at the positions of these peaks, were cut out from the micrographs.

References for particle picking can be obtained either from a 2D classification (see later) or from a known low resolution 3D model. Depending on the contrast of the micrographs and the chosen threshold, many false positives are usually selected. To detect and remove false positive particles, a sorting algorithm was used (Scheres, 2015): for each extracted particle image a corresponding template in a given in-plane orientation was subtracted and the resulting difference image was analyzed by a number of statistical parameters. In case of true particle, the difference image would contain only background noise, and if a false positive was picked, the difference image would have features that cannot be described by noise alone. As a result, the sorting algorithm assigns a certain score for each particle: particles with high score are often high contrast ice contaminants, protein aggregates etc.; low-score particles are mostly noise that was wrongly correlated with a template. Nevertheless, visual inspection of all particles is necessary to assure the good quality of the picking procedure. In addition, extracted particles are thresholded (to remove very bright or very dark pixels, coming from artifacts of the detector) and normalized to average zero value and standard deviation of one. Both of these steps are important for correct correlation calculations during all further alignment steps.

2.3.6.3. Alignment of single particle images: two different approaches.

Before noisy single particle images can be averaged to improve SNR, they have: (i) to represent the same view (projection) of the same 3D object, (ii) to be aligned to each other both rotationally and translationally. Two different approaches are widely used for image alignment: maximal cross-correlation (maxCC) and maximum-likelihood (ML).

The maxCC approach, in the simplest case of 2D reference-based alignment (i.e., only in-plane rotation), includes the following steps (fig. 38a): 1) start from some initial guess about the structure: a single 2D reference – it is usually an average of randomly selected particles in random orientations; 2) for each experimental image calculate CC as a function of rotation angle; 3) find the maximum of CC for each experimental image; 4) rotate all experimental images by angle, that has the maximum CC value; 5) average aligned images; 6) use the average as a new reference and iterate all the steps. By iterating, a local maximum of CC function is approached. When SNR is high enough, correct alignment will be found and the final average will represent a closest (in least squares sense) estimate of a true structure. However, typical cryo-EM images are very noisy, with low SNR. In that case, false CC peaks may lead to errors in alignment and the algorithm will not necessarily converge to the correct structure.

ML approach was found extremely useful for alignment of noisy and/or small particles with low SNR (Sigworth, 1998). Instead of finding a single optimal orientation for each image, the algorithm calculates a weighted sum over all possible orientations. In this case, each particle will contribute to the new average not only once (as in maxCC approach, with optimal rotation), but as many times as number of sample rotation angles (fig. 38b). Each contribution will have a certain weight, that is essentially a probability of a particle to be in certain orientation. The function optimized and maximized here is called likelihood, which is the probability that a given data set of experimental images arises from a 3D map (reference) with a certain parameter set (rotations). The ML approach was found less sensitive to the choice of initial reference than the standard maxCC approach, especially in the case of the alignment of images with low SNR. It is worth to note that both maxCC and ML algorithms become equivalent at high SNR. The power of ML algorithm can be extended further for multiple-reference 2D and 3D alignments, also this approach was also found very promising when dealing with heterogeneous data sets. The only disadvantage of this algorithm is high computational cost, in comparison with relatively fast maxCC calculations. In this work, all 2D and 3D refinements were performed in Relion program (Scheres, 2012), that implements a ML approach for cryo-EM structure determination. The only exception was SAGA HAT negatively stained data set, that was analyzed in IMAGIC (image alignment based on maxCC; van Heel and Keegstra, 1981) and XMIPP (ML-based 3D alignment and classification; Scheres et al., 2009) software.

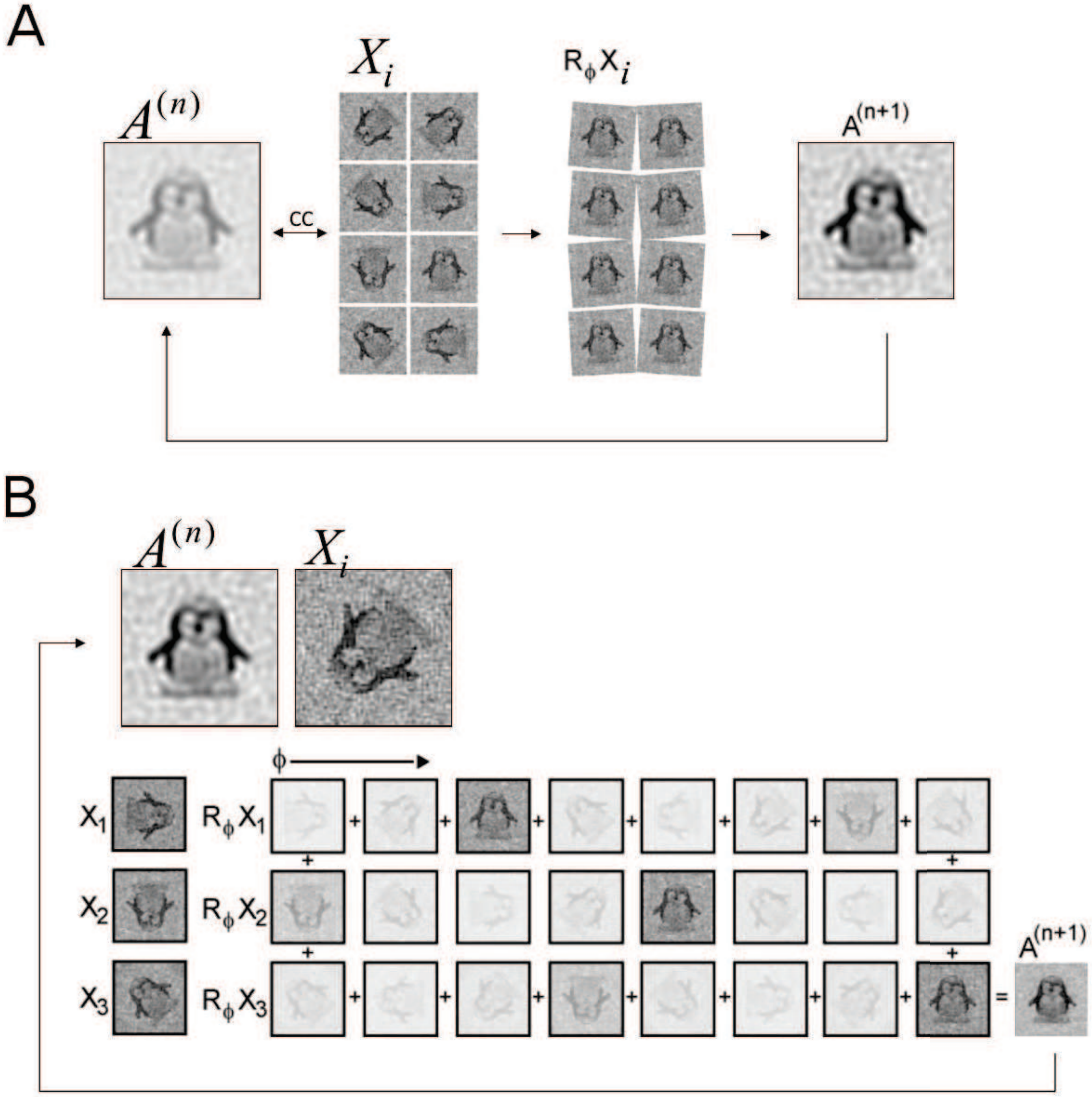

Fig. 38. Comparison of maxCC and ML approaches.

(a) In maxCC approach an initial reference $A^{(n)}$ *is cross-correlated with every experimental particle* X_i *, that is rotated by vector* R_φ *. One optimal value of* φ *is found, that corresponds to a maximal CC. Particles* X_i *are rotated by optimal angles and averaged, producing an improved reference* $A^{(n+1)}$*. (b) In ML approach an initial reference* $A^{(n)}$ *is compared to experimental particle* X_i *in statistical manner: e.g., particle* X_1 *is rotated by vector* R_φ *and all possible orientations* φ *are considered – each orientation is given a different weight, depending on likelihood between reference and rotated particle. All particles* X_i *, rotated by all possible angles, are averaged, taking their weight into account. New reference is* $A^{(n+1)}$*. Different weights are depicted as particles with different transparency level. From S. Scheres, MRC EM course 2014.*

2.3.6.4. 2D alignment and classification.

After the extracted particles are thresholded and normalized, they are usually submitted to iterative process of 2D classification. In real data set particles exhibit different orientations (either on carbon support or in ice), producing different views, that have to be distinguished. During classification, the images are aligned translationally (in X and Y directions) and rotationally (in-plane rotation) and also classified in order to separate distinct views of the object. Reference-free classification starts from generating a set of N references from randomly oriented unaligned particles, thus avoiding any bias during further alignment. The only parameter that has to be provided is a number of classes, N. The 2D classification provides powerful means to clean up the data set by throwing away bad particles that usually do not average well together. It is normally repeated several times to be sure that all bad images have been discarded from the further analysis. The class averages with high SNR can be used to assess quality and possible heterogeneity of the data set.

2.3.6.5. Initial model problem in EM.

The projection direction of each 2D experimental particle is not known a priori. However, this information is necessary to calculate a 3D map. The biggest challenge is to determine the angular assignments for a data set of an unknown 3D structure. Once the initial model is available, it can be refined. Numerous methods of initial model generation have been developed and they all can be separated into two groups: experimental methods, mostly based on tilting the specimen inside the microscope (single particle tomography (Walz et al., 1997), random conical (Radermacher, 1988) and orthogonal tilt methods (Leschziner and Nogales, 2006)); and computational methods (many different approaches, based on a common line method (Van Heel, 1987), random model generation, computer-generated shapes etc.). Details of these methods are beyond the scope of this text and described elsewhere.

In this work, we used random conical tilt method (RCT) to determine the initial model of SAGA HAT module. RCT method relies on a simple geometrical relationship between two particle views related by a known angle. Initially developed for samples adopting a preferred orientation on the support, RCT includes the following steps:

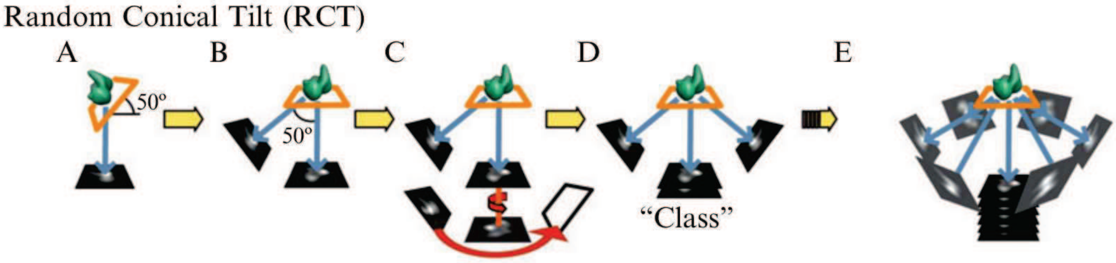

Fig. 39. The principle of RCT method.

See the details in the text. From Leschziner, 2010.

- (i) images of the same specimen area are taken at 0° and high tilt (45°-60°). First, a tilted image is taken, that will be used for 3D reconstruction (fig. 39A). The untilted image, used for alignment and classification, is therefore exposed twice (fig. 39B).
- (ii) tilt-pair particles are picked and extracted from both tilted and untilted images. In addition, the exact tilt angle and tilt axis position are determined, based on geometrical relationship between neighboring particle pairs.
- (iii) untilted particles are subjected to 2D classification and alignment, producing 2D class averages that represent certain characteristic views. The in-plane rotation angles, determined for untilted images, are used to determine relative positions of corresponding tilted particles (fig. 39C and D).
- (iv) in-plane rotation angles are combined with tilt geometry parameters determined earlier to produce a set of Euler angles (fig. 40) that is assigned to the tilted images.
- (v) 3D reconstructions are calculated from tilted particles of each "class", determined during 2D classification for untilted particle pairs (fig. 39E).

In theory, three orthogonal projections are needed to reconstruct a 3D volume. Each projection direction is described by two angles (fig. 40), in-plane rotation of an experimental image is represented by a third angle, completing the position determination in a 3D coordinate system. In case of RCT, the first two angles are determined by tilt geometry and the last one is determined during 2D alignment. Unfortunately, one cannot acquire a tilted image at 90°, therefore all RCT reconstructions are usually distorted in the direction, parallel to the incident beam. This effect is known as "missing cone" due to its shape in Fourier space. The standard method of addressing this problem is combining independent RCT reconstructions. Because missing data depends on orientation of the molecules relative to the electron beam, two independent reconstructions will be missing information in different regions of the Fourier space. Thus, after alignment and averaging of 3D reconstructions the missing cone can be compensated.

However, this problem might be further complicated by a presence of heterogeneous 3D structures in the sample. In this work, we have used XMIPP software (Scheres et al., 2009) to perform simultaneous 3D classification and alignment of RCT reconstructions, at the same time taking into account the "missing cone".

2.3.6.6. Three-dimensional reconstruction and refinement.

The main challenge of SPA is a problem of 3D reconstruction from 2D projections. For each experimental particle five parameters have to be determined: three Euler angles (fig. 40) and two in-plane translations. If no prior information is available about the structure (i.e., low resolution EM structure, X-ray structure etc.), then determination of angular orientations becomes more difficult. However, when an initial low-resolution model is known, it can be subjected to iterative projection matching (fig. 36b): (i) projections of the current model are generated over all possible orientations, (ii) each experimental image is compared with each of the projections, (iii) each image gets probability-weighted angular assignment (in case of ML approach) since projection orientations are known, (iv) all images are back-projected to reconstruct an improved 3D model.

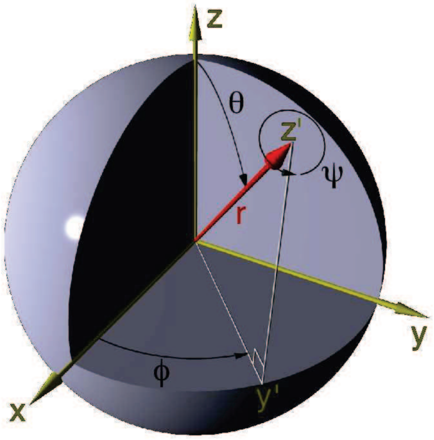

Fig. 40. Euler angles describe an orientation of the object.

Orientation is defined by three successive rotations: by φ around z-axis, followed by θ around new y'-axis, followed by ψ around new z''-axis. Resulting vector r shows final orientation in 3D space. From http://www.ebi.ac.uk

2.3.6.7. Detecting structural heterogeneity.

Most data sets are heterogeneous. Heterogeneity can be either compositional (due to aggregation or missing subunits / factors) or conformational (flexible domains etc.). The detection of a structural heterogeneity is not trivial. Already during 2D classification the problem arises: how to distinguish different views of the same 3D structure from views of a different structure? Large variability in size (i.e., bound antibodies or missing subunits) can be dealt in 2D, but it is usually not enough to separate different conformations. In 3D, the problem become even more complex: here the 2D projections can not be compared directly to each other even if they arise from the structurally unique 3D object. One group of methods to deal with heterogeneity in 3D is unsupervised classification, based on the multireference projection matching described before (fig. 36b). Similarly to the 2D case, the data set is randomly split in N parts and refined for one iteration against initial low-pass filtered model. Afterward, all N produced models are refined simultaneously for several iterations. It was shown that ML statistical procedure, applied to such iterative projection matching, is able to separate different conformations.

2.3.6.8. Resolution estimation in cryo-EM.

In electron microscopy several resolution-limiting factors exist:

- wavelength of electrons, that depends on voltage (0.037 Å for 100 keV, 0.02 Å for 300 keV electrons)
- quality of electron optics (aberration corrections, different factors that dampen high resolution)
- defocus of the objective lens: high-resolution information is dampened at high defocus values
- SNR of the data, that depends on many factors: electron fluence, radiation damage, sample composition, detector used etc.
- alignment accuracy, that ultimately depends on SNR at low and intermediate resolution

In single particle cryo-EM there is no “external” standard, that can be used to evaluate the resolution, therefore only internal consistency of the results can be evaluated. One of the resolution measures currently used in the field is Fourier shell correlation (FSC), that is a function plotted versus spatial frequency. FSC values are CC coefficients computed between the

Fourier transforms of two 3D volumes over the shells of equal resolution. FSC shows how well two 3D maps correlate with each other at a certain resolution level.

Over the years it was noticed, that the iterative refinement with single 3D map is prone to overrefinement, when the bias towards noise in the map may inflate resolution estimates. Overrefined structures often have "hairy density", when high-resolution features are superimposed on a low-resolution ghost of the initial model. To prevent overfitting, a so-called gold-standard refinement scheme is used nowadays (Scheres and Chen, 2012): (i) an initial reference is strongly low-pass filtered, (ii) the data set is randomly split in two halves, (iii) both halves are refined separately, (iv) FSC is calculated between two EM maps arising from each half, (v) if refinement has converged, both halves of the data set are combined, a single 3D map is reconstructed and filtered to the resolution, estimated before with FSC. If the refinement has not converged, both maps from each half data set are filtered according to FSC and the next cycle of refinement is started.

Typically, FSC is calculated between two completely independent maps arising from two halves of the data set, and the target resolution is where FSC drops below 0.143 (Rosenthal and Henderson, 2003). However, FSC provides only a single number that describes the overall resolution of the map. In practice, due to heterogeneity and image processing errors, the local resolution of the maps may vary (Kucukelbir et al., 2014).

2.3.6.9. 3D map validation.

Every 3D EM map has some uncertainty. Therefore, it is essential to provide an assessment of map errors, that can arise from incorrect initial model, wrong hand determination, alignment errors, overrefinement etc. In last few years, the cryo-EM community came up with a list of recommendations (Henderson et al., 2012) to estimate both map validity in general and reported resolution. They can be summarized as follows:

- ensure that you can see particles in the raw micrographs. This is especially important for small proteins, that have intrinsically low contrast. Every automated picking program produce false positives, thus a manual inspection of particles is necessary
- resolution below 10 Å can be roughly estimated by simply looking at the map: at 6-8 Å protein α-helices should be visible, at the resolution below 5 Å one should see β-sheets
- absolute hand and angular orientation determination (up to the level of 15-20 Å) can be validated by tilt-pair validation test (Rosenthal and Henderson, 2003; Henderson et al.,

2011), where few tilt pair micrographs (0° and 10°) are collected and particle pairs are picked. Then projection matching is used to determine the angles of untilted particles using the 3D map. Application of tilt transformation to the orientations of untilted particles gives predictable orientations for tilted ones. If the map has the correct hand, its projection will agree with the tilted image after rotating by the tilt angle used in the experiment.

- always use gold-standard refinement and FSC calculation for two half-sets refined independently
- high-resolution noise substitution (Chen et al., 2013) allows to estimate amount of signal and noise from the overfitting present in a 3D map. In this procedure, the information beyond a certain resolution (usually where FSC drops below 0.75) is substituted by random noise and a new FSC curve is calculated, that is then compared to original FSC of the 3D map. Afterwards, an unbiased FSC can be calculated from two curves.

The two last algorithms are fully implemented in Relion software, used in this work.

2.3.6.10. Image processing protocol used in this work.

A common workflow for image analysis was used for all negative stain and cryo-EM data (the only exception is SAGA HAT RCT data set). Cryo-EM data set was collected only in “movie” mode, where the frames 2-8 (out of 17) and the total 1 s exposure (all frames summed together without alignment) were saved in single stacks. All images in the stack were thresholded, normalized and aligned to each other using Motioncorr program (Li et al., 2013). This allowed to correct for specimen drift and global particle movement during movie acquisition. The aligned stack was summed resulting in a new total sum image (high-dose image with high contrast and corrected beam-induced global drift), that was used further throughout all steps until movie processing of individual single particles in Relion.

After CTF estimation, all micrographs were screened for signs of high astigmatism or drift and bad images were discarded. Automated particle picking was performed in gEMpicker (Hoang et al., 2013) using the 2D class averages from a small, manually picked data set. Extracted particles were coarsened 2 or 4 times to minimize computational costs of the first steps of the refinement. Afterward, the particles were sorted in Relion (see section 2.3.6.2) or/and by several consecutive 2D classification runs. After each round, the particles within the classes, displaying high-resolution features, were selected. All the following steps were performed in

Relion software, including: 3D refinement using a known low-resolution structure of SAGA (Wu et al., 2004; Durand et al., 2014) as a reference; 3D classification to sort out structural heterogeneity (see Results for specific data sets). Again, the best 3D classes (displaying recognizable features) were selected and subjected to further cycles of 3D refinement. Resolution was estimated with FSC criteria for two independent halves of the data sets (at FSC = 0.143), according to the gold standard, as implemented in Relion. Final structures were masked with soft Gaussian mask and filtered to the final estimated resolution.

Movie processing of individual single particles in Relion was done at the last stage of the refinement only for cryo-EM data and included two steps: (i) translational alignment of the particles between each frame (only frames 2-8, total exposure not included) in movie stack, that led to improved "average particles" from all frames; (ii) particle-polishing procedure, that is used to account for substantial amount of noise present in single frames during the previous movie alignment. The latter procedure, in turn, consists of three steps:

- linear movement tracks for per-frame particles are fitted from previous translational alignment. Multiple neighboring particles are considered simultaneously, increasing robustness for the noise present in single frames
- individual reconstructions are calculated from each movie frame (2-8), FSC is calculated between these reconstructions and used to devise a resolution and dose-dependent weighting scheme for each frame. This accounts for increasing dose and decreasing high-resolution signal for each following frame
- particles from all frames are summed together, with both movement tracks and frame weighting scheme taken into account

The output is "shiny" particles with increased SNR, that are further submitted to 3D refinement, yielding a structure with improved resolution.

Most important parameters of the data sets acquired and processed in this work are summarized in Table 11.

Data set	**Microscope and voltage**	**Nominal magnification**	**Pixel size, Å/px**	**Particle box size, px**	**Number of micrographs**	**Defocus range (negative)**	**Number of picked particles**	**Final number of particles**	**Resolution**
Human SAGA HAT (neg. stain)	Tecnai F20 200 kV	40.000 62.000	2.6 1.7	100 200	220 tilt pairs 2952	1 – 2 µm	6813 23.787	6006 14.003	40 Å
P. pastoris SAGA (neg. stain)	Tecnai F20 200 kV	40.000	2.6	192	900	0.8 – 1.7 µm	58.024	32.803	26 Å
P. pastoris SAGA (cryo-EM)	Titan Krios 300 kV	47.000	1.41	400	2326[1] 5854[2]	1.4 – 3.5 µm[1] 1 – 3 µm[2]	77.318[1] 528.945[2]	48.464[1] N/A[2]	18 Å* N/A[2]
P. pastoris SAGA-TBP (neg. stain)	Tecnai F20 200 kV	40.000	2.6	192	1250	1 – 1.9 µm	108.876	38.678	26 Å
S. cerevisiae SAGAΔAda2 (neg. stain)	Tecnai F20 200 kV	62.000	1.71	320	3600	0.6 – 1.6 µm	59.221	19.284	32 Å

Table 11. Summary of all EM data sets acquired and processed in this work.

*[1] supported cryo-EM, [2]unsupported cryo-EM. * Resolution for the full SAGA complex. N/A – refinement in progress.*

2.4. Mass spectrometry and cross-linking.

2.4.1. The principle of mass spectrometry.

Mass spectrometry (MS) is an analytical technique that allows to identify and quantify sample composition via ion separation by their mass-to-charge ratio (m/z). Today it represents a complementary approach to a structural biology tool, providing the information about stoichiometry, subunit interaction and organization of cellular complexes. A typical MS experiment for proteomic analysis includes the following steps (fig. 41): (a) sample ionization, (b) mass analysis and (c) ion detection.

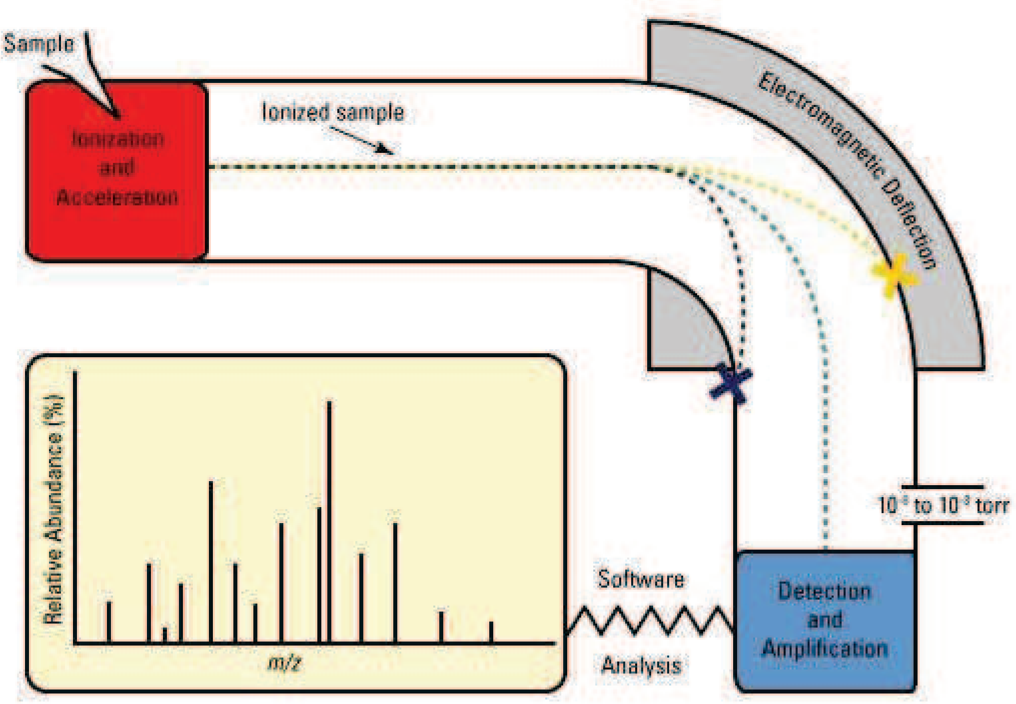

Fig. 41. Diagram of a simple mass spectrometer.

After the sample is ionized, the ions are separated by mass and charge via electromagnetic deflection and then detected and amplified. The system is kept under high vacuum to avoid ion collisions with molecules of the air. From http://lifetechnologies.com

The two most popular techniques for sample ionization are electrospray ionization (ESI) and matrix-assisted laser desorption/ionization (MALDI). ESI produces ions using electrospray, when a high voltage is applied to the liquid to create an aerosol. Advantage of this method is that it can produce multiply charged ions (+2, +3 etc.), extending the mass range for ion detection (kDa – MDa), effectively allowing to analyze both very small and very large molecules often present in protein complexes. However, this is a “soft” ionization technique, so most of the molecules do not fragment completely, hence very little structural information can be gained from a simple mass spectrum obtained. Therefore, multiple rounds of sample fragmentation are normally used (see below). In MALDI, the sample is mixed in a drop with a UV-absorbing crystalline matrix material and applied to a metal plate. The plate is placed inside the mass spectrometer under vacuum and irradiated by UV laser, causing the matrix molecules to absorb radiation, heat and cause ablation and ionization of the sample. Generally, MALDI generates single charged ions (+1), which simplifies the analysis but making it difficult to detect big protein ions.

After ionization, charged molecules are accelerated using electrical and/or magnetic field and then separated by their m/z ratio: heavy ions with small charge (high m/z) are deflected the least by the magnetic field, and ions with low m/z are most deflected. Afterward, the deflected ions hit a detection plate, producing a tiny electrical current that is amplified by the detector. Mass spectrometer is connected to a computer with a software, that analyzes detected data and produces graphs showing the ions organized by their detected abundance and m/z ratio. This data is further processed through databases, allowing to predict molecule identity by its m/z ratio.

2.4.2. Tandem mass spectrometry.

Usually, a single mass analyzer can isolate one peptide from many entering a mass spectrometer. Tandem mass spectrometry, where multiple rounds of fragmentation and mass analysis are performed sequentially, allows a direct sequencing of peptides. MS/MS employs two stages of mass analysis in order to analyze selectively the fragmentation of particular ions in a complex mixture. It has become particularly useful in “bottom-up” proteomics approach, when prior to MS analysis, proteins are subjected to proteolytic digestion (by proteases, such as trypsin). The proteins might be first separated by gel electrophoresis, either one or two-dimensional, then the protein bands are excised from the gel and digested. Alternatively, the digestion can be performed directly in a crude protein extract. In both cases, a complex mixture of protein peptides is separated by reversed-phase liquid chromatography system, directly coupled to a mass spectrometer, that performs MS/MS analysis.

2.4.3. MS/MS analysis of *S. cerevisiae* SAGAΔAda2 complex.

In this work, MS analysis of yeast SAGAΔAda2 was done as follows. After running a one-dimensional SDS-PAGE, the gel was silver stained and protein bands were excised. In-gel reduction, alkylation and destaining was performed with acetonitrile (ACN). Protein digestion was done with endoproteinase LysC (Roche) and trypsin. Digested peptides were desalted and further separated on C18 spin column using nanoRSLC U3000 liquid chromatography system, coupled to LTQ Orbitrap Elite mass spectrometer (Thermo Scientific) with a nano-ESI source. MS/MS data was treated using Swissprot database (*S. cerevisiae* 2014-05). The peptides were filtered with 1% FDR, rank 1 and minimum peptide length of 6 amino acids.

2.4.4. Chemical cross-linking in combination with MS.

Mass spectrometry is able to identify and quantify even small sequence modifications of the proteins. Such high-resolution is particularly useful for protein interaction studies, when MS is combined with cross-linking (Rappsilber, 2011). Chemical cross-linking converts non-covalent protein interactions or simply their proximity into covalent bonds. These artificially created cross-links can be efficiently identified by MS and reveal precisely which amino acid residues are involved in a cross-link. Such information can provide a number of spatial restraints that can be further used for generating residue level topology data. This technique has been successfully applied to study numerous protein-ligand (Müller and Sinz, 2012; Schwarz et al., 2013) and multiprotein complexes (Murakami et al., 2013; Han et al., 2014; Knutson et al., 2014; Greber et al., 2015).

A typical workflow for a cross-linking/MS experiment includes the following steps: proteins are cross-linked in a solution and then digested to give a mixture of peptides, some of which will be cross-linked. This mixture is analyzed by LC-MS/MS, cross-linked peptides are identified and the linked residues are detected. Most commonly used cross-linkers contain two reactive groups (fig. 42), separated by a spacer. Reactive groups can target primary amino groups found in the side chains of lysines, that are highly frequent in many proteins. Reactivity and selectivity of the cross-linkers may vary, defining the success of cross-linking reaction; spacer length determines the maximum distance between two cross-linked residues. Identification of cross-linked peptides is still very challenging, since they are often present in very low concentrations compared to unmodified peptides. Multiple cross-links or, on the contrary, an incomplete cross-linking reaction further complicate the analysis. Many different approaches were proposed to facilitate the detection of cross-links. In this work, an isotope labeled cross-linking (Müller et al., 2001) was used, where the cross-linking reaction was performed with a mixture of light and heavy forms of cross-linker reagent. This led to identification of specific doublets in the spectra of MS and MS/MS, thus significantly simplifying cross-link identification.

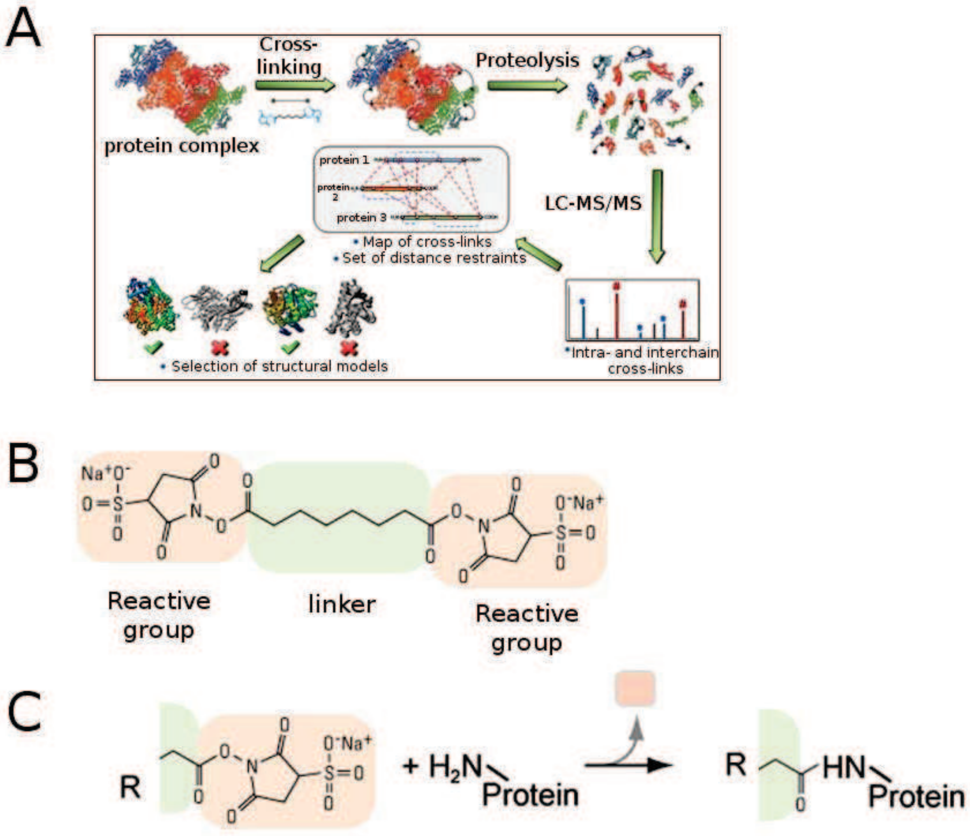

Fig. 42. Workflow of a cross-linking/MS experiment.

(a) Outline of the experiment: the target complex is cross-linked, digested and subjected to liquid chromatography coupled to a tandem mass spectrometer. The fragmentation spectra of all peptides are subjected to database search for cross-links identification. Identified cross-links are then used for building a topology model of the protein complex, based on distance restraints. (b) Bis[sulfosuccinimidyl]suberate (BS3) cross-linker, used in this study, has two reactive groups, separated by a spacer. (c) The reactive groups target mostly primary amines of lysine amino acids. Adapted from *http://daltonlab.iqm.unicamp.br* *and Rappsilber, 2011.*

2.4.5. Cross-linking/MS analysis and homology modeling of human SAGA HAT subcomplex.

The details of this analysis are described in the paper accompanying this work (page 131).

Chapter 3. Results.

3.1. Architecture of SAGA HAT module and its localization in *S. cerevisiae* SAGA complex.

To determine the molecular organization of the HAT module and its position within the full SAGA complex, we have first ensured that the HAT subcomplex can be assembled *in vitro*, when recombinantly expressed, and then purified the HAT from baculovirus-infected insect cells. We have analyzed the HAT subunit interaction network by cross-linking coupled to MS and built a 3D model of the module using homology modeling and rigid body docking methods. Furthermore, we have studied the structure of the HAT by electron microscopy and localized the module within the complete *S. cerevisiae* SAGA complex by analyzing the mutant, in which the HAT module was removed.

3.1.1. Pull-down assays of different SAGA HAT subunits.

Application of eukaryotic protein expression system gives several advantages over the standard *E. coli* system, especially for the production of human proteins: unlimited molecular size, presence of eukaryotic chaperones for proper folding, specific post-translational modifications. For recombinant expression of SAGA HAT module we have used insect cells to produce human HAT subunits (see part 2.1). Before the construction of the baculovirus for recombinant subcomplex production, to ensure that HAT can be assembled *in vitro*, several co-expression tests were done in Imre Berger's lab by Simon Trowitzsch. In the first series of experiments, histidine-tagged Gcn5 was expressed in isolation or with different combinations of ADA2B, ADA3 and SGF29. The resulting subcomplexes were pulled-down by a nickel chelating column as described in Methods. ADA2B and ADA3 subunits were shown to significantly incorporate into the GCN5-subcomplex only when both were expressed together (fig. 43a, lane 6). SGF29 did not seem to co-express with other subunits on its own, it was integrated into the subcomplex only when other 3 subunits were expressed together (fig. 43a, lane 7).

In a second experiment, the His tag was placed on SGF29 and same series of pull-downs was performed (fig. 43b). ADA3 co-expressed with SGF29 (fig. 43b, lane 3), but not with the other subunits. GCN5 incorporated only into the trimeric complex of ADA2/ADA3/SGF29. Looking at the intensity of the bands (fig. 43b, lane 7), His-SGF29 seemed to be significantly overexpressed in comparison with other subunits, thus it was decided to use the His tag on GCN5 due to apparently more stoichiometric HAT composition.

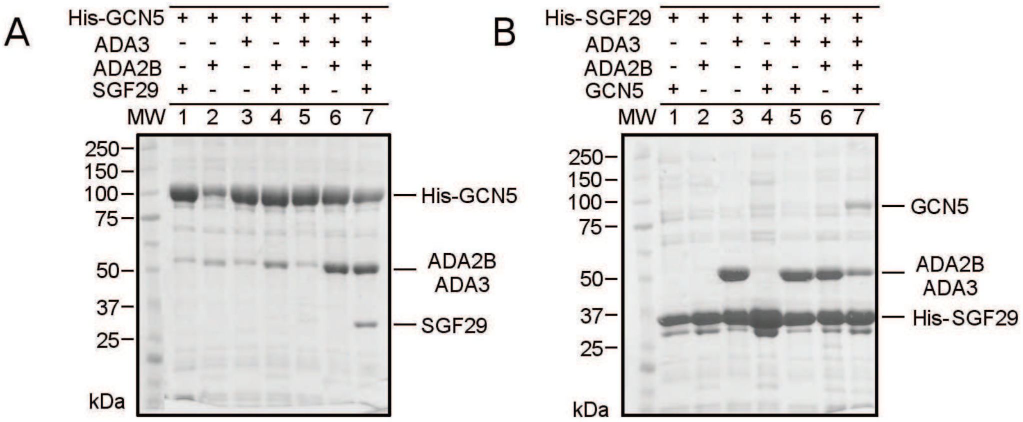

Fig. 43. His tag pull down assays.

To confirm that SAGA HAT can form a stable recombinant subcomplex, His tag was placed on GCN5 (a) or SGF29 (b) and a series of pull-down experiments was performed with different subunit combinations.

3.1.2. Production of recombinant SAGA HAT in insect cells.

The HAT subcomplex, comprised of 4 subunits (hGCN5, hADA2B, mADA3, hSGF29) was expressed from a single recombinant baculovirus using the Multibac system (see part 2.1.1). Plasmids for each subunit were kindly provided by Laszlo Tora's group (IGBMC, France). The mouse ADA3 (mADA3) sequence is 99.3% identical to its human homologue with the exception of three mutations: T110A, V126L, I226M. The multigene construct and initial virus were assembled in the group of Imre Berger (EMBL, France) by Simon Trowitzsch. DH10EMBacY *E. coli* strain, containing the Multibac bacmid with a YFP reporter gene co-expressed with the bacmid, was used for transposition. All subsequent steps of cell infection, protein production and purification were done in our lab. We have followed the expression of the subcomplex in insect cells after viral infection by several parameters: insect cell size and shape, number of cells and YFP fluorescence (fig. 44). The infection was successful when cells showed an increase in size and a round shape. Fluorescence of YFP, co-expressed together with HAT, has increased as more cells became infected. Different times of harvesting after infection were tested. Protein production was checked every 24 hours after infection by Bradford assay and SDS-PAGE of the following sample aliquots (fig. 45): whole cell extract after lysis and soluble extract fraction (supernatant after clearing by centrifugation). After 72 hours cell shape started to deteriorate, their average size decreased, suggesting that cell lysis is taking place. Protein amount was maximal at this point, thus the culture was stopped and harvested.

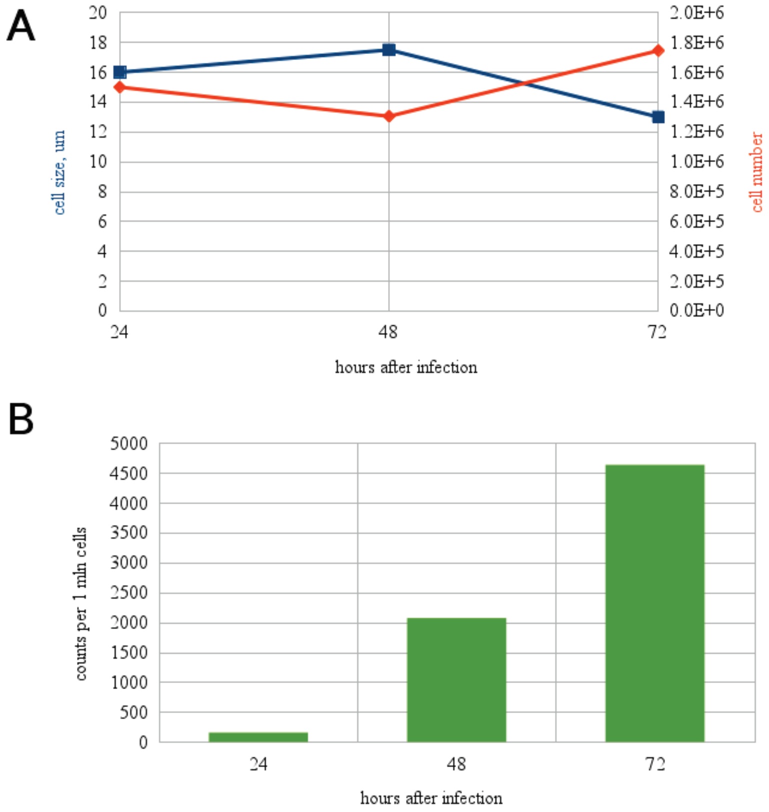

Fig. 44. Estimation of cell parameters after infection.

*(a) After successful infection, cell size normally increases until the cells start to lyse and die. Cells were counted in order to maintain the culture below 1.5*10^6 cells, otherwise the culture was diluted with the media. (b) YFP fluorescence was counted every day after infection and also followed by visual inspection in a light microscope. After 72 h all cells had YFP color and since they have started to lyse, the culture was stopped.*

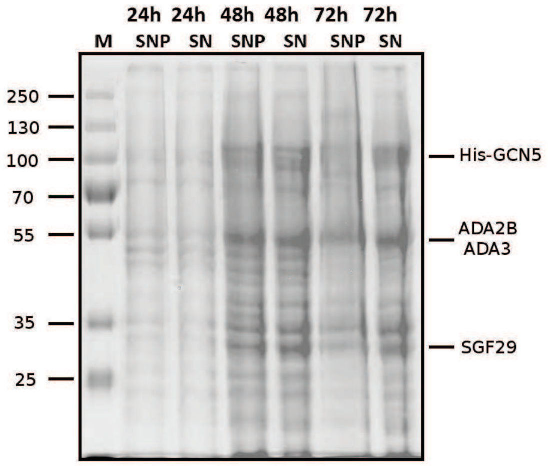

Fig. 45. Expression test of SAGA HAT after insect cells infection.

Every 24 h two aliquots were checked on the SDS-PAGE gel: M – molecular weight marker, SNP – whole cell extract after lysis, SN – soluble part of the whole cell extract.

3.1.3. Purification of recombinant SAGA HAT subcomplex.

The original purification protocol was developed by Simon Trowitzsch in Imre Berger's group. However, this purification yielded only 0.1-0.5 mg/ml of SAGA HAT of low quality (fig. 46). Several changes and optimizations discussed below were implemented to increase both quality and quantity of purified SAGA HAT.

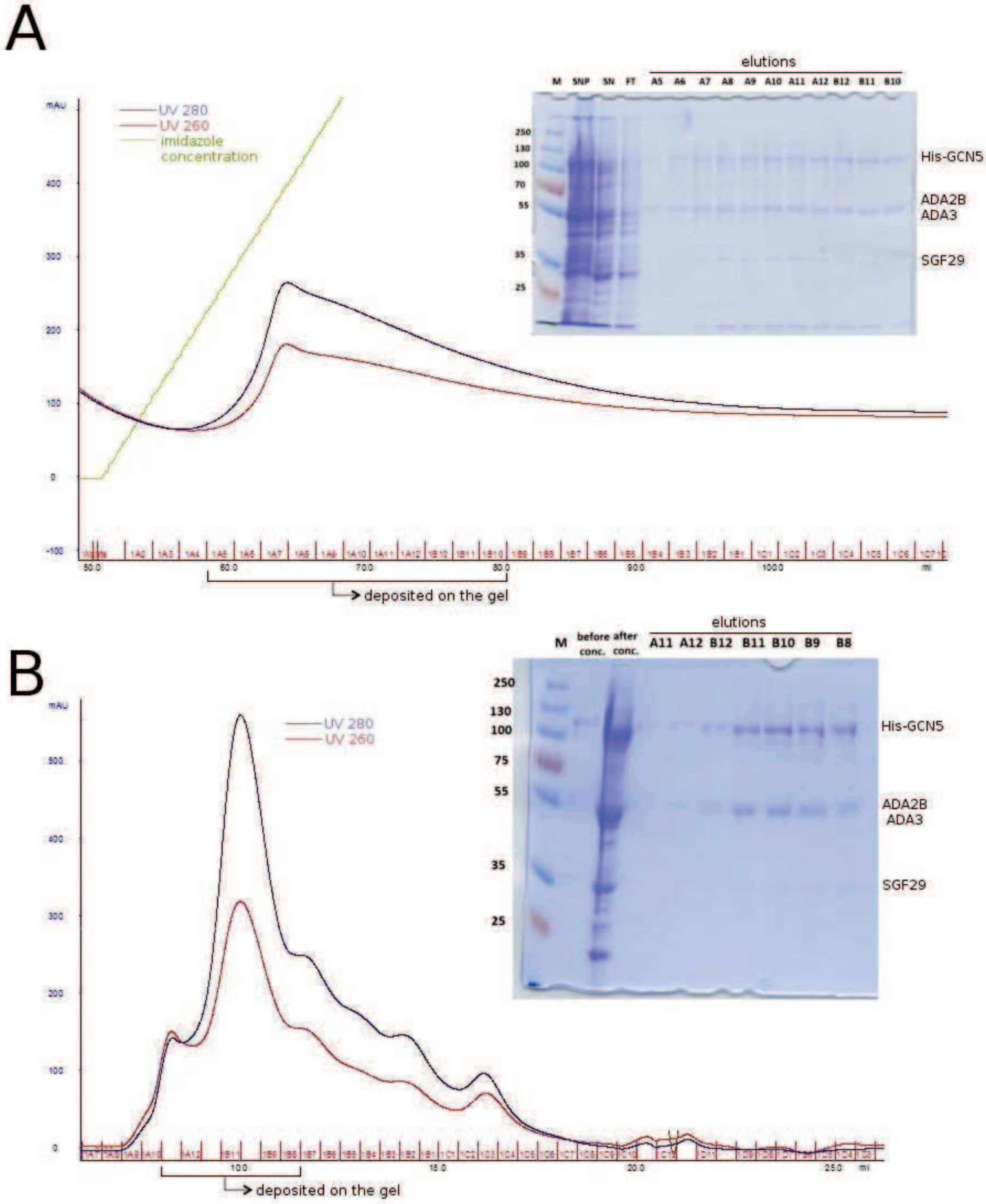

Fig. 46. Chromatograms and SDS-PAGE gels of the first SAGA HAT purification, performed according to S. Trowitzsch protocol.

(a) Affinity chromatography with TALON (Co^{2+}) resin binding His tag on GCN5. Elution displays a wide peak and indicates inefficient binding of the subcomplex. (b) Gel filtration of pooled and concentrated fractions from the previous step. Multiple peaks are observed during elution, indicating many co-purified contaminants and/or aggregates. M – molecular weight marker, SNP – total cell extract, SN – soluble cell extract, FT – unbound fraction.

3.1.3.1. Cell lysis optimization.

To determine the optimal time of sonication used for cell lysis, the total protein extract concentration was measured. Every minute during the lysis an aliquot was taken and checked using the Bradford method. Full cell lysis was achieved after 5 min (Table 12), thus 5-6 min sonication time was further used to ensure complete cell lysis.

Lysis time, min	1	2	3	4	5	6
Concentration, mg/ml	12.25	13.00	14.35	16.63	20.48	19.32

Table 12. Total protein concentration estimated by Bradford in whole cell extracts.

3.1.3.2. Metal affinity chromatography (IMAC).

In the original protocol, the first step of SAGA HAT purification was done with Co^{2+}-containing TALON resin: after binding of His-tagged subcomplex and several washes, the elution was performed with an imidazole gradient (5-250 mM). Such elution displayed a relatively wide peak on the chromatogram (fig. 46a). The cobalt resin is known to have a low affinity but relatively specific binding of histidine, reducing the presence of contaminating proteins in the eluate. In order to increase histidine binding, a nickel resin with lower specificity but higher affinity was chosen over the cobalt as a compromise. At first, the elution with imidazole gradient from the nickel column has revealed the presence of contaminants of high molecular weight (fig. 47a). To reduce this contamination, instead of a gradual increase of imidazole concentration during the elution, a step-by-step elution with extensive washes was applied, and every step was checked by SDS-PAGE. This procedure revealed that SAGA HAT is eluted only at the highest imidazole concentration (250 mM) with much less contaminants, revealing a sharp and strong peak on the chromatogram (fig. 47b). The total protein amount after nickel affinity column, measured using the Bradford method, increased by 3 times (comparing to typical yield after cobalt column), reaching 10-15 mg.

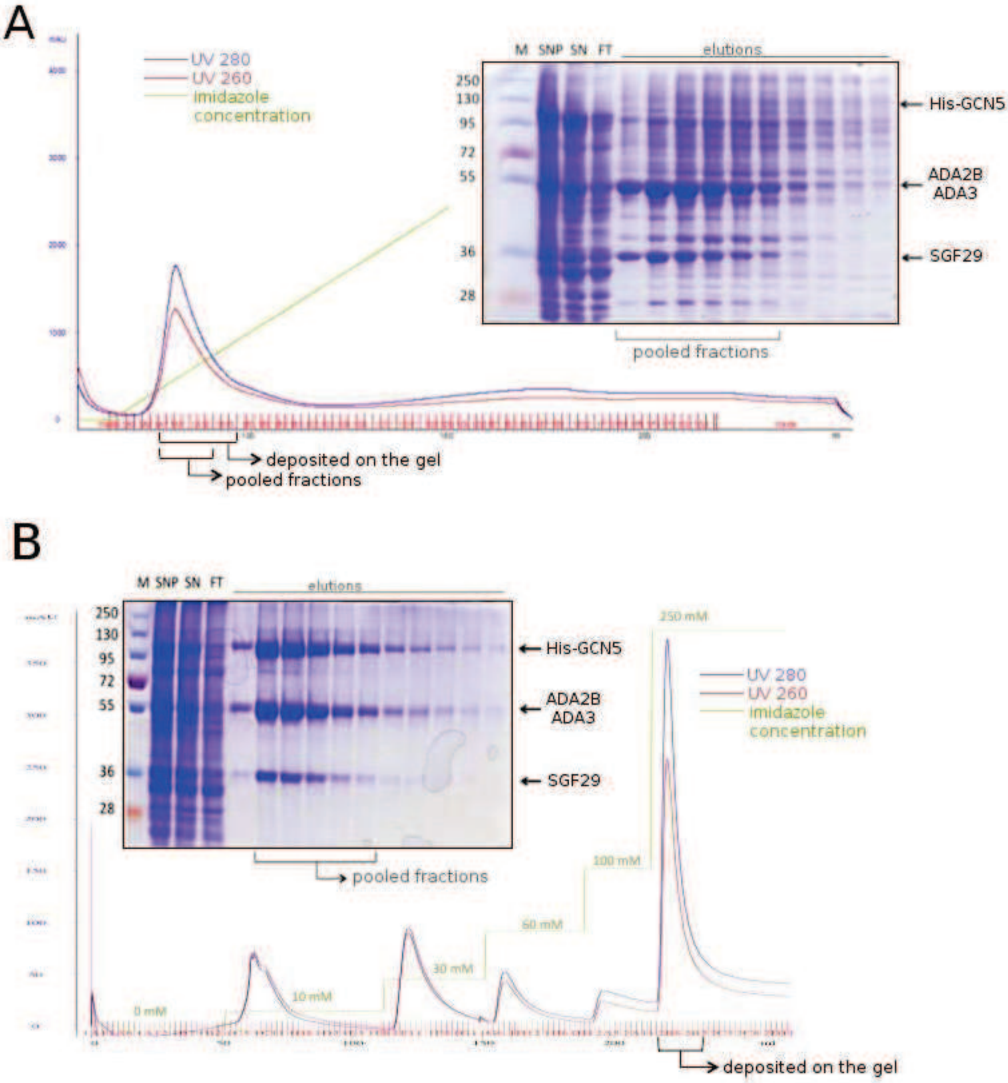

Fig. 47. Comparison of the nickel affinity purification with linear gradient elution or step-by-step gradient elution.

(a) The linear gradient leads to many contaminants present in the elution. (b) The step-by-step gradient elution with extensive washes leads to the clean elution of SAGA HAT subcomplex, as seen on the gel. M - molecular weight marker, SNP - total cell extract, SN - soluble cell extract, FT - unbound fraction.

In the original protocol, the purification buffer (50 mM Tris-HCl, 400 mM NaCl) was optimized only for the cobalt resin. However, the amino groups of Tris could interact with nickel, decreasing His tag binding efficiency. Thus, Tris buffer was replaced by HEPES, that doesn't have any reactive amino groups. Additionally, the salt concentration was increased to 500 mM and 10% glycerol was added to increase subcomplex solubility, since SAGA HAT has displayed aggregation tendency at high concentrations (see further).

3.1.3.3. Size exclusion chromatography.

In the original protocol, the sample was concentrated after affinity purification before loading on a gel filtration column. However, when the sample after nickel column was concentrated, more than 50% protein loss due to probable aggregation was observed by measuring the final protein concentration. To confirm that the concentration at this step indeed leads to protein aggregation, the samples before and after concentration were loaded on the gel filtration column. The gel filtration profile in both cases showed three distinct peaks (fig. 48) that were checked on the SDS-PAGE: the first peak corresponded to aggregates (eluted before the column exclusion volume), the second peak – to the purified SAGA HAT subcomplex, and the small third peak – to isolated His-tagged GCN5. Nevertheless, the ratio between first and second peaks was different: in the case of the concentrated sample, the “aggregates” peak was clearly higher than the “protein complex” peak (fig. 48a). Thus, the concentration step was eliminated in further purifications.

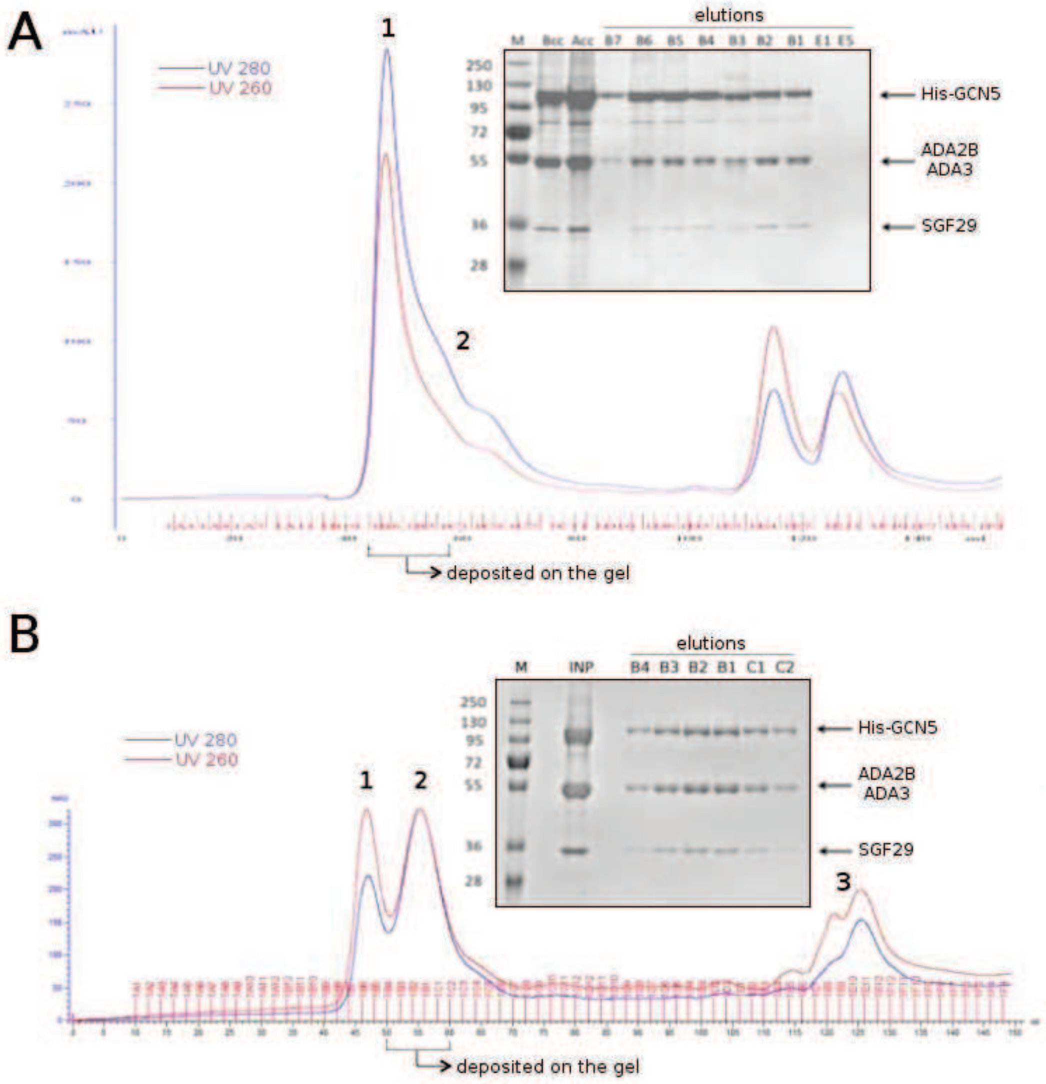

Fig. 48. Gel filtration chromatograms and corresponding SDS-PAGE gels for the concentrated (a) and non-concentrated (b) sample.

(a) Concentrated sample after GF displays the first high peak of aggregates, the 2nd peak is severely downweighed. Additional peaks at the end of elution correspond to GCN5 alone or contaminants (gel not shown). (b) Non-concentrated sample after GF displays 3 distinct peaks: 1 – aggregates (gel not shown), 2 – HAT subcomplex, 3 – GCN5 alone (gel not shown). The first peak is lower (UV 280 nm) than the second, however, nucleic acid contaminants (UV 260 nm) are present in both peaks.

To handle large sample volumes coming after IMAC (3-4 ml of pooled fractions), the gel filtration column was changed from Superdex S200 10/30 to Superdex S200 16/60, allowing increased sample loading (up to 4 ml) and higher peak resolution. Additionally, the gel filtration buffer was modified: besides the salt increase and glycerol addition as indicated before, DTT was added to prevent disulfide bond formation, as well as EDTA to reduce protease activity. Both DTT and EDTA were not used during IMAC due to their negative influence on the binding capacity of the nickel resin.

3.1.4. Cross-linking map and homology modeling of SAGA HAT module.

To analyze the architecture of the HAT module we have used chemical cross-linking in combination with mass spectrometry, that allowed to map the protein-protein interaction network. The spatial constraints, obtained in the cross-linking experiment, were used to build an initial low-resolution model of the HAT module, based on the known and homologous structures of HAT subunit domains and rigid body docking approach. The results obtained in this part of the work led to the co-first author publication with Nha-Thi Nguyen-Huynh:

Chemical cross-linking and mass spectrometry to determine the subunit interaction network in a recombinant human SAGA HAT subcomplex.

Nha-Thi Nguyen-Huynh, Grigory Sharov, Clément Potel, Pélagie Fichter, Simon Trowitzsch, Imre Berger, Valérie Lamour, Patrick Schultz, Noëlle Potier, Emmanuelle Leize-Wagner.

Protein Sci., vol. 24 (8), p. 1232-1246 (2015). doi: 10.1002/pro.2676

Publication 1

Chemical cross-linking and mass spectrometry to determine the subunit interaction network in a recombinant human SAGA HAT subcomplex

Nha-Thi Nguyen-Huynh,[1] Grigory Sharov,[2] Clément Potel,[1†] Pélagie Fichter,[2] Simon Trowitzsch,[3] Imre Berger,[3] Valérie Lamour,[2] Patrick Schultz,[2] Noëlle Potier,[1*] and Emmanuelle Leize-Wagner[1]

[1]Laboratoire de Spectrométrie de Masse des Interactions et des Systèmes (LSMIS) UMR 7140 CNRS/Université de Strasbourg - "Chimie de la Matière Complexe", 1 Rue Blaise Pascal, 67008 Strasbourg, France

[2]Integrated Structural Biology Department, Institut de Génétique et de Biologie Moléculaire et Cellulaire (IGBMC), UMR 7104, INSERM U964, 1 rue Laurent Fries, 67404, Illkirch, France

[3]European Molecular Biology Laboratory (EMBL), Grenoble Outstation, 6 rue Jules Horowitz, 38042 Grenoble, France

Received 22 December 2014; Accepted 3 March 2015
DOI: 10.1002/pro.2676

Abstract: Understanding the way how proteins interact with each other to form transient or stable protein complexes is a key aspect in structural biology. In this study, we combined chemical cross-linking with mass spectrometry to determine the binding stoichiometry and map the protein–protein interaction network of a human SAGA HAT subcomplex. MALDI-MS equipped with high mass detection was used to follow the cross-linking reaction using bis[sulfosuccinimidyl] suberate (BS3) and confirm the heterotetrameric stoichiometry of the specific stabilized subcomplex. Cross-linking with isotopically labeled BS3 d0-d4 followed by trypsin digestion allowed the identification of intra- and intercross-linked peptides using two dedicated search engines: pLink and xQuest. The identified interlinked peptides suggest a strong network of interaction between GCN5, ADA2B and ADA3 subunits; SGF29 is interacting with GCN5 and ADA3 but not with ADA2B. These restraint data were com-

---

Abbreviations: amu, atomic mass units; BSA, bovine serum albumin; CX, chemical cross-linking; DOPE, discrete optimized protein energy; EM, electron microscopy; ESI-MS, electrospray ionization mass spectrometry; H/D exchange, hydrogen/deuterium exchange; HCD, higher-energy collisional dissociation; IM, ion mobility; IMAC, immobilized metal affinity chromatography; LC-MS/MS, liquid chromatography coupled with mass spectrometry in tandem; MALDI-MS, matrix assisted laser desorption/ionization mass spectrometry; MOI, multiplicity of infection; NMR, nuclear magnetic resonance; PBS, phosphate buffered saline; SDS-PAGE, sodium dodecyl sulfate polyacrylamide gel electrophoresis; SEC, size exclusion chromatography; TFA, trifluoroacetic acid; XRD, X-ray diffraction.

Additional Supporting Information may be found in the online version of this article.

[†] *Present address:* Clément Potel's present address is Biomolecular Mass Spectrometry and Proteomics, Bijvoet Center for Biomolecular Research and Utrecht Institute for Pharmaceutical Sciences, Utrecht University, Padualaan 8, 3584 CH Utrecht, The Netherlands; Netherlands Proteomics Center, Padualaan 8, 3584 CH Utrecht, The Netherlands
Nha-Thi Nguyen-Huynh and Grigory Sharov contributed equally to this work.

Grant sponsors: Fondation pour la Recherche Médicale, the Institut National de la Santé et de la Recherche Médicale, the Centre National pour la Recherche Scientifique, the Strasbourg University, the Association pour la Recherche sur le Cancer and the Agence Nationale pour la Recherche (DiscoverIID grant), the Labex INRT, the French Infrastructure for Integrated Structural Biology (FRISBI) [ANR-10-INSB-05-01]; Grant sponsor: INSTRUCT as part of the European Strategy Forum on Research Infrastructures (ESFRI); Grant sponsor: European Commission (EC) Framework Programme (FP) 7 project ComplexINC; Grant number: 279039.

*Correspondence to: Noëlle Potier; Laboratoire de Spectrométrie de Masse des Interactions et des Systèmes (LSMIS), UMR 7140 CNRS/ Université de Strasbourg - “Chimie de la Matière Complexe”, 1 Rue Blaise Pascal, 67008 Strasbourg, France. E-mail: npotier@unistra.fr

bined to molecular modeling and a low-resolution interacting model for the human SAGA HAT subcomplex could be proposed, illustrating the potential of an integrative strategy using cross-linking and mass spectrometry for addressing the structural architecture of multiprotein complexes.

Keywords: cross-linking mass spectrometry; SAGA HAT subcomplex; proteomics; protein– protein interaction

Introduction

Most if not all biological processes are catalyzed by dedicated interaction networks. Proteins rarely exist alone in the cell but are involved in macromolecular complexes and elucidation of the assembling dynamics of these multiprotein blocks is of major importance to understand and modulate cellular dysfunction.

In the last few years, integrative approaches have attracted considerable interest. This might be partly because of the realization that, despite their own limitations, traditional structural methods have complementary attributes, and combining respective information would definitely help in exploring the architecture, dynamics, and function of macromolecular complexes. Data arising from various biophysical techniques may be gathered on the bound as well as on the free states of the multiprotein system, and may help generating testable structural models or understanding self-assembling of the protein building blocks.[1,2] Input of physical theories and statistical data arising from sequence or structural databases homology studies, might also be added to best refine the proposed model.[3] In this context, mass spectrometry (MS) was shown as a high-potential technique for characterizing intact multiproteins assemblies, especially for determining the complex stoichiometry or monitoring dynamic changes.[4] More recently, MS-based technologies, such as IM,[5] H/D exchange,[6] chemical cross-linking (CX),[7] have emerged to assist these structural building activities. In particular cross-linking, by providing important spatial restraint information, is emerging as a promising approach for generating residue level topological data.[8] Revealing precisely which residues are linked yields proximity information that might increase the resolution of the proposed structural model. An increasing number of studies were reported during the last decade,[9–11] including very large multiprotein complexes such as RNA polymerase II[12] and protein–drug interactions[13] or illustrating the integration of CX-MS to theoretical structural knowledge arising from modeling,[14,15] EM,[16] or XRD.[17] The choice of cross-linkers is crucial because the success of cross-linking reaction depends on their reactivity and selectivity as well as their spacer-arm lengths which rule the maximum distance between two cross-linked residues.[18,19] Identifying the cross-linked peptides is also still very challenging since they often are present at a small concentration in solution compared to unmodified peptides. Indeed, multiple cross-linked products may occur for one reactive residue or the reaction may be incomplete. Therefore, several enrichment methods that will improve MS detection have been successfully developed including affinity selection,[20] strong cation exchange (SCX) chromatography,[21] H/D labeling or exclusion list of low charge states peptides. Using cross-link reagent bearing a HCCA (α-cyano-4-hydroxycinnamic acid) moiety might also be very efficient allowing specific MALDI signal enhancement of such cross-linked peptides compared to unmodified peptides.[22,23] Furthermore, the CX-MS workflow requires the use of search engines dedicated to identification of cross-linked peptides. A large number of algorithms are available in the literature[24–27] but the main challenge still resides in assigning the confidence of a match.

The multisubunit Spt-Ada-Gcn5 Acetyltransferase (SAGA) complex investigated here is a paramount example of a transcriptional coactivator.[28] Transcription of protein coding genes by the RNA polymerase II (RNAPII) enzyme is a tightly regulated process that requires the assembly of a Pre-Initiation Complex (PIC) at the promoter of transcribed genes. Sequence-specific transcriptional activators and post-translational modifications of nucleosomal histones contribute to the recruitment of multisubunit coactivator complexes acting as bridging factors between the activators and the PIC.[29,30] In humans, the 1.8 MDa SAGA coactivator complex is composed of 18 distinct subunits[31,32] and adopts a modular organization as evidenced by genetic complementation studies,[33] and by electron microscopy models.[34,35] More recently, quantitative proteomics established a subunit interaction network and segmented SAGA into four stable modules.[36] SAGA harbors a Histone Acetyl Transferase (HAT) activity carried by the Gcn5 subunit and modulated by the Ada2 and Ada3 subunits. Together with Sgf29, these subunits form the HAT module which is a major regulator of histone H3 acetylation in yeast cells.[37,38] Gcn5 contains a bromodomain which binds acetylated lysines in histone tails,[39] while Sgf29 contains a Tudor domain capable of binding H3K4me2/3, another hall mark of actively transcribed chromatin.[40]

In this paper, we describe how CX-MS has been successfully used to determine the binding stoichiometry and topology of the human SAGA HAT subcomplex. We integrated these data with molecular modeling in order to determine the molecular interactions between the four subunits.

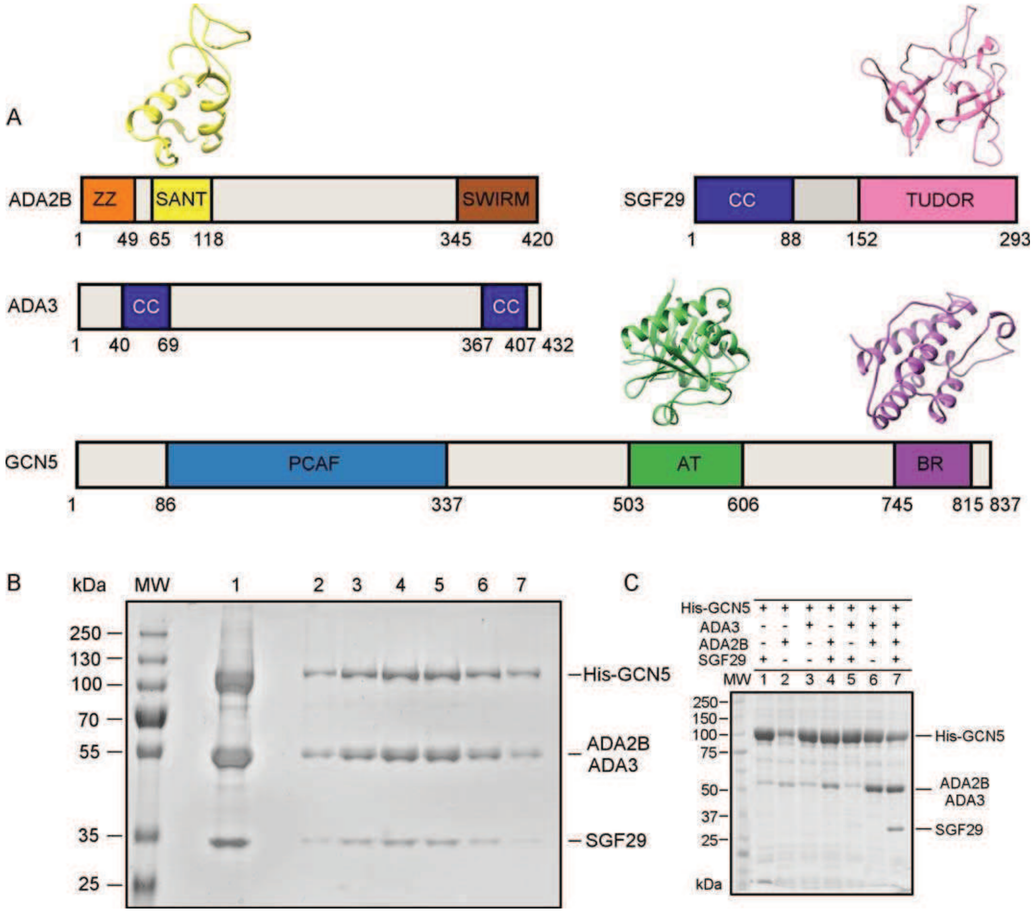

Figure 1. Purification and biochemical characterization of the SAGA HAT subcomplex. (A) Maps of the conserved structural domains within the SAGA HAT subunits. The domains are identified as follows: ZZ—type zinc finger domain; SANT—histone binding domain found in Swi3, ADA2, N-Cor and TFIIIB; SWIRM—eukaryotic domain found in proteins implicated in chromatin remodeling and gene expression; CC—coiled-coil domain; TUDOR—structural motif involved in binding methylated arginine or lysine residues; PCAF—P300/CBP-associated factor homology domain; AT—acetyltransferase domain; BR—bromodomain. The known atomic structures of the conserved domains are shown (SANT—PDB 1X41, TUDOR—PDB 3MEA, AT—PDB 1Z4R, BR—PDB 3D7C). (B) SDS-PAGE analysis of the purified SAGA HAT module. MW molecular weight marker, lane 1 purified SAGA HAT after the metal chelating affinity column, lanes 2–7—peak fractions of the size exclusion chromatography. (C) Co-expression of histidine-tagged GCN5 with different combinations of the other SAGA HAT subunits followed by pull down experiments.

Results

Production of the human SAGA HAT module

The sequence analysis of the SAGA HAT subunits revealed conserved homology domains and for some of them an atomic model has been determined by X-ray crystallography [Fig. 1(A)]. To analyze the SAGA HAT subcomplex, we produced recombinant hGCN5, hADA2B, mADA3, and hSGF29 from a single virus in insect cells using the Multibac technology.[41] The mouse ADA3 sequence is 99.3% identical to its human homologue with the exception of three mutations (T110A, V126L, and I226M). The complex was histidine-tagged on GCN5 and purified over a metal chelating affinity column followed by size exclusion chromatography (SEC). SEC of the HAT complex showed a single peak consistent with an homogeneous protein assembly and the SDS-PAGE analysis indicated that all four polypeptides co-elute in the same fractions [Fig. 1(B)]. Because of their close molecular weight ADA2B and ADA3 co-migrate on a 10% acrylamide gel.

Co-expression experiments

In order to investigate the interactions between the catalytic subunit GCN5 and the other subunits of the SAGA HAT module, the histidine-tagged GCN5 was co-expressed in isolation or with different combinations of the ADA3, ADA2B, or SGF29 subunits. The resulting complex was purified and analyzed as described above. The ADA2B and ADA3 subunits interact poorly in isolation with GCN5 [lanes 2 and 3 in Fig. 1(C)] but form a heterotrimeric complex when expressed together [lane 6 in Fig. 1(C)]. SGF29 incorporates only in a complex formed by GCN5, ADA2B, and ADA3 [lane 7 in Fig. 1(C)] and does not bind stably to GCN5 alone [lane 1 in Fig. 1(C)].

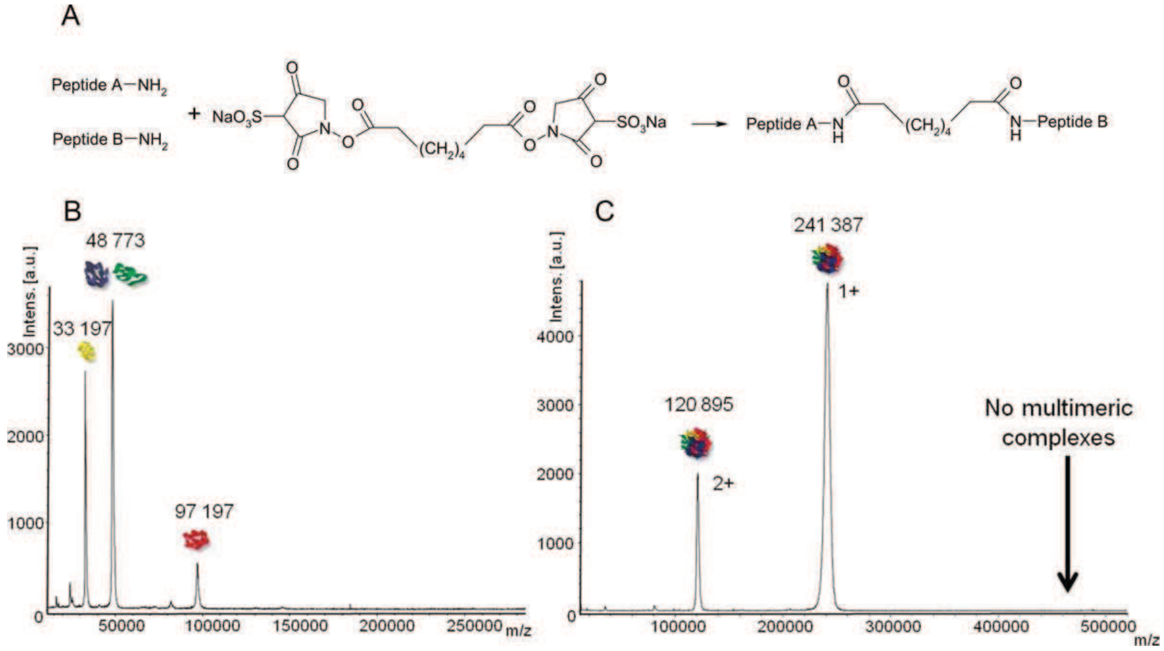

Figure 2. (A) Chemical formula of two peptide chains connected through the BS3 cross-linker. (B,C) MALDI-MS spectra of the complex before (B) and after (C) cross-linking with BS3. A unique stoichiometry at 241,387 Da is detected after cross-linking.

Characterization of the cross-linking reaction by MALDI-MS

The cross-linking reaction was performed with bis[sulfosuccinimidyl] suberate (BS3). The reaction conditions (cross-linker to protein ratios, times, and temperatures) were optimized so that cross-linker reacts on the whole sample, but at a quite low cross-linker concentration to avoid perturbation on structure which can lead to aggregation or nonspecific interactions[42] (see Materials and Methods, Supporting Information Fig. 1). The complex stoichiometry as well as the cross-linking kinetics were followed by MALDI-MS analysis using high mass detection.[43] Before cross-linking, ions at 33,197, 48,773, and 97,197 ± 10 Da were detected, which corresponds to the theoretical masses of the four subunits (ADA2B and ADA3 are co-detected since their mass difference is too small to be distinguished by MALDI-MS) [Fig. 2(A)]. After cross-linking, molecular mass of about 241,387 Da ± 200 Da was measured. From this mass, two possible stoichiometries could be formally deduced, i.e. 1/1/1/1 (of each subunit) or 1/2/1 (GCN5/ADA3 or ADA2B/SGF29) [Fig. 2(B)]. A bottom-up approach by LC-MS/MS (see further) will definitively allow to conclude in favor of the first hypothesis. The signal corresponding to the cross-linked entire complex is quite broad, which results in a quite low mass accuracy but sufficient to propose a tetrameric binding stoichiometry without ambiguity. This may be because of the heterogeneity of the cross-linked molecules but also to a certain amount of salt adducts arising from the 100 mM NaCl containing buffer. No signal corresponding to other stoichiometries (homomers or high-order multimers) neither to intermediates reflecting incomplete cross-linking reaction or assembly process was observed. This suggests that there was no residual aggregation and that the detected complex with the expected stoichiometry is formed by specific interactions.

Determination of the proximity map in SAGA HAT by LC-MS/MS

In order to get insight into the "in-proximity areas" (through identification of the spatially close lysines), the cross-linked complex was digested with trypsin and the resulting peptide mixture was then directly analyzed by LC-MS/MS. As described by Müller *et al.,*[44] a mixture of the light form (d0) and the isotopically labeled form (d4) of BS3 was used for the cross-linking reaction in order to facilitate the detection of low abundant cross-linked products. In this method, peptides containing the cross-linker will appear as a doublet in the mass spectrum with a mass shift of 4.025 amu.

Trypsin activity did not appear to be affected by cross-linking as shown by the high sequence coverages obtained after classical Mascot database search (from 55% for SGF29 up to 91% for ADA3). Whereas peptides that have been modified by a partially hydrolyzed cross-linker (type 0 cross-link or dead-end peptides[18]) are easily identified by performing standard database search in Mascot, two open-source algorithms, pLink[26] and xQuest,[27] were chosen, for both intra- and interlinkages determination. Whereas pLink is dedicated to the identification of peptides modified with either light or heavy cross-linker, xQuest is based on the determination of

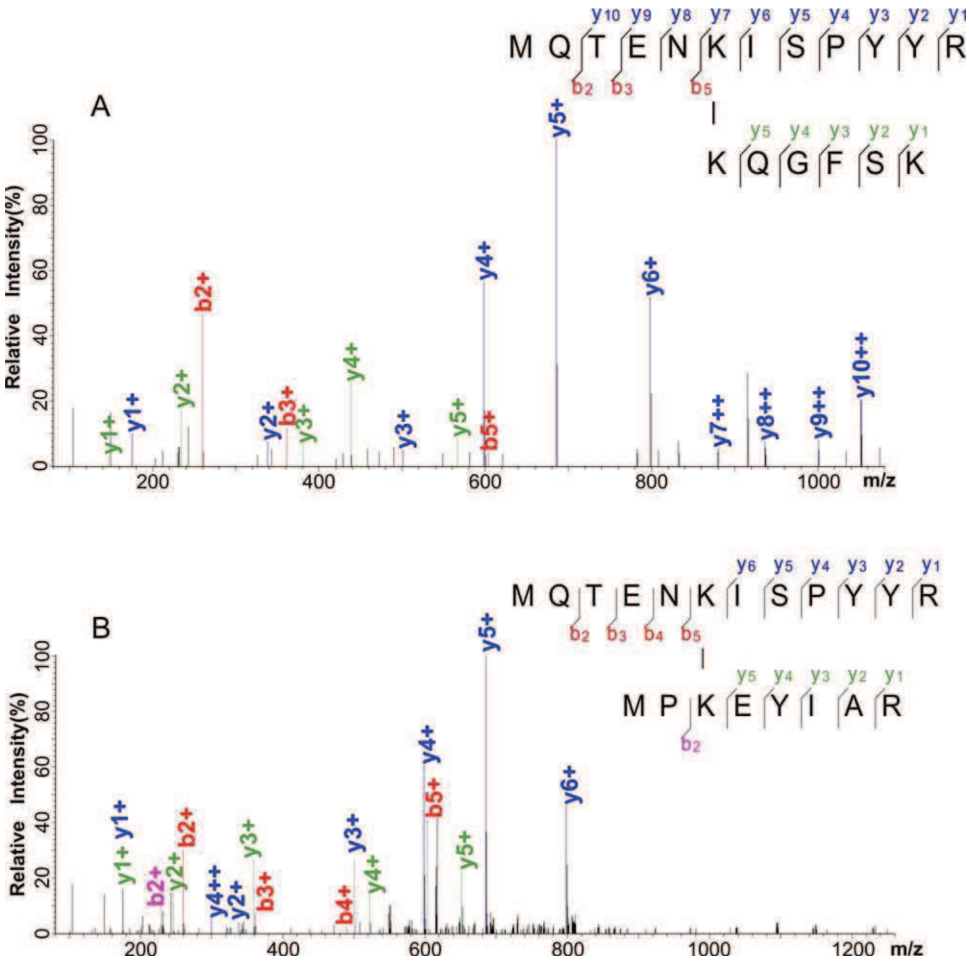

Figure 3. MS/MS spectra of two intercross-linked peptides as identified by pLink (score E-10). The K–K bridge may be (A) or not (B) detected as an ion fragment.

peptides modified with both light and heavy cross-linkers. The request was repeated using a database containing random protein sequences (instead of the sequences of four subunits of SAGA HAT) and false cross-linker mass in order to evaluate false positive matches as well as the associated scorings for both algorithms. From such experiment, background threshold scores were extracted, under which there is a serious risk that a good fit with the MS/MS data occurred by chance. Scores higher than 25 for xQuest and E-06 for pLink were then chosen as thresholds for positive identification. In addition, manual validation was systematically performed to further validate the identification (about 10% were manually removed). Cross-linked peptides follow the general rules of CID-peptide fragmentation[45] and thus, their fragmentation spectra should provide fragment ions arising from both bridged peptides upstream and downstream of the linkage. Two examples of MS/MS spectrum assignment by pLink are shown in Figure 3. In both cases, the corresponding peptides were unambiguously identified as reflected by the large series of consecutive y-ions and b-ions. In some cases, ions containing the linkage product will also be detected increasing the identification relevance as shown in Figure 3(A) at $m/z = 879.4828$.

A total number of 97 unique inter-cross-links (indicating the linkage of unique lysine–lysine pairs) have been found with pLink and xQuest (Table I). Among them, 23 were identified by xQuest only and 43 by pLink only. These results are the combination of three technical replicates (displaying 80% overlap) and one biological replicate (40% overlap). It should be noticed that the proximity areas were strictly equivalent in all replicates. The overlap discrepancy might come from: (i) a lower MS/MS intensity spectrum that could not be unambiguously assigned, (ii) a flexibility of the complex structure that slightly modify the distance between reactive lysines leading to different links, or (iii) from a competitive reactivity of two neighboring lysines, yielding cross-linked peptides with different masses.

A linkage map has been produced from these results to visualize intrasubunit links and intersubunit links (Fig. 4). This map clearly highlights several discrete inter-linked areas with high

Table I. *List of intermolecular cross-links between SAGA HAT subunits. No score means the cross-linked peptide was not identified or not validated manually.*

GCN5-ADA3

Linked K in GCN5	Linked K in ADA3	Score in pLink	Score in xQuest	Mass (Da)	Mass precision (ppm)
314	85	5.26E-14	55	2614.3754	−0.4
314	147	3.65E-12		2905.4569	1.1
314	221		45.71	2654.3527	−0.9
314	354	1.56E-13	44.66	3218.6297	−0.7
361	85	5.47E-10		1803.0770	0.8
366	78	4.89E-07		1229.6615	−0.7
366	85	7.13E-07		1491.8961	0.7
366	421	5.47E-09		1400.7874	−0.1
536	85	1.74E-06		2835.6122	0.2
561	78	2.27E-07		1622.8750	−0.1
561	85	1.15E-08	50.64	1877.0595	−0.1
561	109	1.15E-09		2560.3219	1.3
561	147	1.37E-13		2172.1661	−0.6
580	85	1.28E-11	45.79	1856.1238	−0.5
649	85	8.37E-10	51.03	1563.9136	−0.7
649	124	2.08E-10	45.04	1630.9042	−0.3
649	304		36.54	2227.1581	−7.6
657	405		27.06	1937.1456	−7.9
660	85	1.29E-06		1455.8925	−0.8
707	85	1.03E-08		1617.0026	−0.3
713	364		35.84	2059.1279	−0.5
748	418		32.19	1957.0869	−0.8
750	418		44.43	1957.0843	−2.1
784	78	9.97E-07		2081.0428	2.5
816	85		35.8	1959.1605	−3.3

GCN5-ADA2B

Linked K in GCN5	Linked K in ADA2B	Score in pLink	Score in xQuest	Mass (Da)	Mass precision (ppm)
314	328		44.78	2626.3341	−0.4
314	331	5.50E-07	57.22	2514.2217	−0.8
314	378	7.20E-09	42.5	3089.6396	−3.3
314	390		40.44	2753.3972	−0.5
361	390	4.30E-06		1942.0999	−2.7
536	214	2.43E-06		2651.5058	−0.7
572	214	2.85E-08		1664.9613	−1.1
572	220	7.75E-13		1720.9987	−0.8
580	250		38.81	2154.2570	−7.7
580	250		35.23	1870.0746	−1.6
702	250		29.07	2645.4865	−3.5

GCN5-SGF29

Linked K in GCN5	Linked K in SFG29	Score in pLink	Score in xQuest	Mass (Da)	Mass precision (ppm)
314	40	4.31E-11		3585.7811	−1.3
314	50	2.09E-08	38.84	3410.6459	0.2
561	40	2.71E-16		2852.4903	1.1
561	50	5.72E-12	44.79	2673.3288	−0.2
580	40	5.47E-09		2827.5303	1.9
580	50	5.78E-10	51.09	2652.3947	0.1
649	40	1.06E-16	34.72	2539.3443	−0.2
649	50	4.39E-10	43.37	2360.1818	−0.7
649	79	1.49E-09		1404.7976	−0.5
649	112	6.92E-08		1365.7438	1.2
816	40	4.67E-07		2934.5975	−6.7

Table I. *Continued*

GCN5-SGF29

Linked K in GCN5	Linked K in SFG29	Score in pLink	Score in xQuest	Mass (Da)	Mass precision (ppm)
816	50	1.39E-14		2755.4366	0.5
855	50	1.85E-06		2638.3675	−0.3

ADA3-ADA2B

Linked K in ADA3	Linked K in ADA2B	Score in pLink	Score in xQuest	Mass (Da)	Mass precision (ppm)
147	235	7.92E-13	45.89	2644.4160	0.4
147	390		42.1	2033.1002	−0.8
194	331	3.05E-09	41.76	2069.1205	−0.1
194	378	7.49E-06		2640.5114	0.0
194	390		35.29	2304.2634	−3.0
210	319	3.62E-06		2446.2276	0.6
210	319	3.33E-08	30.06	3172.6412	−6.4
210	322	1.63E-10	39.13	2884.4199	−0.1
210	322	2.88E-12	40.76	3016.5401	−0.1
210	328	4.43E-06	56.43	2673.3658	0.2
210	331	7.99E-08	50.25	2557.2294	−0.1
210	378	8.13E-06		3132.6454	−0.6
210	390		33.39	2796.4038	−0.2
221	319	4.79E-07		1431.8233	1.6
221	322	9.36E-09	297.22	1874.0409	1.1
221	322	6.31E-15	45.29	1998.1108	−0.4
221	328	3.39E-14	42.52	1654.9365	0.5
221	331	1.54E-09	44.44	1542.8210	−2.8
221	373	1.63E-08		2087.2605	0.8
221	378		30.64	2118.2407	−0.3
221	390		41.6	1781.9974	−1.4
223	328	3.72E-07		2270.3145	0.1
223	378	2.86E-06		2729.5941	0.0
242	328	1.06E-06		1783.0202	−0.8
242	378	2.96E-11	45.58	2246.3254	0.2
242	390	9.38E-06	44.71	1910.0836	0.5
304	378	4.59E-06		2741.5227	−2.4
304	390		42.97	2405.2822	0.3
312	390		26.06	2977.5704	−1.0
354	390		40.72	2346.2754	−0.6
364	282	3.79E-07		1597.9274	0.2
372	390	2.14E-07	41.31	1940.0690	0.1
397	403		29.07	2645.4865	1.0

ADA3-SGF29

Linked K in ADA3	Linked K in SGF29	Score in pLink	Score in xQuest	Mass (Da)	Mass precision (ppm)
77	50	6.23E-14	33.99	3586.8165	0.2
78	50	4.17E-07		2141.0575	1.3
85	40	2.34E-08		2574.4029	−0.1
85	50	1.70E-09	47.29	2399.2672	0.0
85	50		39.28	2555.3450	−9.1
97	50		34.14	3176.6059	−0.6
109	40	4.30E-08		3257.6652	−0.3
147	40	1.57E-11		2869.5094	0.4
147	50	4.08E-10	45.14	2690.3476	−0.4
372	40	1.38E-11	30.73	2772.4517	−1.3
372	50	1.06E-06		2597.3158	−0.6
397	50	4.09E-07		2355.2078	−0.1
421	40	6.71E-06		2487.3193	0.3
421	50	7.04E-07		2312.1834	0.4
421	99	2.34E-07		2501.3789	0.6

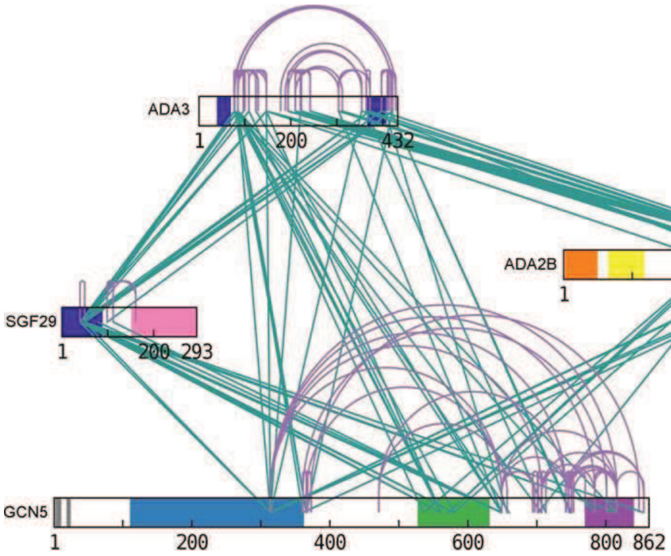

Figure 4. Inter- (green) and intra- (purple) molecular cross-linking map within the SAGA HAT subunits obtained with the BS3 d0-d4 cross-linking agent. The represented cross-links are deduced from both xQuest and pLink identification. Colors are consistent with Figure 1(A).

significance as shown by many contacts that have been detected for each of them. In particular, extensive cross-linking was observed between ADA3 and ADA2B suggesting a strong network of connections between these two subunits while no cross-links were detected between SGF29 and ADA2B. Remarkably, among the 22 lysines equally distributed on the SGF29 sequence, two lysines (K40 and K50) appeared to be mostly involved in inter-cross-linked products, suggesting that the bridging reaction occurred in a very specific way. On the other hand, no cross-links could be detected for the N-terminal half of GCN5 containing most of the PCAF domain, the Tudor domain of SGF29, and the ZZ-SANT domains of ADA2B.

Molecular modeling

Atomic structures of several domains composing the HAT subunits are available [SANT, Tudor, acetyltransferase, and bromo domains, Fig. 1(A)] or can be predicted by homology modeling of closely related sequences (PCAF, ZZ, SWIRM). Intrasubunit cross-links were used as additional spatial restraints helping to predict tertiary models of the SAGA HAT subunits since the relative positions of their domains are not known. For each subunit, domains with known or modelled structure were positioned in a way that minimizes the distances between the cross-linked lysines. That was achieved with a rigid body docking approach where the domain structure was not deformed.

For GCN5 the intricate interaction network has allowed to position the PCAF, acetyltransferase and bromo domains satisfying the maximum theoretical distances between cross-linked lysines [Fig. 5(A)]. The only exception is the connection between lysines K361 and K801 [Fig. 5(B)]. K361 residue is situated within a disordered region between PCAF and acetyltransferase domains, that region is flexible and thus the lysine residue may occupy a different position. The whole GCN5 subunit is heavily cross-linked with the exception of N-terminus including mostly PCAF domain. According to the secondary structure prediction this part of the protein is mostly helical. We note that about 40% of GCN5 sequence, comprising the gaps between three known domains, is unstructured and we could not predict tertiary structure of these areas. In case of ADA2B protein most of the cross-links were detected in C-terminal part. ZZ and SANT domains that are known to stimulate nucleosome binding are not involved in any interactions [Fig. 5(C,D)]. The middle part of ADA2B does not have a specific structure and represent around 45% of total sequence. The complete structure of ADA3 was predicted by comparative modeling and docking of four domains (see Materials and Methods section) since no homologs with known structures were found. Secondary structure prediction showed that this protein is mostly helical with two short coiled-coil domains at the termini [Fig. 5(E,F)]. In the case of SGF29 we have modelled the structure of the coiled coil domain (3–88), but unfortunately this model is of low quality since the protein does not display enough intralinks that would allow more accurate structure modeling.

In order to obtain an initial low-resolution model of SAGA HAT subcomplex, all four subunits were positioned one with respect to the others by minimizing the distances between inter-subunit cross-linked lysines and by avoiding any steric clashes between the atomic models [Fig. 6(A,B)]. This model of the SAGA HAT module is still missing

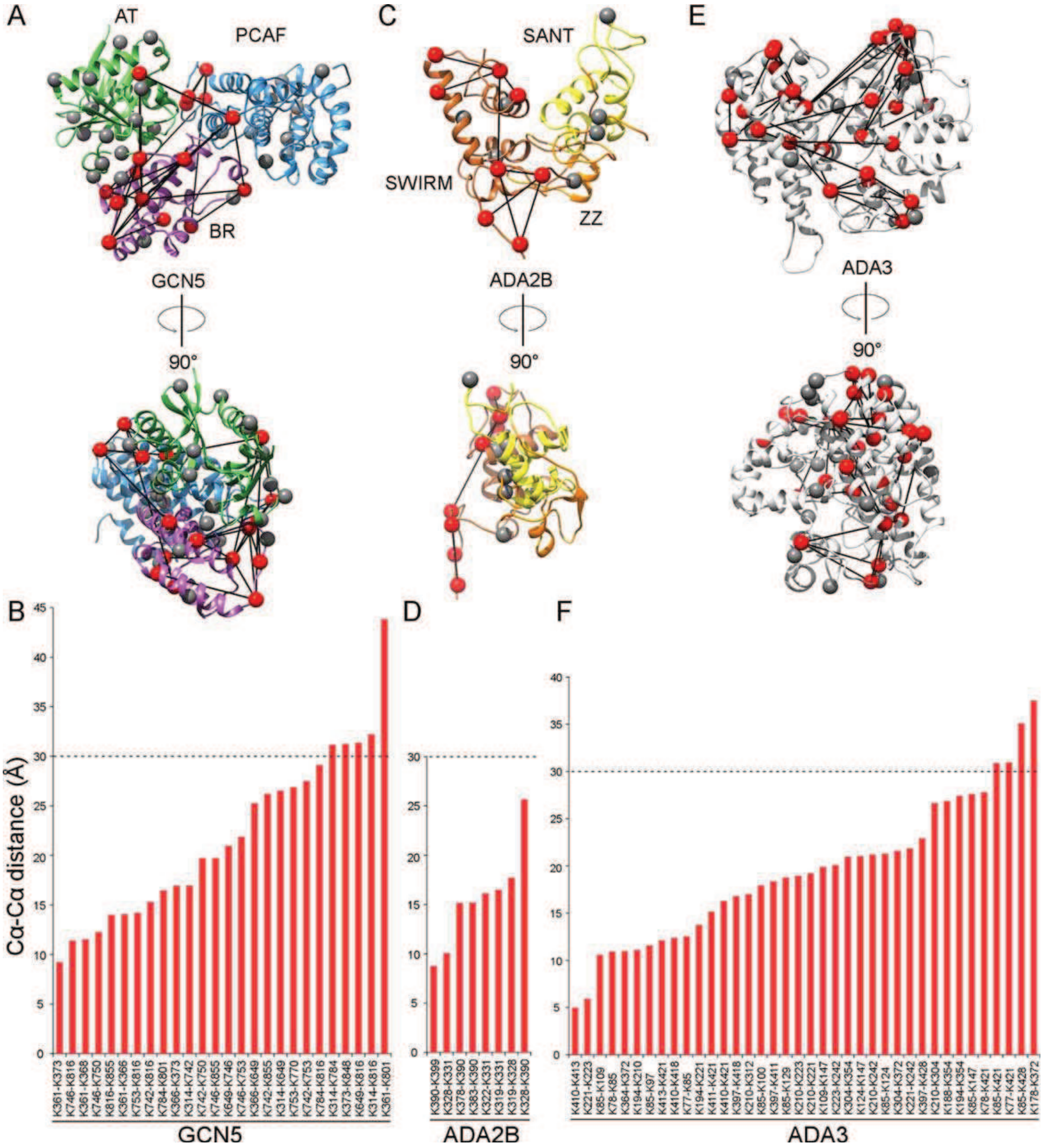

Figure 5. Tertiary subunit architecture modelled by intrasubunit cross-links. (A, C, E) Modeling of the conserved domains of GCN5, ADA2B, and ADA3, respectively. Red spheres correspond to lysine residues involved in intrasubunit cross-links, grey spheres show the nonmodified lysines and black lines represent the distance between the cross-linked residues. (B, D, F) Calculated distances between all cross-linked residues of GCN5, ADA2B, and ADA3, respectively. The 30Å threshold is indicated by a dotted line.

several unstructured parts of subunits and might be further improved by molecular dynamics methods since here only rigid-body docking was used.

Discussion

In this project, chemical cross-linking-mass spectrometry was used in combination with modeling in order to build a low-resolution structural model of the SAGA HAT subcomplex. Even with advances in mass spectrometry technology or with cross-linking reagents improvements, data interpretation is still challenging and the problem of extracting effective cross-linked peptide from random matching remains.

Strategy and validation of the identification of cross-linked peptides

MALDI-TOF allowed a straightforward comparison between control and cross-linked samples and thus a direct determination of the binding stoichiometry. While ESI-MS is shown to be an appropriate method to characterize noncovalent complexes in terms of existence and stoichiometry,[46] it could not be used in this particular case, since the SAGA HAT subcomplex could not be transferred properly to a volatile ammonium acetate buffer. Chemical cross-linking appeared as a nice alternative by covalently stabilizing the complex before its characterization by mass

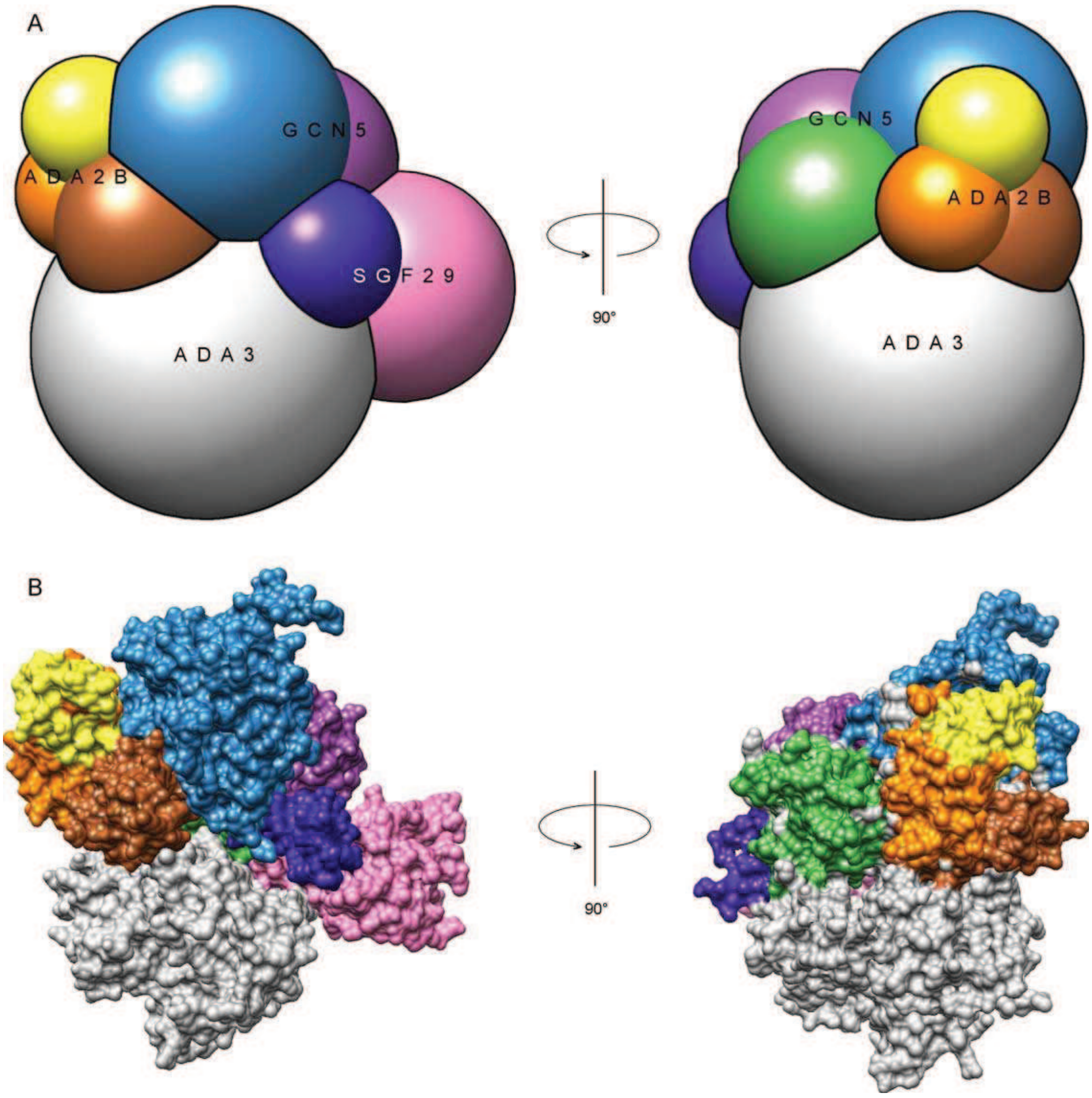

Figure 6. Quaternary organization of the SAGA HAT module. (A) Schematic representation of the 3D organization of the conserved SAGA HAT subunit domains obtained upon distance minimization of intersubunit cross-linked lysine residues. (B) Proposed 3-D model of the SAGA HAT module in which the atomic models of the conserved domains are positioned and oriented according to the cross-linking network and avoiding structural overlap. The model represents 63% of the SAGA HAT residues.

spectrometry so that it could withstand denaturing conditions and be analyzed by methods that usually lead to complex dissociation. A bottom-up approach was then used to look for the cross-linked peptides and establish a linkage map. As already described in the literature,[47,48] mass spectrometer yielding high-resolution measurement of peptide and fragment masses but also able to precisely determine the charged states, appeared to be essential for the success of the database search. The large number of bridged products that are arising from all possible pair-wise combination of peptides containing a lysine should be considered, which generates a very large theoretical database and increases the risk of random matches if mass accuracy is not sufficient. In our case, the parameters of mass spectrometer (mass range of 300–1200 *m/z*) allowed us to analyze a large majority of digested peptides present in solution without compromising the high mass accuracy and high resolution. All matched peptides displayed a mass accuracy within 6 ppm from which 97% were lower than 2 ppm.

Interestingly, the overall connectivity pattern of the pLink- and xQuest-extracted maps is very convergent (Supporting Information Fig. 4). Each of the putative linked areas has been observed with both algorithms and combining the results of both searches yielded a more expressive network map. Some limitations of one algorithm might be compensated by the other (pLink requires generous fragmentation spectra of the cross-linked peptide, whereas xQuest needs quality fragmentation spectra for both d0 and d4 modified peptide).

Several observations strengthen the validation of our putative linkage map. (i) Each linked area is defined by multiple peptide pairs (for example, 6 and 7 links respectively are detected to propose a connection between the N- and C-termini of ADA3 with the N-terminal domain of SGF29). (ii) Most of the time, there is a correlation between inter and intrasubunit linkages. For example, the lysine K40 of SGF29 is in the vicinity of the lysines K85 and K421 of ADA3 whereas intrasubunit link between K85 and K421 of ADA3 strongly suggests that these two lysines are brought close to each other by the tertiary fold of the subunit. (iii) The analysis of deletion mutants have shown an interaction between yAda2b residues 167–434 (including the SWIRM domain) with C-terminus of Ada3[49] which is fully consistent with our observations. (iv) Yeast two hybrid approaches demonstrated an interaction between the N-termini of SGF29 and of ADA3[50–53] which is also revealed by our experiments.

The absence of observed cross-linked peptides might also be worth to exploit but it is not straightforward to unambiguously correlate no cross-linking with distant lysines (i.e. higher than the known 30 Å maximum distance of Cα–Cα atoms of peptides that can be cross-linked by BS3[54,55]). Indeed, various situations could prevent reaction with the cross-linker: (i) there is no lysine in a particular domain (only four lysines in the ADA2B N-terminus domain); (ii) the lysines may not be accessible to the cross-linker (buried in the structure) or not reactive because of chemical environment. Looking at dead-end peptides could likely help to address this purpose (see Supporting Information Table 3). In our case, 53% of the lysines in the GCN5 PCAF domain have been shown to possibly react with BS3, meaning that they are accessible to the solvent but without forming cross-links. On the other hand, only two lysines among 12 in SGF29 Tudor domain have been detected as dead-end modified lysine, strongly suggesting that these lysines are not available for cross-linking reaction. Interestingly these noncross-linked domains are often involved in interactions with modified histone tails and are likely to be exposed to bind nucleosomes thus being less prone to interactions with subunits of the HAT module.[37,56] In such situation, the use of cross-linkers containing other types of reactive functions (photoreactive cross-linkers, sulfhydryl cross-linkers etc.) is also conceivable. In these cases, it may be important to adapt the cross-links identification software to be used. (iii) The mass/charge of cross-linked peptides may be higher than 1200 *m/z* to be selected by the MS for fragmentation, or too short (less than three amino acids) to give sufficient information for being recognized during database search; (iv) the fragmentation mass spectrum is of poor quality to be unambiguously interpreted.

Biological relevance of our linkage map

We compared our results with the recently published linkage map of the full yeast SAGA.[57] The yeast and human subunits show a large degree of sequence conservation except for hGCN5 which contains an additional PCAF domain at its N-terminus and the yAda3 subunit which is twice as large as hADA3. Despite a different lysine distribution in the sequences from two organisms, the network of domain interactions is highly conserved with the notable exception of ySgf29 which shows a large number of intrasubunit links and only interacts with yAda3 while hSGF29 has a strong interaction network with the most domains of hGCN5 (Fig. 4). The interaction between yeast Sgf29 and Gcn5 was previously demonstrated by systematic deletion analysis.[58] The same regions of hGCN5 make extensive cross-links with hADA3 along its whole sequence while in yeast the link between these two subunits is limited to a single domain in the C-terminal part of yAda3. Rather than reflecting the differences in primary sequence between the two organisms, we suggest that these differences in the cross-linking network reflect a conformational change between the free HAT module and when it is incorporated into the complete SAGA complex. Interestingly, the domains that show less cross-links within the HAT module in the yeast complex are strongly involved in interactions with other SAGA subunits. As shown by Han *et al.,*[57] in yeast Ada3 has few intra- and interlinks with subunits of HAT module, at the same time revealing numerous cross-links with other proteins like Taf5, Taf6, Taf12, Spt7, Sgf73, and Ada1. We show that in human ADA3 is extensively cross-linked to the other HAT subunits: ADA2B, SGF29. and GCN5. Altogether these results suggest that the HAT module is reorganized upon integration in the SAGA complex.

The histidine pull down assays [Fig. 1(C)] do not completely recapitulate the cross-linking experiments and it appears that proximity does not imply interaction. Cross-linking results indicate that the N-terminus of SGF29 is close to GCN5 but this proximity is not enough to form a stable pair-wise interaction. SGF29 requires the presence of both ADA3 and ADA2B to establish a network of interactions leading to a HAT subcomplex. Likewise the ADA2B and ADA3 subunits incorporate significantly into the subcomplex only when they are co-expressed together with GCN5. The last observation is partially supported by the fact that yAda2 plays a critical role in the formation of the HAT module and its association with the overall SAGA complex.[36]

The interaction between CX-MS and molecular modeling is very promising. The data generated by CX-MS will provide associations and restraints information that will help to eliminate inconsistent theoretical models. In combination with complementary techniques, CX-MS is moving to integrate a general pipeline that aim to iteratively improve the accuracy of the proposed structural model with a final objective of determining the closest one to the real state of the complex at a particular moment of its life. Large, heterogeneous or dynamic complexes would benefit from such an alternative approach since traditional structural biology methods are often difficult to carry through.

Materials and Methods

SAGA HAT expression and purification

The constructs for the expression of the HAT module of SAGA (hGCN5, hADA2B, mADA3, hSGF29) were generated using baculovirus Multibac system.[41] Briefly, hGCN5 with TEV-cleavable 6xHis-tag on N-terminus was cloned into pFL acceptor plasmid, the other three genes encoding hADA2B, mADA3, and hSGF29 were cloned into donor plasmids (pIDS, pIDC and pIDK respectively), which were subsequently combined via Cre-Lox recombination. The total construct was integrated via Tn7 transposition into a bacmid with YFP at the backbone as an expression level marker. The bacmid was isolated and used for infection of 1 L Sf21 insect cells (160 mL of virus for 1 L of SF21 insect cells in Sf900II serum-free media, MOI 15–20). YFP fluorescence was used to monitor overall expression levels after the day of infection. Cells were harvested in ice cold PBS with 10% glycerol when YFP levels reached a plateau (72 h). All further purification steps were done at 4°C.

Cell pellet from 1 L of culture was dissolved in 40 mL of lysis buffer (20 mM HEPES-KOH pH 8.0, 500 mM NaCl, 10% glycerol, Protease Inhibitor Cocktail (Roche)) and mildly sonicated on ice for 3 min. Cell extract was further cleared by ultracentrifugation for 1 h at 50,000 g, HAT subcomplexes were purified by His-tag using HiTrap Chelating HP 5 mL column (GE Healthcare), previously loaded with nickel. After washing steps with increasing concentration of imidazole, HAT subcomplexes were eluted at 250 mM imidazole as confirmed by 10% SDS-PAGE. Peak elutions were pooled and further purified on Superdex S200 16/60 GL (GE Healthcare) gel filtration column. Complexes were eluted in a similar buffer (20 mM HEPES-KOH pH 8.0, 500 mM NaCl, 10% glycerol, 1 mM DTT, 0.1 mM EDTA) and analyzed by 10% SDS-PAGE. Peak fractions were pooled, concentrated on Amicon 30 kDa column, then flash frozen in liquid nitrogen and stored at −80°C.

During purification steps, protein concentrations were controlled by Bradford assay,[59] their quality by SDS-PAGE[60] and the identity of the subunits by Western Blotting.[61]

Pull-down assays

Constructs for expression of different subunit combinations were generated using Multibac system and expressed in insect cells as described before. Small scale purifications using histidine-tag on GCN5 were performed in batch and eluates were analyzed on 10% SDS-PAGE.

Cross-linking and MALDI-MS data acquisition

Before cross-linking reaction, the complex was subjected to buffer exchange to 20 mM HEPES, 100 mM NaCl in 4°C to eliminate glycerol (Micro Bio-Spin chromatography column, exclusion limit 6 kDa). An equimolar mixture of the BS3 cross-linking reagent (Thermo Scientific) in the normal form (d0) and isotopically labeled form (d4) was prepared by dissolving in water. Cross-linking reactions were carried out on 30 μL of a 2 mg/mL protein complex at various protein to cross-linker molar ratios namely 1:50, 1:75, 1:200 at two incubation times (30 min, 2 h). From these assays, the optimal conditions for cross-linking reaction were fixed at a ratio protein:cross-linker 1:75 (equivalent to a lysine:crosslinker molar ratio of 1:0.5 since SAGA HAT subcomplex contains 164 lysine residues) for 30 min at room temperature. The reaction was quenched with Tris 25 mM final concentration at pH 7.5. The intact and cross-linked complexes were analyzed by MALDI-MS (Autoflex II TOF/TOF, Bruker) equipped with a HM1 high mass detector (CovalX, Switzerland). Sinapinic acid (SA) matrix at 10 mg/mL in H_2O/ACN/TFA (50/50/0.3) and a dried drop spotting method were chosen. The analyses were performed in a linear acquisition mode and the MALDI calibration was achieved with a solution of BSA at 5 μM using the monocharged ions of the monomer and dimer of BSA (Sigma-Aldrich).

LC-MS/MS data acquisition

After cross-linking, the protein complex was digested with trypsin (Sequencing Grade Modified, Promega) at a 1:5 enzyme to substrate molar ratio and incubated for 3 h at 37°C. The reaction was quenched with addition of TFA at final concentration of 0.1%. MALDI analysis was performed as a control after digestion, in order to ensure that there was no more signal corresponding to the entire complex. LC-MS/MS data were acquired on an LTQ Orbitrap Elite mass spectrometer (Thermo Scientific) equipped with a nano electrospray ionization source in positive ion mode coupled with a Rapid Separation LC system (nanoRSLC U3000, Thermo Scientific), using HCD fragmentation mode. The isolation window of precursor ion was fixed at 2 Da, and the HCD collision energy at 30% for the fragmentation. A full-scan survey MS spectrum (300–1200 *m/z*) was carried out with resolution of 240,000 at *m/z* 400, then the fifteen most abundant ions detected were submitted to a MS/MS experiment in Orbitrap at resolution 15,000.

MS data analysis

Two search engines, dedicated to cross-linked peptides xQuest (v2.1.1, http://proteomics.ethz.ch/cgi-bin/xquest2_cgi/index.cgi) and pLink (v1.15, http://pfind.ict.ac.cn/software/pLink), were used. In the xQuest searches, the precursor mass tolerance was set at 10 ppm as well as the fragment mass tolerance for common- and cross-linked ions. In pLink, the tolerance for MS1 matching was fixed at 10 ppm and the filter for peptide tolerance at 10 ppm. Each data set was produced using a FPR of 5%. All spectra of putative cross-linked peptides were manually controlled before positive identification. The bridged peptides were confidently identified when the majority of the observed fragment ions were assigned and when more than three consecutive fragment ions were matching for both linked peptides. The interaction maps between subunits were drawn via xiNET-Crosslink Viewer (http://crosslinkviewer.org), developed in Rappsilber laboratory.

SAGA HAT sequence analysis and structural modeling

Multisequence alignment of SAGA HAT subunits from different species allowed to identify the conserved structural domains. Homology modeling for the domains without known structure was used to predict the atomic models, otherwise the available crystal structures were found in PDB database. Intramolecular cross-linking data provided the distance restraints for further improving the homology models of each domain. Complete models of GCN5, ADA2B, and ADA3 were generated using rigid-body fitting with imposed cross-linking restraints (maximum distance between cross-linked lysines was set to 30 Å). The quality of the final models was verified using several bioinformatics algorithms, all models received good to fair scores (Supporting Information Table 4). The best predicted atomic models of all subunits were positioned and oriented according to the intersubunit cross-linking network and avoiding structural overlap, thus resulting into the 3-D model of the SAGA HAT module.

Disclosure: The authors declare no competing financial interest and no conflict of interest.

Acknowledgments

N-T. N-H. is supported by a PhD Scholarship from the French MESR (Ministère de l'Enseignement Supérieur et de la Recherche). The authors would like to thank the Proteomic Platform of IGBMC where the purchase of the LTQ Orbitrap Elite mass spectrometer was supported by the ARC Foundation, and the Proteomic Platform of IBMC (Institut de Biologie Moléculaire et Cellulaire) in Strasbourg, France, for their contribution to this project.

References

1. Purdy MD, Bennett BC, McIntire WE, Khan AK, Kasson PM, Yeager M (2014) Function and dynamics of macromolecular complexes explored by integrative structural and computational biology. Curr Opin Struct Biol 27:138–148.
2. Göbl C, Madl T, Simon B, Sattler M (2014) NMR approaches for structural analysis of multidomain proteins and complexes in solution. Prog Nucl Magn Reson Spectrosc 80:26–63.
3. Rodrigues JPGLM, Bonvin AMJJ (2014) Integrative computational modeling of protein interactions. FEBS J 281:1988–2003.
4. Heck AJR (2008) Native mass spectrometry: a bridge between interactomics and structural biology. Nat Methods 5:927–933.
5. Lanucara F, Holman SW, Gray CJ, Eyers CE (2014) The power of ion mobility-mass spectrometry for structural characterization and the study of conformational dynamics. Nat Chem 6:281–294.
6. Konermann L, Tong X, Pan Y (2008) Protein structure and dynamics studied by mass spectrometry: H/D exchange, hydroxyl radical labeling, and related approaches. J Mass Spectrom 43:1021–1036.
7. Merkley ED, Cort JR, Adkins JN (2013) Cross-linking and mass spectrometry methodologies to facilitate structural biology: finding a path through the maze. J Struct Funct Genomics 14:77–90.
8. Stengel F, Aebersold R, Robinson CV (2012) Joining Forces: Integrating Proteomics and Cross-linking with the Mass Spectrometry of Intact Complexes. Mol Cell Proteomics 11:R111.014027.
9. Singh P, Panchaud A, Goodlett DR (2010) Chemical Cross-Linking and Mass Spectrometry As a Low-Resolution Protein Structure Determination Technique. Anal Chem 82:2636–2642.
10. Leitner A, Walzthoeni T, Kahraman A, Herzog F, Rinner O, Beck M, Aebersold R (2010) Probing Native Protein Structures by Chemical Cross-linking, Mass Spectrometry, and Bioinformatics. Mol Cell Proteomics 9:1634–1649.
11. Rappsilber J (2011) The beginning of a beautiful friendship: Cross-linking/mass spectrometry and modelling of proteins and multi-protein complexes. J Struct Biol 173:530–540.
12. Murakami K, Elmlund H, Kalisman N, Bushnell DA, Adams CM, Azubel M, Elmlund D, Levi-Kalisman Y, Liu X, Gibbons BJ, Levitt M, Kornberg RD (2013) Architecture of an RNA polymerase II transcription pre-initiation complex. Science 342:1238724.
13. Müller MQ, Sinz A (2012) Chemical cross-linking and high-resolution mass spectrometry to study protein–drug interactions. Methods Mol Biol 803:205–218.
14. Kalisman N, Adams CM, Levitt M (2012) Subunit order of eukaryotic TRiC/CCT chaperonin by cross-linking, mass spectrometry, and combinatorial homology modeling. Proc Natl Acad Sci 109:2884–2889.
15. Lössl P, Kölbel K, Tänzler D, Nannemann D, Ihling CH, Keller MV, Schneider M, Zaucke F, Meiler J, Sinz A (2014) Analysis of nidogen-1/laminin γ1 interaction by cross-linking, mass spectrometry, and computational modeling reveals multiple binding modes. PLoS ONE 9:e112886.
16. Thierbach K, von Appen A, Thoms M, Beck M, Flemming D, Hurt E (2013) Protein Interfaces of the Conserved Nup84 Complex from Chaetomium thermophilum Shown by Crosslinking Mass Spectrometry and Electron Microscopy. Structure 21:1672–1682.

17. Kim SJ, Fernandez-Martinez J, Sampathkumar P, Martel A, Matsui T, Tsuruta H, Weiss TM, Shi Y, Markina-Inarrairaegui A, Bonanno JB, Sauder JM, Burley SK, Chait BT, Almo SC, Rout MP, Sali A. (2014) Integrative Structure–Function Mapping of the Nucleoporin Nup133 Suggests a Conserved Mechanism for Membrane Anchoring of the Nuclear Pore Complex. Mol. Cell. Proteomics 13:2911–2926.
18. Paramelle D, Miralles G, Subra G, Martinez J (2013) Chemical cross-linkers for protein structure studies by mass spectrometry. Proteomics 13:438–456.
19. Sutherland BW, Toews J, Kast J (2008) Utility of formaldehyde cross-linking and mass spectrometry in the study of protein–protein interactions. J Mass Spectrom 43:699–715.
20. Rhode BM, Hartmuth K, Urlaub H, Lührmann R (2003) Analysis of site-specific protein-RNA cross-links in isolated RNP complexes, combining affinity selection and mass spectrometry. RNA 9:1542–1551.
21. Fritzsche R, Ihling CH, Götze M, Sinz A (2012) Optimizing the enrichment of cross-linked products for mass spectrometric protein analysis. Rapid Commun Mass Spectrom 26:653–658.
22. Lascoux D, Paramelle D, Subra G, Heymann M, Geourjon C, Martinez J, Forest E (2007) Discrimination and selective enhancement of signals in the MALDI mass spectrum of a protein by combining a matrix-based label for lysine residues with a neutral matrix. Angew Chem Int Ed 46:5594–5597.
23. Paramelle D, Cantel S, Enjalbal C, Amblard M, Forest E, Heymann M, Geourjon C, Martinez J, Subra G (2009) A new generation of cross-linkers for selective detection by MALDI MS. Proteomics 9:5384–5388.
24. Götze M, Pettelkau J, Schaks S, Bosse K, Ihling CH, Krauth F, Fritzsche R, Kühn U, Sinz A (2011) StavroX—A Software for Analyzing Crosslinked Products in Protein Interaction Studies. J Am Soc Mass Spectrom 23:76–87.
25. Rasmussen MI, Refsgaard JC, Peng L, Houen G, Højrup P (2011) CrossWork: Software-assisted identification of cross-linked peptides. J Proteomics 74:1871–1883.
26. Yang B, Wu YJ, Zhu M, Fan SB, Lin J, Zhang K, Li S, Chi H, Li YX, Chen HF, Luo SK, Ding YH, Wang LH, Hao Z, Xiu LY, Chen S, Ye K, He SM, Dong MQ (2012) Identification of cross-linked peptides from complex samples. Nat Methods 9:904–906.
27. Rinner O, Seebacher J, Walzthoeni T, Mueller L, Beck M, Schmidt A, Mueller M, Aebersold R (2008) Identification of cross-linked peptides from large sequence databases. Nat Methods 5:315–318.
28. Grant PA, Duggan L, Côté J, Roberts SM, Brownell JE, Candau R, Ohba R, Owen-Hughes T, Allis CD, Winston F, Berger SL, Workman JL (1997) Yeast Gcn5 functions in two multisubunit complexes to acetylate nucleosomal histones: characterization of an Ada complex and the SAGA (Spt/Ada) complex. Genes Dev 11: 1640–1650.
29. Grünberg S, Hahn S (2013) Structural insights into transcription initiation by RNA polymerase II. Trends Biochem Sci 38:603–611.
30. Hahn S, Young ET (2011) Transcriptional regulation in Saccharomyces cerevisiae: transcription factor regulation and function, mechanisms of initiation, and roles of activators and coactivators. Genetics 189:705–736.
31. Nagy Z, Tora L (2007) Distinct GCN5/PCAF-containing complexes function as co-activators and are involved in transcription factor and global histone acetylation. Oncogene 26:5341–5357.
32. Spedale G, Timmers HTM, Pijnappel WWMP (2012) ATAC-king the complexity of SAGA during evolution. Genes Dev 26:527–541.
33. Eisenmann DM, Chapon C, Roberts SM, Dollard C, Winston F (1994) The Saccharomyces cerevisiae SPT8 gene encodes a very acidic protein that is functionally related to SPT3 and TATA-binding protein. Genetics 137:647–657.
34. Brand M, Leurent C, Mallouh V, Tora L, Schultz P (1999) Three-dimensional structures of the TAFII-containing complexes TFIID and TFTC. Science 286: 2151–2153.
35. Wu P-YJ, Ruhlmann C, Winston F, Schultz P (2004) Molecular Architecture of the S. cerevisiae SAGA Complex. Mol Cell 15:199–208.
36. Lee KK, Sardiu ME, Swanson SK, Gilmore JM, Torok M, Grant PA, Florens L, Workman JL, Washburn MP (2011) Combinatorial depletion analysis to assemble the network architecture of the SAGA and ADA chromatin remodeling complexes. Mol Syst Biol 7:503.
37. Bian C, Xu C, Ruan J, Lee KK, Burke TL, Tempel W, Barsyte D, Li J, Wu M, Zhou BO, Fleharty BE, Paulson A, Allali-Hassani A, Zhou JQ, Mer G, Grant PA, Workman JL, Zang J, Min J (2011) Sgf29 binds histone H3K4me2/3 and is required for SAGA complex recruitment and histone H3 acetylation. EMBO J 30: 2829–2842.
38. Durant M, Pugh BF (2006) Genome-wide relationships between TAF1 and histone acetyltransferases in Saccharomyces cerevisiae. Mol Cell Biol 26:2791–2802.
39. Hassan AH, Prochasson P, Neely KE, Galasinski SC, Chandy M, Carrozza MJ, Workman JL (2002) Function and selectivity of bromodomains in anchoring chromatin-modifying complexes to promoter nucleosomes. Cell 111:369–379.
40. Vermeulen M, Eberl HC, Matarese F, Marks H, Denissov S, Butter F, Lee KK, Olsen JV, Hyman AA, Stunnenberg HG, Mann M (2010) Quantitative interaction proteomics and genome-wide profiling of epigenetic histone marks and their readers. Cell 142:967–980.
41. Trowitzsch S, Bieniossek C, Nie Y, Garzoni F, Berger I (2010) New baculovirus expression tools for recombinant protein complex production. J Struct Biol 172:45–54.
42. Lee YJ, Lackner LL, Nunnari JM, Phinney BS (2007) Shotgun Cross-Linking Analysis for Studying Quaternary and Tertiary Protein Structures. J Proteome Res 6:3908–3917.
43. Wenzel R, Röhling U, Nazabal A, Hillenkamp F (2009) A detector device for high mass ion detection, a method for analyzing ions of high mass and a device for selection between ion detectors. Available from: Patent WO/2009/086642.
44. Müller DR, Schindler P, Towbin H, Wirth U, Voshol H, Hoving S, Steinmetz MO (2001) Isotope-tagged cross-linking reagents. A new tool in mass spectrometric protein interaction analysis. Anal Chem 73:1927–1934.
45. Gaucher SP, Hadi MZ, Young MM (2006) Influence of crosslinker identity and position on gas-phase dissociation of Lys–Lys crosslinked peptides. J Am Soc Mass Spectrom 17:395–405.
46. Hilton GR, Benesch JLP (2012) Two decades of studying non-covalent biomolecular assemblies by means of electrospray ionization mass spectrometry. J R Soc Interface 9:801–816.
47. Mann M, Hendrickson RC, Pandey A (2001) Analysis of proteins and proteomes by mass spectrometry. Annu Rev Biochem 70:437–473.

48. Schubert P, Hoffman MD, Sniatynski MJ, Kast J (2006) Advances in the analysis of dynamic protein complexes by proteomics and data processing. Anal Bioanal Chem 386:482–493.
49. Candau R, Berger SL (1996) Structural and functional analysis of yeast putative adaptors. Evidence for an adaptor complex in vivo. J Biol Chem 271:5237–5245.
50. Kurabe N, Katagiri K, Komiya Y, Ito R, Sugiyama A, Kawasaki Y, Tashiro F (2007) Deregulated expression of a novel component of TFTC/STAGA histone acetyltransferase complexes, rat SGF29, in hepatocellular carcinoma: Possible implication for the oncogenic potential of c-Myc. Oncogene 26:5626–5634.
51. Ito T, Chiba T, Ozawa R, Yoshida M, Hattori M, Sakaki Y (2001) A comprehensive two-hybrid analysis to explore the yeast protein interactome. Proc Natl Acad Sci U S A 98:4569–4574.
52. Rolland T, Taşan M, Charloteaux B, Pevzner SJ, Zhong Q, Sahni N, Yi S, Lemmens I, Fontanillo C, Mosca R, Kamburov A, Ghiassian SD, Yang X, Ghamsari L, Balcha D, Begg BE, Braun P, Brehme M, Broly MP, Carvunis AR, Convery-Zupan D, Corominas R, Coulombe-Huntington J, Dann E, Dreze M, Dricot A, Fan C, Franzosa E, Gebreab F, Gutierrez BJ, Hardy MF, Jin M, Kang S, Kiros R, Lin GN, Luck K, MacWilliams A, Menche J, Murray RR, Palagi A, Poulin MM, Rambout X, Rasla J, Reichert P, Romero V, Ruyssinck E, Sahalie JM, Scholz A, Shah AA, Sharma A, Shen Y, Spirohn K, Tam S, Tejeda AO, Trigg SA, Twizere JC, Vega K, Walsh J, Cusick ME, Xia Y, Barabási AL, Iakoucheva LM, Aloy P, De Las Rivas J, Tavernier J, Calderwood MA, Hill DE, Hao T, Roth FP, Vidal M (2014) A proteome-scale map of the human interactome network. Cell 159:1212–1226.
53. Rual J-F, Venkatesan K, Hao T, Hirozane-Kishikawa T, Dricot A, Li N, Berriz GF, Gibbons FD, Dreze M, Ayivi-Guedehoussou N, Klitgord N, Simon C, Boxem M, Milstein S, Rosenberg J, Goldberg DS, Zhang LV, Wong SL, Franklin G, Li S, Albala JS, Lim J, Fraughton C, Llamosas E, Cevik S, Bex C, Lamesch P, Sikorski RS, Vandenhaute J, Zoghbi HY, Smolyar A, Bosak S, Sequerra R, Doucette-Stamm L, Cusick ME, Hill DE, Roth FP, Vidal M (2005) Towards a proteome-scale map of the human protein–protein interaction network. Nature 437:1173–1178.
54. Fischer L, Chen ZA, Rappsilber J (2013) Quantitative cross-linking/mass spectrometry using isotope-labelled cross-linkers. J Proteomics 88:120–128.
55. Merkley ED, Rysavy S, Kahraman A, Hafen RP, Daggett V, Adkins JN (2014) Distance restraints from crosslinking mass spectrometry: Mining a molecular dynamics simulation database to evaluate lysine–lysine distances. Protein Sci 23:747–759.
56. Gamper AM, Kim J, Roeder RG (2009) The STAGA subunit ADA2b is an important regulator of human GCN5 catalysis. Mol Cell Biol 29:266–280.
57. Han Y, Luo J, Ranish J, Hahn S (2014) Architecture of the Saccharomyces cerevisiae SAGA transcription coactivator complex. EMBO J 33:2534–2546.
58. Costanzo M, Baryshnikova A, Bellay J, Kim Y, Spear ED, Sevier CS, Ding H, Koh JL, Toufighi K, Mostafavi S, Prinz J, St Onge RP, VanderSluis B, Makhnevych T, Vizeacoumar FJ, Alizadeh S, Bahr S, Brost RL, Chen Y, Cokol M, Deshpande R, Li Z, Lin ZY, Liang W, Marback M, Paw J, San Luis BJ, Shuteriqi E, Tong AH, van Dyk N, Wallace IM, Whitney JA, Weirauch MT, Zhong G, Zhu H, Houry WA, Brudno M, Ragibizadeh S, Papp B, Pál C, Roth FP, Giaever G, Nislow C, Troyanskaya OG, Bussey H, Bader GD, Gingras AC, Morris QD, Kim PM, Kaiser CA, Myers CL, Andrews BJ, Boone C (2010) The genetic landscape of a cell. Science 327:425–431.
59. Bradford MM (1976) A rapid and sensitive method for the quantitation of microgram quantities of protein utilizing the principle of protein–dye binding. Anal Biochem 72:248–254.
60. Laemmli UK (1970) Cleavage of Structural Proteins during the Assembly of the Head of Bacteriophage T4. Nature 227:680–685.
61. Burnette WN (1981) "Western Blotting": Electrophoretic transfer of proteins from sodium dodecyl sulfate-polyacrylamide gels to unmodified nitrocellulose and radiographic detection with antibody and radioiodinated protein A. Anal Biochem 112:195–203.

3.1.5. Structural analysis of recombinant SAGA HAT module by negative stain EM.

We have used negative stain EM to analyze the purified recombinant SAGA HAT module. Since the subcomplex displayed high aggregation tendency during sample preparation for EM (data not shown), we have proceeded with glycerol gradient centrifugation coupled with mild cross-linking (GraFix, see Kastner et al., 2008). This step has allowed to separate the subcomplex from the aggregates and improve its stability. Considering the fact that there is no known structure of the HAT module, we have decided to perform an RCT analysis (see part 2.3.6.5) to obtain an initial model of the module. During this analysis, we have collected several hundred tilt pair images with 45° angle difference (fig. 49) and manually picked 6813 particle pairs. Selected SAGA HAT particles significantly varied in size from 9 to 14 nm. Subsequent 2D classification of untilted particles led to class averages of different size and shape, indicating a highly heterogeneous particle population (fig. 50).

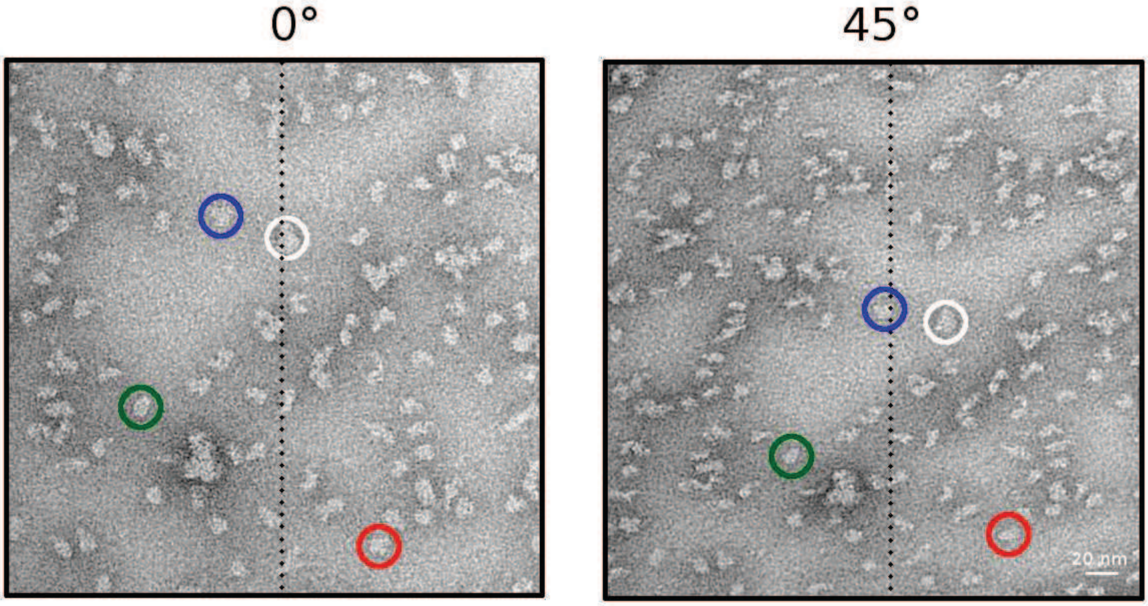

Fig. 49. Tilt-pair images of negatively stained SAGA HAT module.

The image on the right is 45° tilted with respect to the left image. Tilt axis is shown by a dotted line, tilt-pair particles are shown by circles. Scale bar is 20 nm.

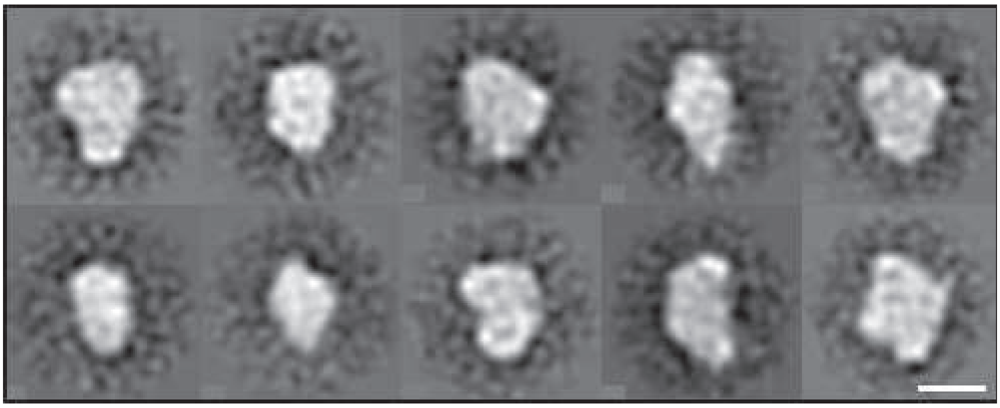

Fig. 50. Representative 2D class averages of the untilted SAGA HAT data set.

2D class averages demonstrate evident heterogeneity of the data set. Scale bar is 10 nm.

Tilted particle images, corresponding to 150 good 2D classes, were selected and 3D reconstructions have been produced for each class. In the next step, all 3D structures were subjected to 3D alignment and classification to compensate for the missing cone of information and separate different conformations. In the end, three most distinct 3D classes were iteratively refined until no improvement of the structure was noticed (fig. 51). Unfortunately, the final 3D structures had a resolution of 30-40 Å only and did not show distinctive structural features. The most compact structure had a globular shape of about 10 nm in diameter, which is consistent with the molecular weight of the SAGA HAT complex (241 kDa).

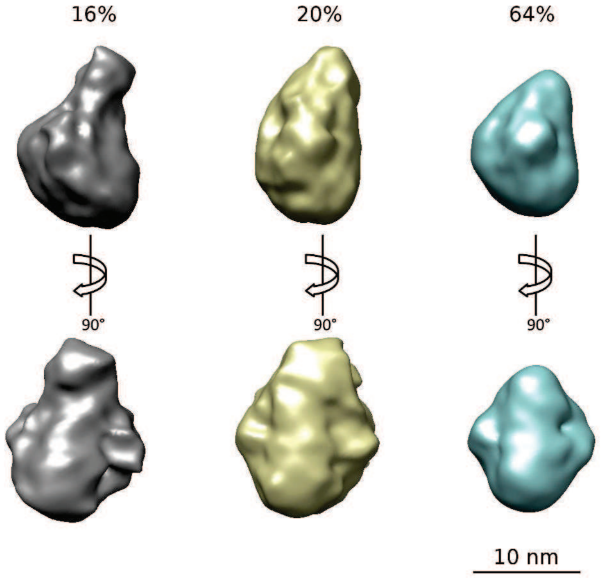

Fig. 51. Three most distinct 3D classes of SAGA HAT data set are shown.

Subsequently, we have collected a larger untilted data set (fig. 52) and semi-automatically selected about 24.000 single particles. The 2D class averages, obtained after 2D classification, displayed similar features as before, ranging in size and shape (fig. 53). We have used the most abundant SAGA HAT 3D structure (fig. 51, 64% of the RCT data set), obtained in the previous step, as an initial model for 3D refinement and reconstruction. Unfortunately, we could not improve the resolution of this model using the new large data set.

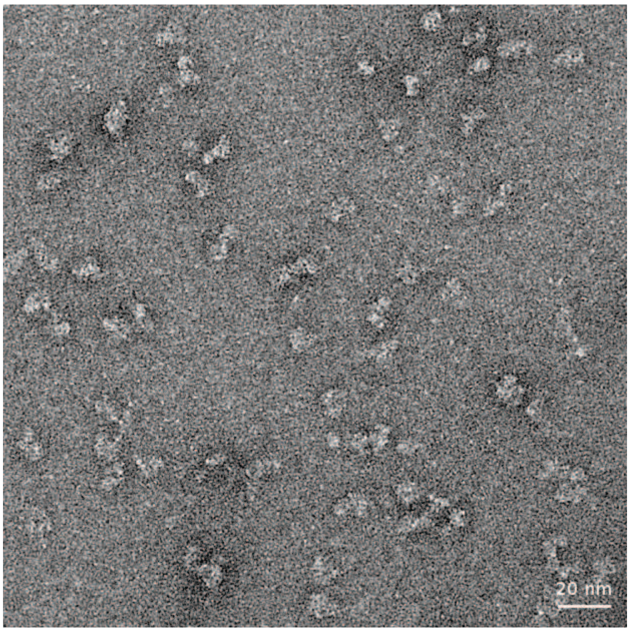

Fig. 52. A typical micrograph of negatively stained SAGA HAT subcomplex.

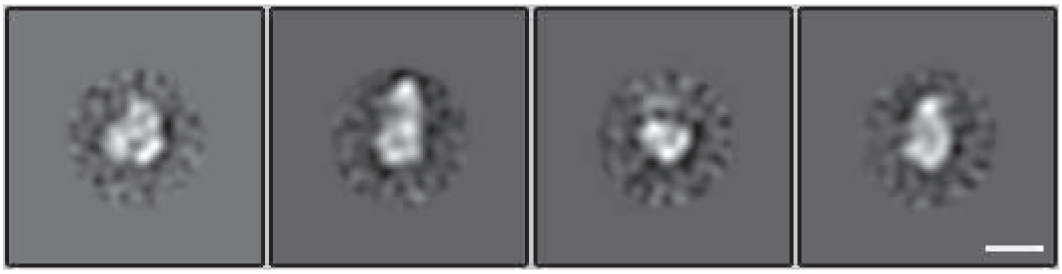

Fig. 53. Several distinct 2D class averages obtained from a larger HAT data set.

The size of HAT particles varies from 8 x 7 nm (3rd *image in the row) to 15 x 7 nm (2*nd *image in the row). Scale bar is 10 nm.*

3.1.6. Localization of the HAT module in *S. cerevisiae* SAGA complex.

3.1.6.1. Purification and characterization of ΔAda2 mutant SAGA complex.

A protocol for SAGA purification from *S. cerevisiae*, based on TAP tag method, was developed by A. Durand in our group (Durand et al., 2014). This protocol was used for the purification of Ada1-TAP-tagged SAGAΔAda2 mutant. Previous studies have shown that Ada2 deletion detaches the HAT module from SAGA, leaving the rest of the complex intact (Lee et al., 2011). The amount of purified complex from 6 liters (14 g) of cells was roughly estimated at 2-3 µg. The sample composition was checked by SDS-PAGE and LC-MS/MS and its structural integrity was assessed by electron microscopy observation using negative staining. The purified proteins were loaded onto a 4-15% acrylamide gradient gel, that was silver stained after running the SDS-PAGE (fig. 54). The gel revealed multiple bands corresponding to the multiprotein SAGA complex as observed before (Durand et al., 2014). Both in WT and mutant strains (SAGAΔAda2), the Spt7 protein migrated as two bands, corresponding to full length Spt7 protein and a truncated version, where the C-terminus was cleaved. The latter form is known to be incorporated into SLIK complex. Thus, both SAGA and SLIK were present in the purified samples in roughly equivalent amounts.

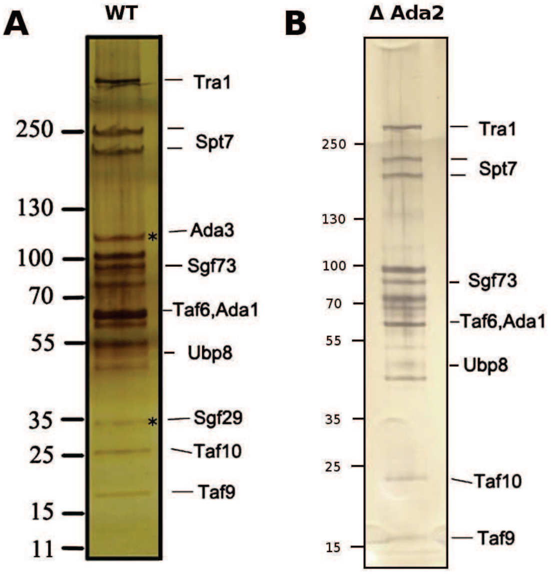

Fig. 54. Silver stained SDS-PAGE gels of purified S. cerevisiae SAGA complexes.

(a) WT SAGA. Note the presence of two Spt7 bands. (b) ΔAda2 mutant, in which HAT module is removed. Asterisk (*) indicates proteins present only in the wild-type SAGA. Panel (a) is taken from Durand et al., 2014.

From the SDS-PAGE gel it can be deduced that the Ada3 and Sgf29 subunits, known to be a part of the HAT module, are clearly absent in the ΔAda2 mutant strain. Ada2 and Gcn5 subunits have molecular weight very similar to other SAGA subunits, thus their identification is not straightforward. MS analysis has confirmed that indeed all four HAT subunits were missing in ΔAda2 strain (Table 13).

Subunit name	Coverage, %	Number of unique peptides	Molecular weight, kDa
Tra1	30.96	103	432.9
Spt7	55.41	77	152.5
Taf5	53.38	39	88.9
Taf12	61.78	32	61.0
Spt20	69.54	28	67.8
Sgf73	54.34	25	72.8
Ada1	42.62	18	54.4
Spt8	44.19	17	66.1
Taf6	35.27	17	57.9
Ubp8	39.70	14	53.6
Spt3	30.56	11	38.8
Taf9	56.69	7	17.3
Taf10	33.50	6	23.0
Sus1	63.54	4	11.0
Sgf11	39.39	3	11.3

Table 13. Sequence coverage of SAGAΔAda2 complex, identified by MS.
No peptides corresponding to Gcn5, Ada2, Ada3 or Sgf29 were detected.

3.1.6.2. *Negative stain EM analysis of ΔAda2 mutant SAGA complex and HAT localization.*

To determine the position of the HAT module inside the full SAGA complex, we have obtained the structure of *S. cerevisiae* SAGA without HAT. The purified mutant complex was negatively stained and analyzed by EM. After supervised particle picking, about 60.000 particles were subjected to 2D classification and alignment (see Methods). The 2D class averages have revealed a known two-lobe structure, indicating that SAGA complex integrity was not affected by Ada2 deletion (fig. 55). We have compared our results with previous study of the WT *S. cerevisiae* SAGA complex performed by A. Durand in our lab. In both cases, SAGA complexes were purified using the same TAP tag method. Mutant SAGA adopted preferential orientations on carbon film as did the wild-type (WT) complex. Most of the differences between the WT and mutant data sets were seen in the lobe B: in comparison with the wild-type, no class averages showing the domain V were detected, also the whole lobe B seemed to adopt different orientations.

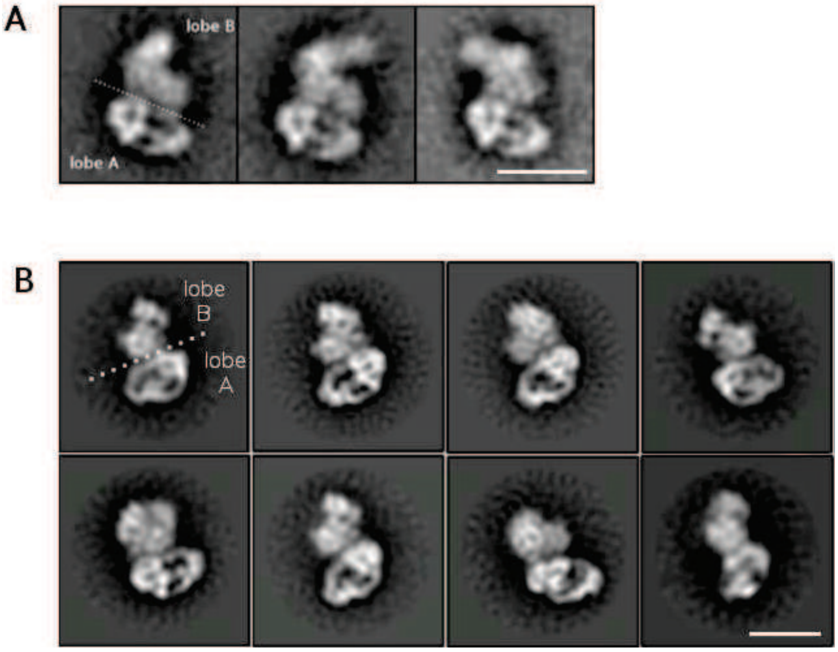

Fig. 55. Representative 2D class averages of WT SAGA (a) in comparison to ΔAda2 (b) mutant.

Lobe B displays large range of different conformations, however the domain V is not visible. Panel A is kindly provided by A. Durand. Scale bar is 20 nm.

We have proceeded with a 3D reconstruction of the mutant SAGA and obtained a model at 27 Å resolution. This model displayed almost no distinct features, especially in the lobe B of the complex (fig. 56).

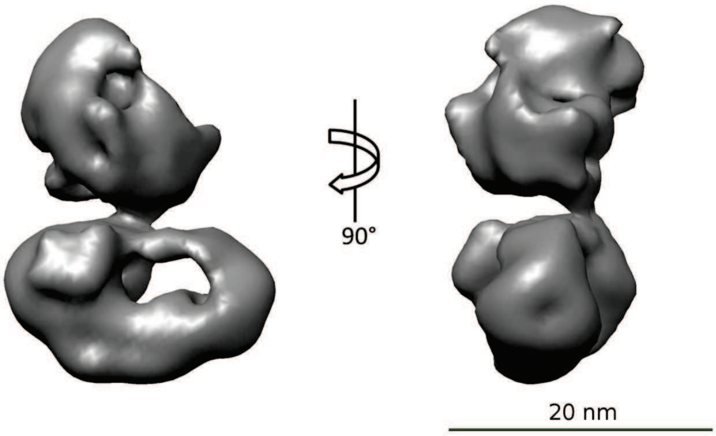

Fig. 56. 3D reconstruction of the mutant SAGA complex.

We have thought that the low resolution of the lobe B comes from its high flexibility, thus to sort out different conformations we have performed a 3D classification. Unfortunately, the 3D structures of different classes did not reveal separate conformations: all classes looked similar (data not shown).

Therefore, we have first refined the lobe A separately from the lobe B, using a soft mask. The masked refinement has slightly improved the resolution of the lobe A, confirming that this lobe is stable and displays no signs of heterogeneity. Next, we took the alignment parameters from the lobe A refinement and run a 3D classification on the whole complex. Since the lobe A was fixed, this has allowed to sort the different conformations of lobe B relatively to the lobe A. This analysis has shown that the position of the lobe B varied by more than 10° with respect to the lobe A, changing significantly the connection between the domains II and IV (fig. 57).

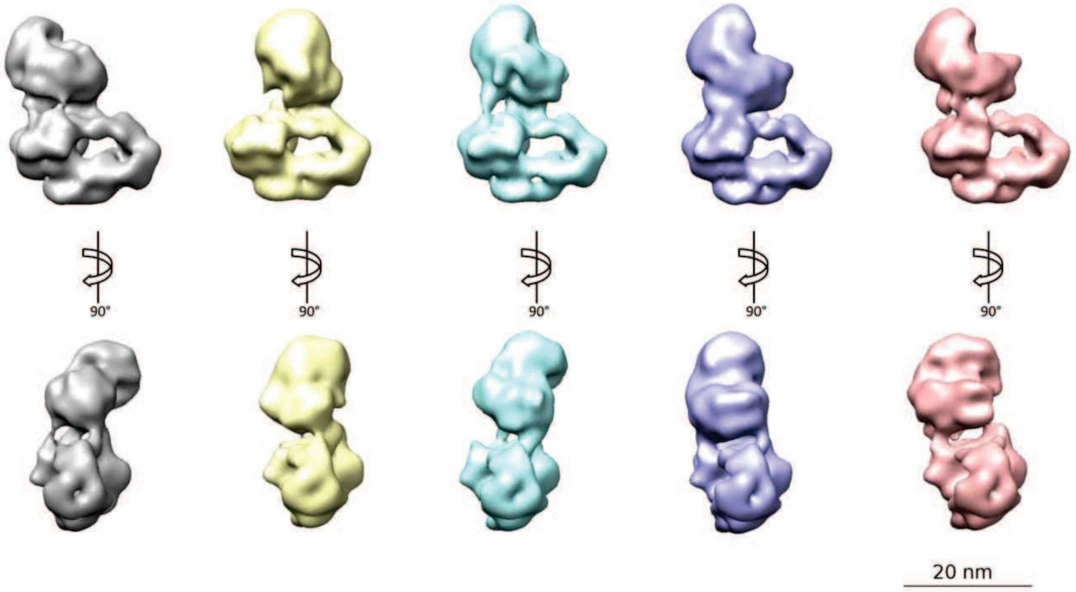

Fig. 57. The lobe A of the mutant complex was fixed and the 3D classification without alignment has been done.

This has allowed to separate the lobe B conformations, revealing its large flexibility.

We have compared the 3D structures of the mutant and the WT SAGA. The mutant structure looked similar to the WT structure from A. Durand (2014) with the exception of lobe B, that showed several density differences (fig. 58).

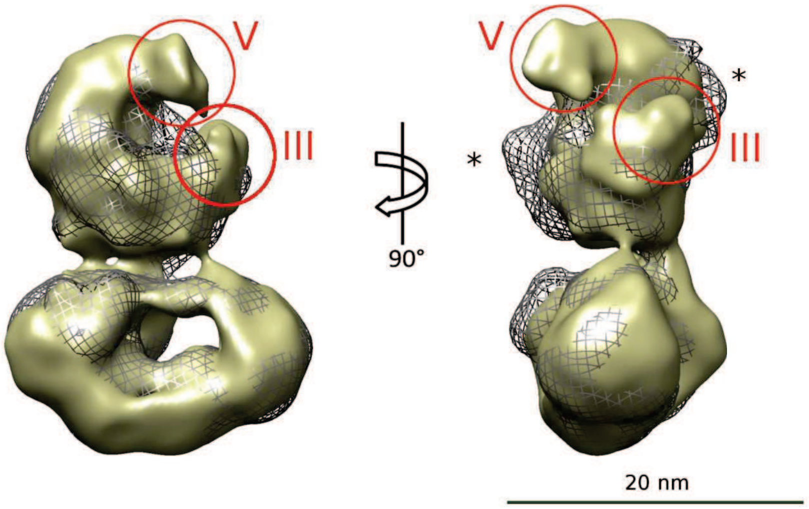

Fig. 58. Comparison of the SAGA structures: the mutant without HAT (gray mesh) and WT SAGA (yellow, from Durand et al., 2014).
Major differences are visible in the domains III and V. Several minor differences (marked by the asterisks) between the 3D maps are also present.

In comparison with the WT structure, two major missing densities were spotted: one corresponding to the position of the dUb module (in the domain III), as determined by A. Durand, and another one, located at the most flexible part of the complex – in the domain V. Mass spectrometry did not reveal the loss of dUb module, so we presume that it might be lost during sample preparation step. Another possibility is that the large conformational flexibility of the lobe B has obscured the identification of missing densities in this lobe. Durand et al. have shown that the dUb module was not very pronounced when the SAGA was analyzed as a whole and appeared more clearly only when the lobe B was refined separately from the rest of the complex. Unfortunately, we did not succeed in separate alignment of the lobe B of the mutant complex to be able to compare it with the WT SAGA, since the model of this lobe displayed very few structural features.

We also cannot exclude the possibility that the removal of the HAT module has induced a conformational change in the lobe B that lead to dissociation of the dUb module. In addition, several minor densities appeared in the mutant structure (fig. 58, marked by asterisks). We suggest that they can be a result of a conformational rearrangement of the lobe B upon HAT removal. Altogether, we are convinced that the HAT module is located in the domain V: the other minor density differences between WT SAGA and the mutant cannot accommodate the size of the HAT module, which is at least 10 nm long. The size of the domain V is approximately 5 x 10 nm, that match the size of the HAT module, determined previously in the EM study (see part 3.1.5).

3.2. Structure of the *P. pastoris* SAGA complex.

3.2.1. Purification and characterization of *P. pastoris* SAGA complex.

During the course of this work a new scheme for the isolation of low abundant multiprotein nuclear complexes has been developed by A. Ben-Shem in our group. Several key elements are listed here:

- (i) switching to *P. pastoris* has allowed to produce much more starting material: 1-2 liters of cell culture gave 20-30 g of cells;
- (ii) careful nuclear extract preparation has allowed to avoid breakage of yeast vacuoles and minimize release of proteases;
- (iii) high-affinity streptavidin-binding peptide was introduced on Sgf73 subunit, ensuring the retention of the dUb module with the rest of SAGA complex during purification;
- (iv) PEG precipitation has enabled efficient separation of SAGA complex from proteases, cell debris and other small protein complexes;
- (v) working at 4C during all steps of purification has assured the highest sample quality;
- (vi) minimization of working volumes (for solubilization etc.) and keeping high protein concentration have led to decrease of sample loss and subunit dissociation at each step of the purification.

The overall yield of the purification was boosted up to 30-70 ug of SAGA complex (from 1-2 liters of yeast cells, corresponding to 20-30 grams of cells). SDS-PAGE gel of the purified complex is shown on fig. 59. The new protocol has allowed to avoid Spt7 cleavage by proteases, preventing co-purification of the SLIK complex. That protocol was used for all further purifications of wild-type SAGA, as well as for the assembly of the SAGA-TBP complex.

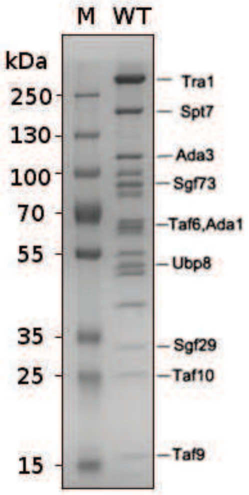

Fig. 59. SDS-PAGE gel of P. pastoris WT SAGA complex, purified by a new method.
Note the presence of a single band corresponding to Spt7, indicating that the SLIK complex was not co-purified. Also, all subunits of the dUb module are retained. Image is kindly provided by A. Ben-Shem.

3.2.2. Structure of SAGA complex by negative stain EM.

Electron micrographs of SBP-tag purified, cross-linked, carbon-adsorbed, negatively stained *P. pastoris* SAGA complex have revealed particles with an elongated shape and a size characteristic to SAGA. The sample displayed minimal aggregation and the stain quality was good, since single particles had high contrast and were fully embedded in uranyl acetate stain, showing no signs of dark halo around the edges of particles. The average size of the SAGA particles was about 27 nm along the long axis and 19 nm in the perpendicular direction (fig. 60). In particular, no dissociation of the two lobes of SAGA was observed, suggesting that the new purification protocol yields a more stable complex.

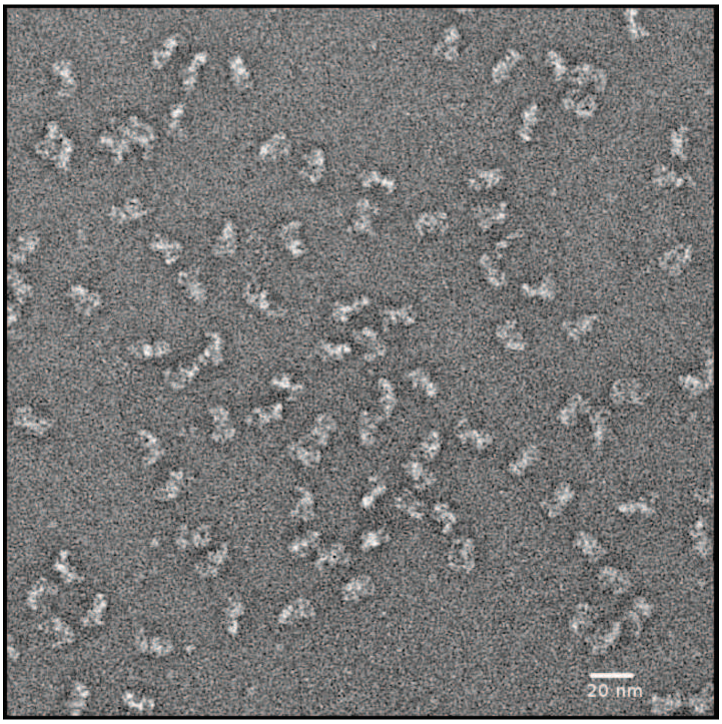

Fig. 60. A typical micrograph of negatively stained SAGA complex.

All micrographs underwent CTF estimation, the average defocus values were 0.8 – 1.7 µm underfocus. A few thousand particles were manually picked and subjected to iterative reference-free 2D alignment and classification with 5-10 classes in order to obtain distinctive views of SAGA complex. Some representative 2D class averages were then used as references for automatic particle picking using the in-house developed gEMpicker software (Hoang et al., 2013), bringing the total number of particles to 58.024. Particle box size and mask size were chosen appropriately to avoid including the neighboring particles and at the same time leaving enough background area around the particles for correct noise estimation. The full data set was subjected to 2D classification and alignment, the main purpose of this step was to identify different views of SAGA, to estimate a portion of preferentially oriented particles adsorbed on the carbon, and to eliminate bad particles from further analysis. To increase the computational speed, at the beginning, the particles were coarsened twice, thus reducing box size and increasing pixel size. Fig. 61 shows typical 2D class averages obtained after analysis of this data set. The complex consists of five domains assembled into two large lobes: the lobe A of 20 by 10 nm size, with the domains I and II; and the lobe B with the domains III, IV and V. The lobe A shows high resolution details and displays a preferential orientation in which the particle maximizes surface contact with the carbon layer of the EM grid. In this view, the fork-shaped domain I and the domain II are clearly distinguished. On the contrary, the lobe B shows lower resolution and adopts different conformations, ranging from round shape of 13 nm in diameter to L-shape of 20 nm long (fig. 61). Most of the time the domain V adopts different positions and seems to be the most flexible part of SAGA as evidenced by its fuzzy appearance. It is difficult to establish whether in some particular views it is absent or simply hidden from observation. We conclude that SAGA adopts a strongly preferred orientation when adsorbed on a carbon film, since one particular view (fig. 61) is overrepresented and includes more than 50% of all particles. Though the 2D analysis provides a powerful means to analyze sample heterogeneity, most of the time it is difficult to distinguish separate conformations (hence different 3D structures) from different orientations of the same 3D structure.

Fig. 61. Some representative 2D class averages of the SAGA complex.

Two lobes can be distinguished: lobe A (domains I and II) and lobe B (domains III, IV, V). Most of the particles adopt the same orientation, however the domain V is flexible. Asterisk (*) marks a clearly closed conformation of the lobe B. Box size is 50 nm.

Since the complex has displayed high heterogeneity, we have decided to perform 3D reconstruction, refinement and classification, using a strongly low-pass filtered structure of SAGA (Durand et al., 2014) as an initial reference, to sort out different conformations. The number of 3D classes was optimized empirically, so that the most abundant conformations could be sorted out. In good agreement with the 2D classification, lobe A did not display significant heterogeneity during 3D classification (fig. 62). However, lobe B revealed tremendous differences within the SAGA data set: (i) the connection between the lobes mediated by the domains II-IV and I-III varies, probably due to flexibility of lobe B with respect to lobe A; (ii) domains III and V form a molecular clamp that can adopt different states from wide open through intermediate to almost closed; (iii) the protein ring formed by the extensions of the domains III and V is not visible in the open conformations and highly pronounced in the closed conformation. We have estimated that 52% of the particles display a clearly open conformation and 34% show a fully closed lobe B. Due to the significant decrease of the number of single particles in each class the resolution of each model has deteriorated, and we have decided to refine separately the two most distinct conformations of SAGA complex: the open and the closed one.

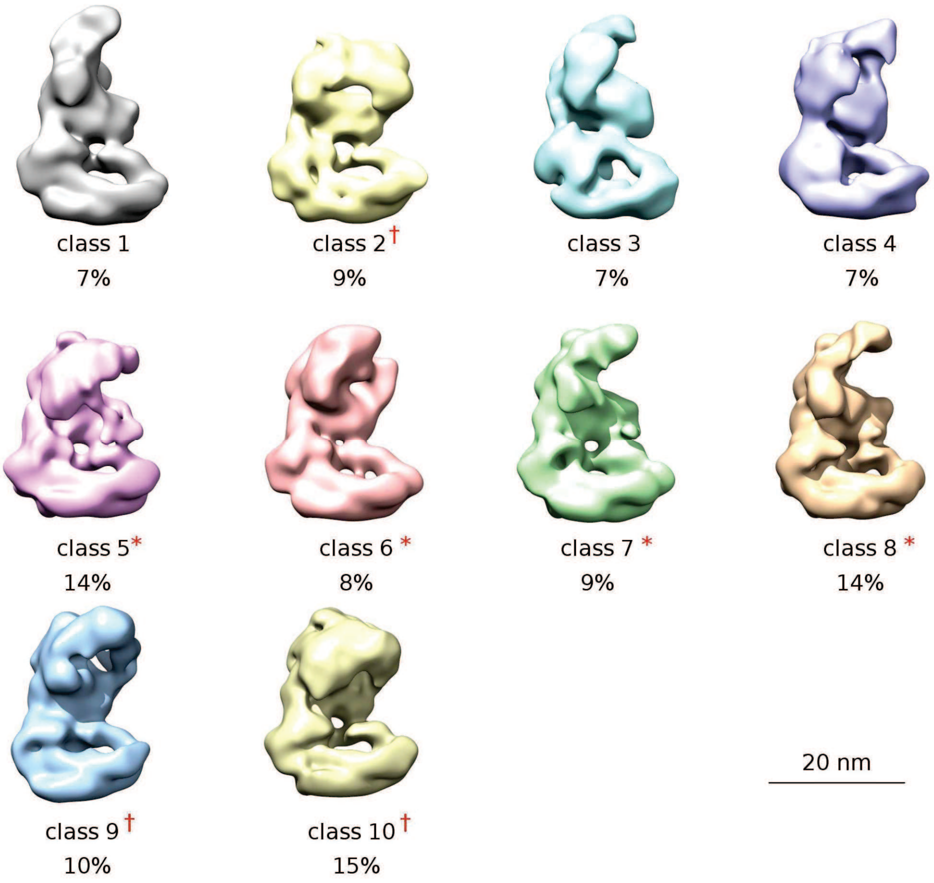

Fig. 62. 3D classification has revealed different conformations of the lobe B.

Percent values reflect the number of particles in each class relatively to the total data set. Asterisk (*) marks the classes assigned to "open state" and the cross (†) marks the "closed state" classes.

We have joined particles from classes 5, 6, 7, 8 and assigned them to an “open state”, while the particles from classes 2, 9, 10 – to a “closed state” (fig. 62). Class 1 was distinct from the other open classes and was not refined further, classes 3 and 4 were also discarded due to their bad quality. The two data sets (“open” and “closed”) were iteratively refined in 3D and produced two structures of similar resolution of about 26 Å (fig. 63). Both models displayed excellent structural overlay with the exception of the protein ring formed by the domains III and V, that seems to be missing in the open state. We suggest that the additional density (green in fig. 63) appearing in the closed conformation of the SAGA is the domain V, that becomes stabilized upon closing of the molecular clamp. In the open conformation, the domain V extension is highly flexible and might be averaged out in our reconstructions.

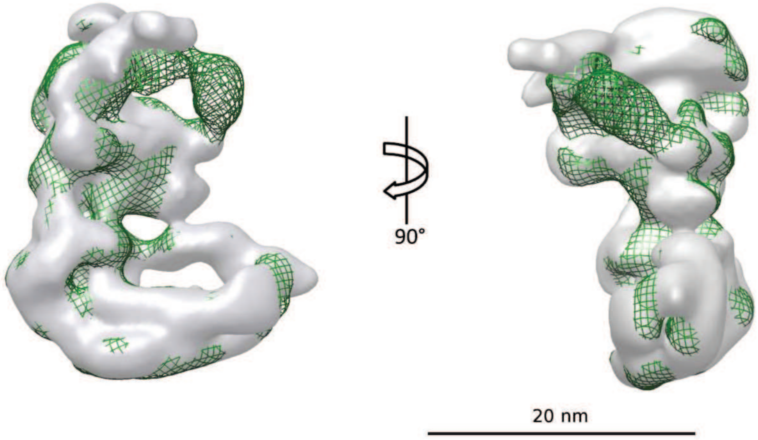

Fig. 63. Comparison between two refined SAGA conformations: open (gray) and closed (green).

Closing of the molecular clamp is mediated by the domains III and V of the lobe B.

We have compared our resulting 3D models (fig. 64) to the existing ones, obtained previously in the lab (Wu et al., 2004; Durand et al., 2014). The new models derived from the *P. pastoris* SAGA preparation clearly display higher resolution than the previous models obtained from *S. cerevisiae* and therefore more details can be observed in both lobes of the complex. As in the model of A. Durand, the two domains I and II of the lobe A are clearly connected by three long links, while lobe B forms a clamp composed of the domains III and V. However, the lobe B looks bigger than in the previous models and the connection between the two lobes is much better defined, notably between the domains II and IV. The difference in the size of the lobe B might be explained by either a better preservation/stoichiometry of all subunits of the complex using the new purification method or smaller conformational flexibility of the lobe B. Indeed, if different positions of lobe B are averaged, the outermost extensions disappear.

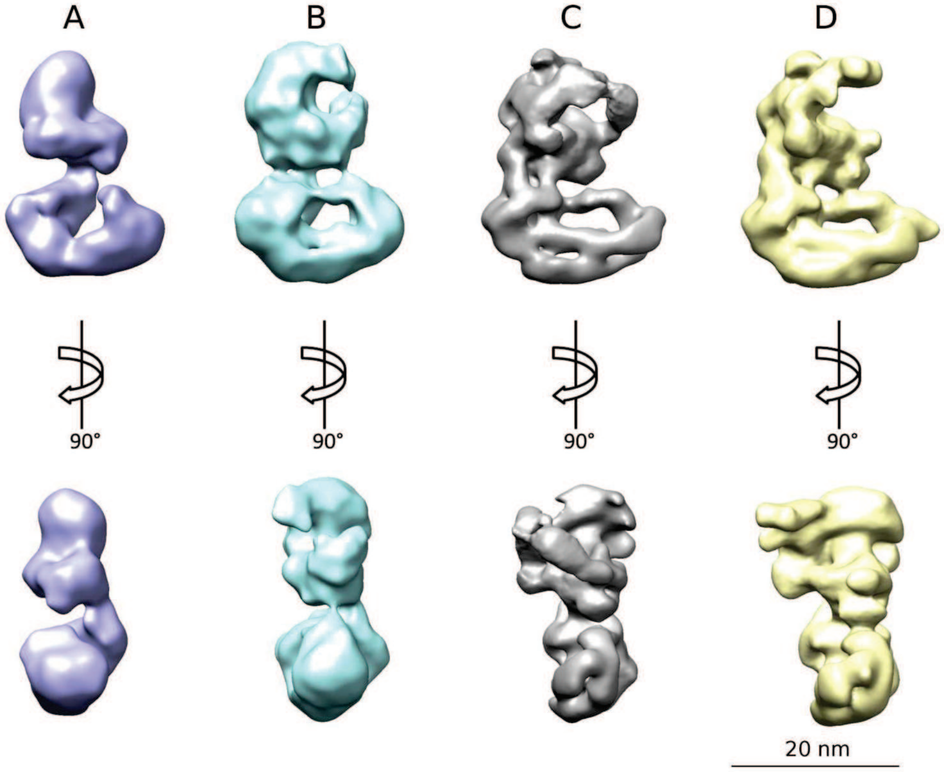

Fig. 64. Gradual improvement in resolution of SAGA structure, obtained by negative staining EM, over the years in our lab.

(a) First 3D model of SAGA (Wu et al., 2004). (b) More recent 3D model from Durand et al., 2014. (c, d) Current 3D models of SAGA (this work), displaying more high resolution features. Two conformations of the lobe B can be clearly distinguished: closed (c) and open (d).

At this point, we have positioned the X-ray structure of the dUb module into the SAGA models in both conformations (fig. 65), based on the previously determined location of this module (Durand et al., 2014). The proposed position of the dUb is based on the residues of Ubp8 near the ubiquitin-binding site, forming a region composed of 3 β-strands. However, a high-resolution information is necessary to confirm this orientation of the dUb module.

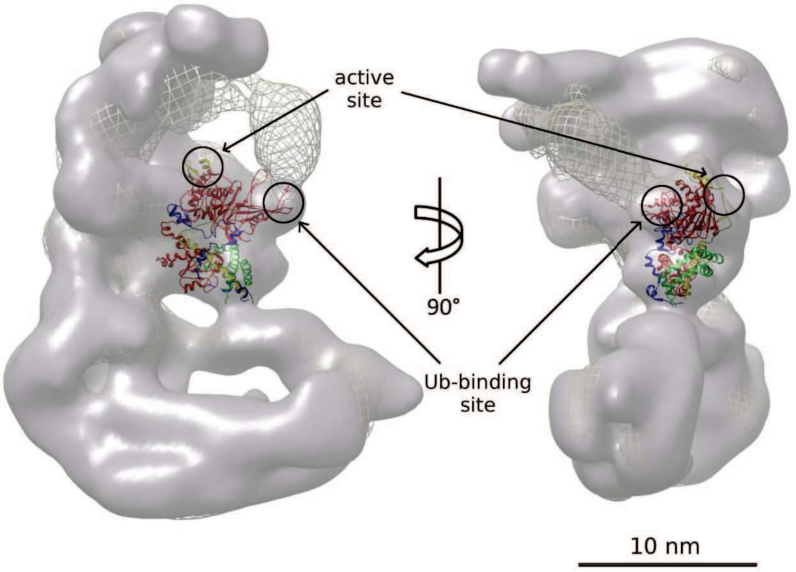

Fig. 65. Fitting of the dUb crystal structure (PDB 3M99) into the density of SAGA models in both open (gray surface) and closed (mesh) conformations.
dUb subunits are the following: Ubp8 (red), Sgf11 (yellow), Sus1 (green), Sgf73 (blue). Both ubiquitin-binding and active sites are shown.

3.2.3. Structure of the SAGA complex by cryo-EM.

3.2.3.1. Sample preparation optimization for cryo-EM.

The negative staining EM analysis of the *P. pastoris* SAGA complex provided indispensable information about its structural organization. However, it is well-known that this procedure has several drawbacks, including possible flattening and distortion of the particles due to dehydration and, of course, the resolution limitations imposed by the nature of the stain. To improve the existing knowledge of SAGA complex, we have determined its structure in near-native conditions using cryo-EM. Previous attempts to resolve SAGA structure were limited by the low yield and heterogeneity of standard TAP-tag purifications and the use of negative staining. The new purification method developed in the lab has improved the homogeneity and stability of endogenous yeast SAGA: for example, we have avoided SLIK complex co-purification, as evidenced by the complete lack of proteolysis in the Spt7 subunit which is prone to the degradation (fig. 59).

In the case of cryo-EM, the presence of sucrose or high salt concentrations in the sample should be avoided due to their detrimental influence on image contrast in cryo conditions. Since the purified sample of SAGA was in 10% sucrose, we have decided to use buffer exchange column to decrease the sucrose amounts to the levels, acceptable for cryo-EM. However, it was noticed that these procedure lead to aggregation and disintegration of SAGA particles (fig. 66). Most likely the fast elimination of sucrose has significantly reduced the sample solubility. Different conditions were screened, including different salts of different concentrations (KOAc, NaCl), detergents (Tween 20, NP-40, DDM) and buffer exchange procedures (columns with different cut-off value, sequential dilutions) in order to have the best preservation of the SAGA complex. Finally, the optimal conditions were found, that gave a good particle distribution on the grid, minimal aggregation and high contrast: cross-linking with 0.1 % of glutaraldehyde in sucrose, followed by either dilution in sucrose-free buffer (supported cryo-EM) or dialysis against sucrose-free buffer (unsupported cryo-EM). It was also noticed that keeping the protein complex constantly at 4°C throughout all sample preparation steps is essential to ensure the best sample quality.

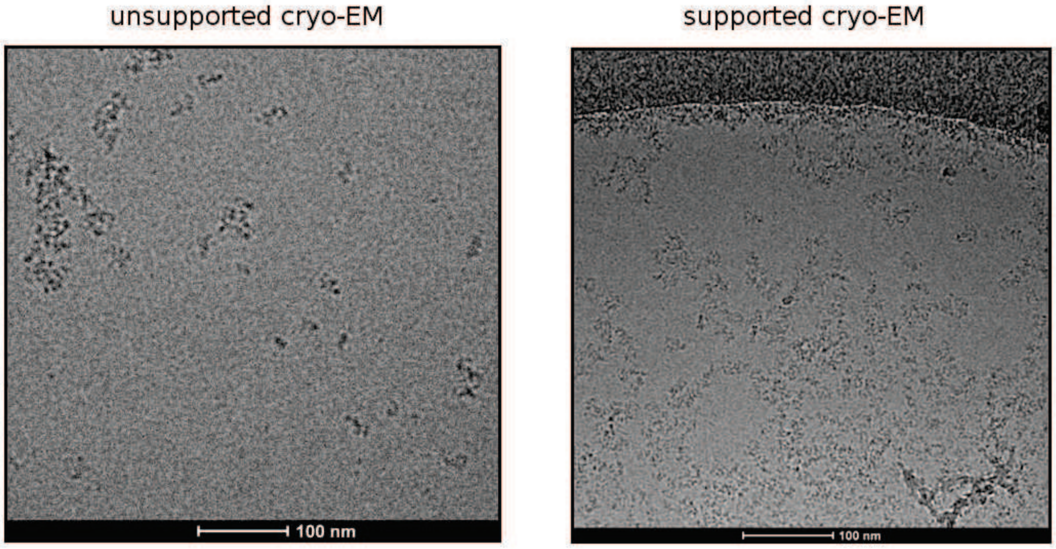

Fig. 66. Direct buffer exchange of non-fixed SAGA leads to the aggregation and disintegration of the protein complex in both unsupported (left panel) and supported cryo-EM (right panel).

3.2.3.2. Single particle cryo-EM image processing of SAGA.

We have set up the conditions for freezing carbon-adsorbed SAGA for supported cryo-EM in order to obtain thin homogenous ice over the grid surface and collected a data set of 2326 micrographs in “movie” mode (see Methods, part 2.3.5) with a total electron fluence of 60 $\bar{e}/Å^2$

(see Table 11). A typical good micrograph of supported, carbon-adsorbed SAGA is shown in fig. 67.

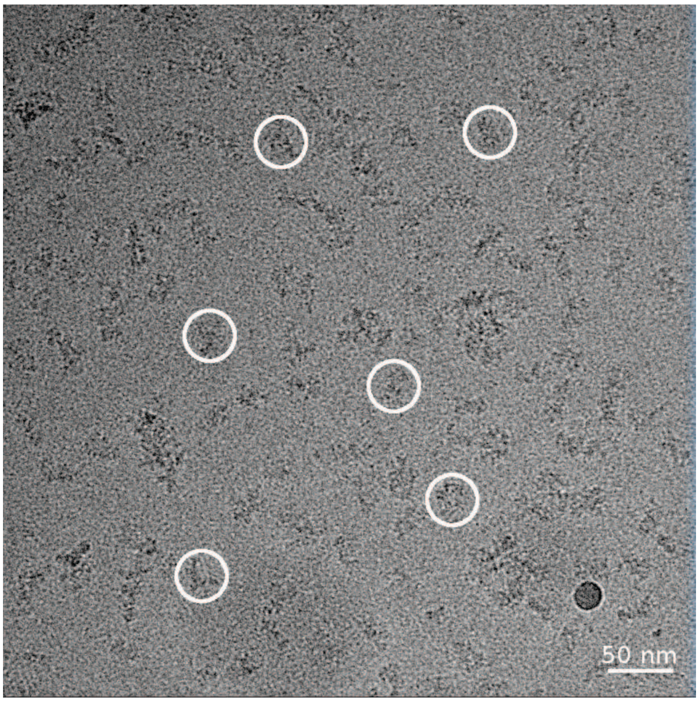

***Fig. 67. A typical supported cryo-EM micrograph with several SAGA complex particles selected*.**

After CTF estimation, all images were visually screened and micrographs with high astigmatism, drift and contamination were discarded. The best images had visible Thon rings up to 4 Å resolution. High electron fluence facilitated single particle picking, that was performed manually on a small subset of micrographs with different defocus values, hence different contrast level. Preliminarily, these micrographs were coarsened 4 times and low-pass filtered to 30-40 Å to improve the contrast before picking. About 2800 particles were selected and subjected to 2D reference-free classification and alignment, producing a few low-resolution (due to the small number of particles) class averages. Several 2D averages were selected (fig. 68) as references for automated particle picking on the whole data set (Hoang et al., 2013). We have also exploited the possibility of using projections of the 3D SAGA model obtained from negatively stained particles (Durand et al., 2014) as references for particle selection.

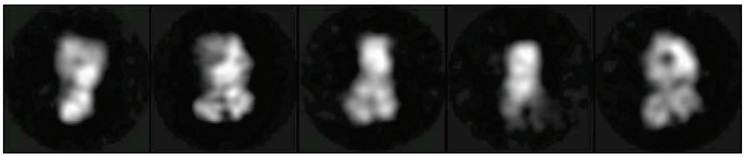

Fig. 68. 2D class averages from a small manually selected data set, used for automatic reference-based particle picking.

Box size is 45 nm.

After sorting, we conserved about 80.000 particles, all of them were subjected to several rounds of iterative 2D classification with removal of bad particles, that yielded 40.000 particle data set, producing high-resolution 2D classes (fig. 69). 2D class averages revealed the known two-lobe shape of the SAGA complex, displaying a preferential orientation on carbon, as in the case of negative stain analysis. Intriguingly, the connection between the two lobes seemed much more stable than before: only the area corresponding to domain V and probably the tip of domain III appeared fuzzy, indicating their inherent flexibility.

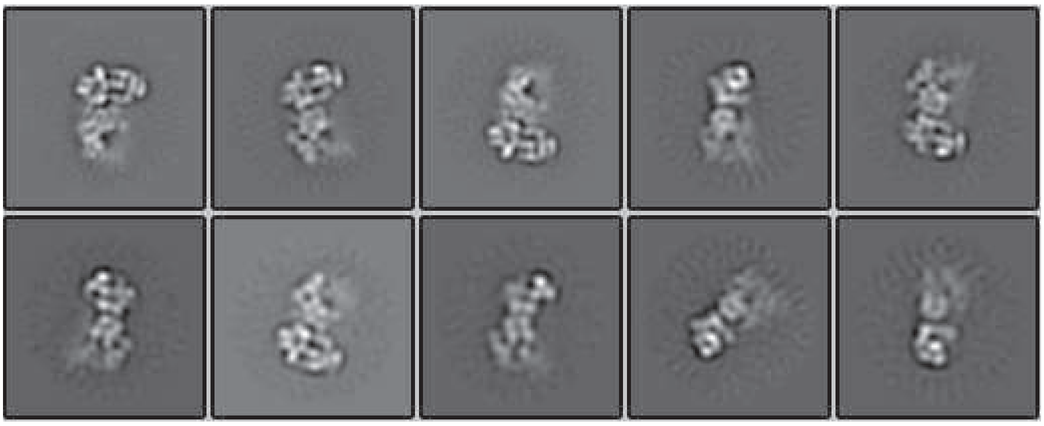

Fig. 69. Representative 2D class averages from ~ 40.000 particles displaying mostly preferential orientation.

Box size is 55 nm.

We have performed a 3D refinement using the known SAGA structure (Durand et al., 2014) and obtained an 18 Å 3D model of the full SAGA complex. When we compared the resulting 3D model to the negative stain structures, it seems that the cryo-EM map mostly resembles the open conformation of SAGA (fig. 70). Though this model encompasses about 40.000 particles, many of them adopt a preferential orientation (fig. 69), severely limiting the final resolution due to overrepresented views. The subsequent 3D classification of the full complex (fig. 71) has revealed a presence of the closed SAGA conformation but which represents only 15% of the full data set. The separation of the open and closed conformations was complicated due to the flexibility of lobe B which thus produced a large number of intermediate states. Therefore, we could not analyze the separate conformations further due to insufficient number of particles in the classes.

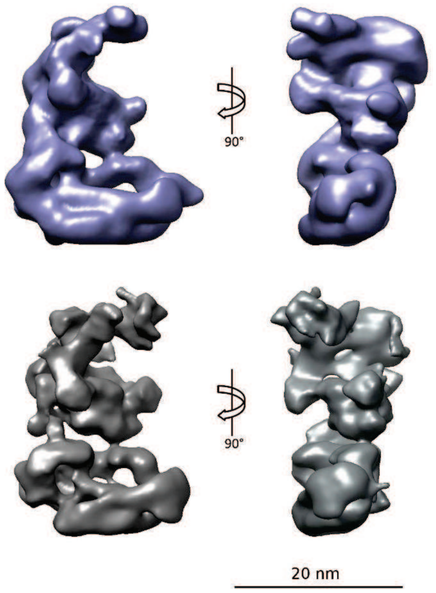

Fig. 70. Comparison of 3D SAGA models obtained by negative staining (blue) and cryo-EM (gray).

Both models show an open conformation of the lobe B.

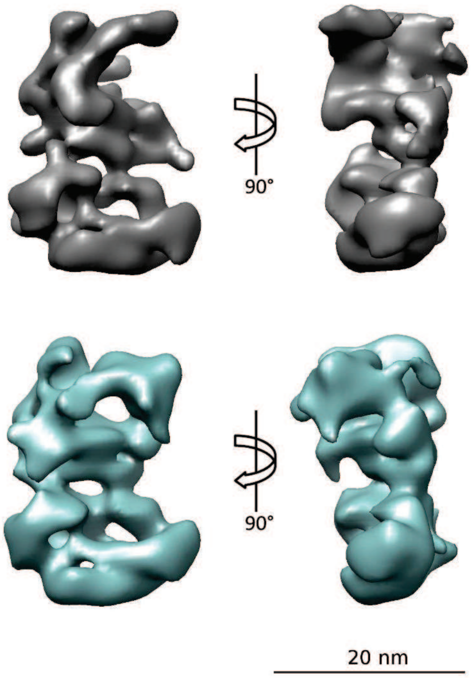

Fig. 71. The two most distinct 3D classes from the supported cryo-EM data set: open (gray) and closed (green) conformations of the lobe B.

Domain V of the lobe B adopts different positions, facilitating the closure of the molecular clamp.

To further improve the resolution we decided to analyze each lobe independently taking advantage of the fact that the particles adsorbed with their long axis parallel to the carbon film thus creating a tomographic series with little overlap of the two lobes. We have proceeded with a separate refinement by applying a soft mask on each lobe of the SAGA complex, taking into account the fact that the lobe B is more flexible than lobe A. We were able to improve the average resolution of lobe A to 11 Å (fig. 72a), as shown on FSC graph (fig. 72b). The local resolution of the model varies from 9 Å in the core of the lobe to 14 Å at the connection between the two lobes. We conclude that this level of resolution is probably the maximum achievable with our data set, considering the severely preferred orientations. Nevertheless, we have started "movie" processing of the lobe A (see details in Methods, part 2.3.6.10), hoping to further improve its structure. This step of refinement is extremely computationally expensive and is still in progress.

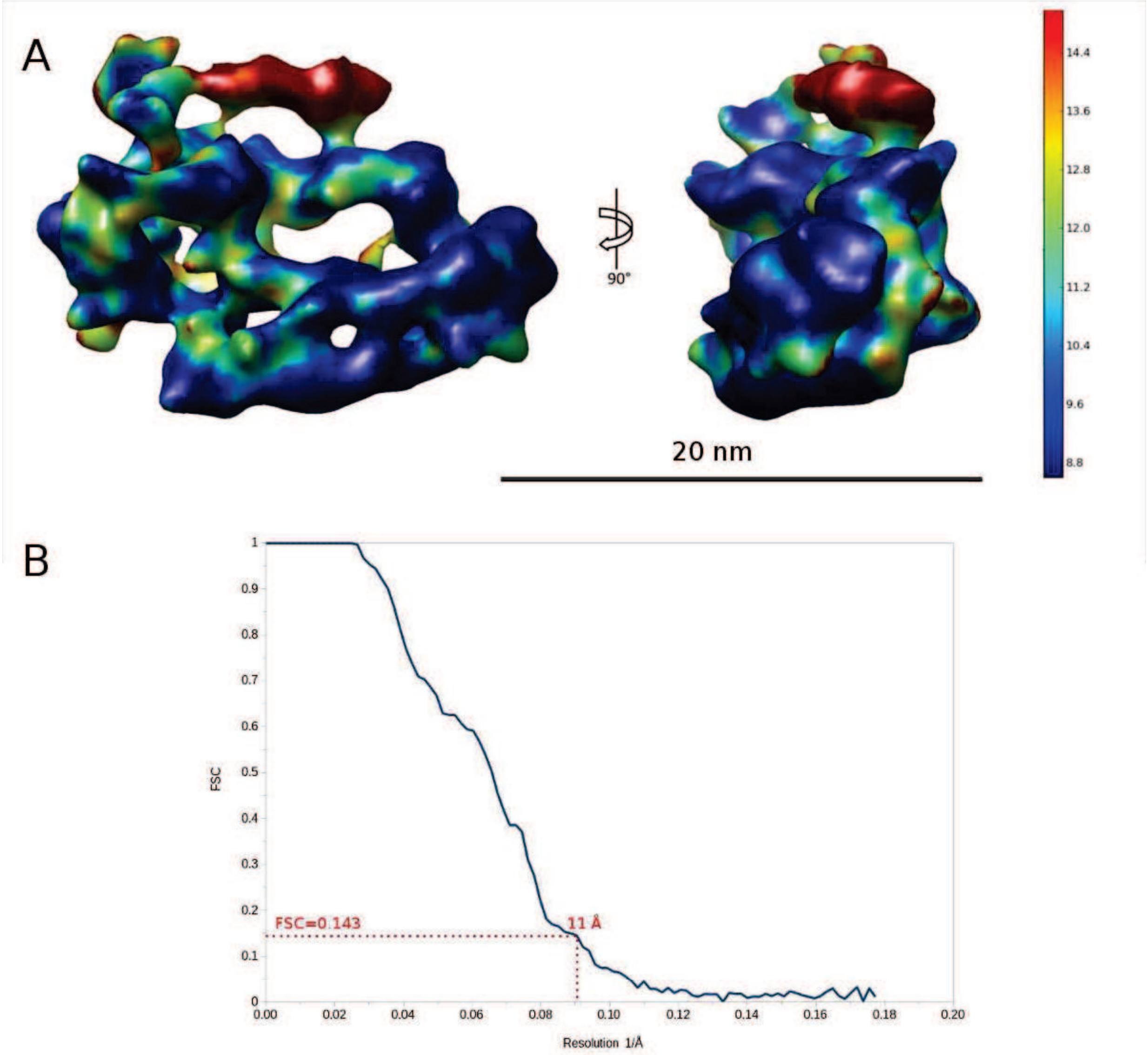

Fig. 72. Lobe A of SAGA complex, refined separately with a mask, has displayed an improved resolution.

(a) The 3D structure of the lobe A colored by the local resolution. (b) The corresponding FSC graph, showing the average resolution of 11 Å.

We have repeated the same procedure for the lobe B. However, its resolution did not improve in comparison with the refinement of the lobe A. This can be explained by the high conformational heterogeneity, that cause averaging of different conformations in the total structure. Subsequent 3D classification of the lobe B with a mask could not separate these conformations. Thus, we have decided to repeat the protocol we used for the HAT data set: we took the alignment parameters from the refinement of the lobe A and classified the whole complex without masking or alignment. We have identified different positions of the lobe B in respect to the fixed lobe A (fig. 73). Lobe B flexibility arises from a multidirectional movement "around" its connection with lobe A, similar to a swirling movement. Up to 15° movement of the lobe B with respect to the lobe A can be detected (corresponding to 5-7 nm). Such a huge flexibility together with the presence of open and closed conformations of the lobe B represent a key obstacle for attaining the higher resolution of the SAGA structure.

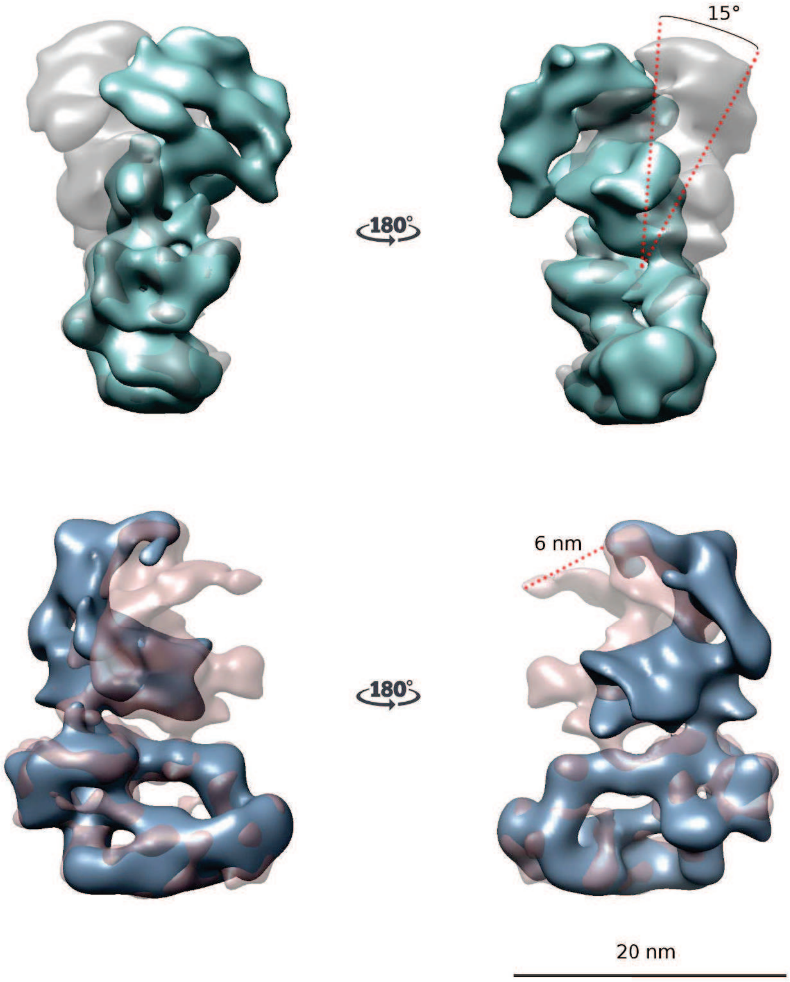

Fig. 73. Overlay between the representative 3D classes, showing the lobe B complex movement.

Upper panel – tilting in 15° range with respect to the lobe A, lower panel – orthogonal view, showing the movement in 5-7 nm range, that changes the connection between the lobes.

The image processing of carbon-adsorbed SAGA complex demonstrated that the complex adopts a preferential orientation on the carbon surface, leading to overrepresentation of particular side views and limiting attainable resolution (see later fig. 77). We have initiated the study of

SAGA by unsupported cryo-EM in order to overcome these problems by embedding single particles of the complex in a thin layer of vitrified ice. First experiments have shown (fig. 74) that the SAGA complex is very fragile even after cross-linking and may undergo disintegration or partial denaturation at the air-water interface, especially if the ice is too thin. Series of sample preparation experiments have been performed using different detergent concentrations and different blotting conditions, that affect ice thickness, number of particles present in the ice layer and their distribution.

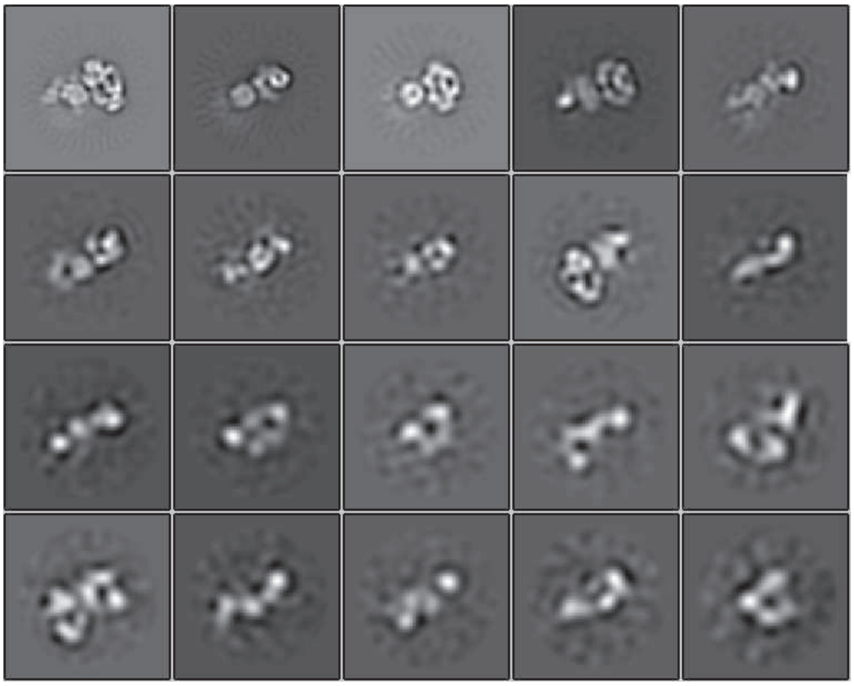

Fig. 74. Example of 2D class averages from unsupported data set, collected from the areas with very thin ice.

Many SAGA particles display smaller size, indicating complex dissociation. Box size is 56 nm.

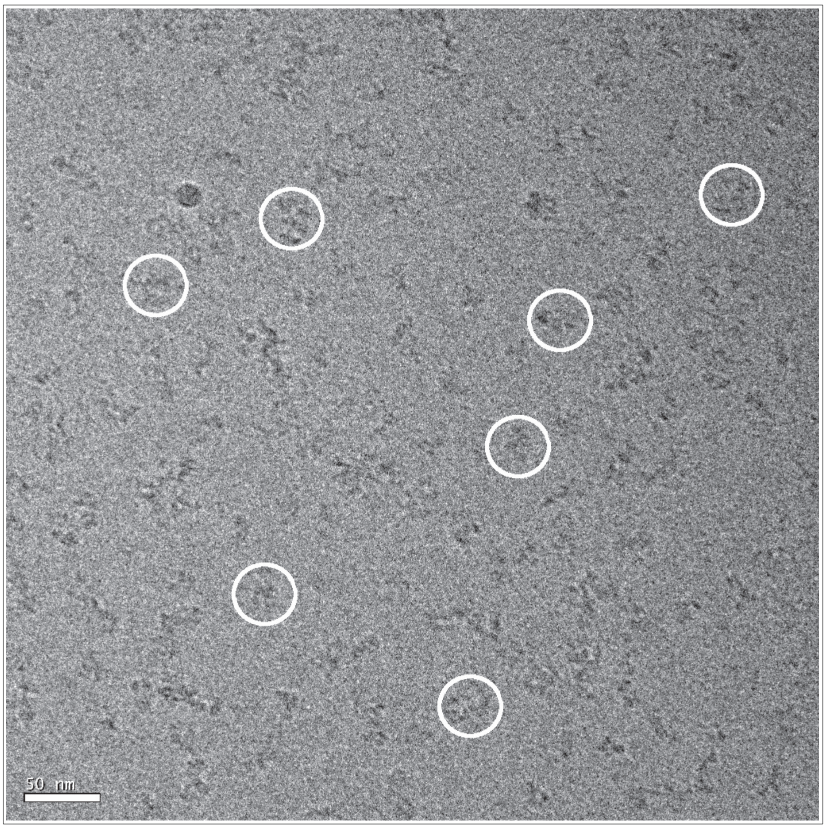

Fig. 75. An example unsupported cryo-EM micrograph with optimal ice thickness.
Several SAGA particles are marked by circles. Scale bar is 50 nm.

So far, we have collected a large data set of 5840 movies (Table 11) recorded in areas where the ice thickness was optimal: not too thin to avoid complex disintegration and not too thick to impede the contrast (fig. 75). We have automatically picked 528.945 particles using projections of SAGA complex from available 3D map. Usually, one should be very careful using such an approach, since this method imposes a severe reference bias during the picking procedure. However, when the particles are clearly visible in the micrographs, this technique allows to detect all possible views of the protein complex, not only the preferential ones. Subsequently, all bad particles containing correlated noise, ice or contaminants can be sorted out by iterative classification. Next, we have decided to proceed directly with a 3D classification of all picked particles. This approach was tested in our lab and has shown to be computationally faster than performing several rounds of 2D classification for the elimination of bad particles. Normally, one starts an initial 3D classification with very coarse angular step (15 or 30 degrees) and global search for particle alignment parameters. Afterward, the angular step can be gradually increased, continuing with local searches over a restricted angular range. This approach has allowed to efficiently remove most of the bad particles from the data set, since many of them have merged into one or two "bad" 3D classes. After removing these unrelated particles we have performed a 3D refinement of the full complex with 435.442 images. The first unsupported cryo-EM 3D structure looks similar to the one of carbon-adsorbed SAGA, with lobe B in open conformation and a pronounced domain V (fig. 76). In addition, the EM map shows less severe preferred orientations in comparison with carbon-adsorbed cryo-EM map of SAGA (fig. 77). We are currently continuing to process this data set.

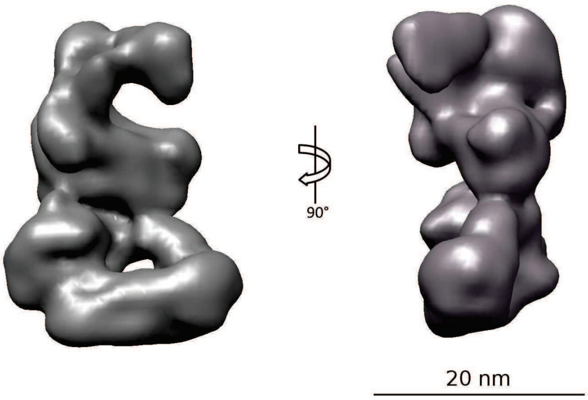

Fig. 76. One of the first SAGA 3D structures from the large unsupported cryo-EM data set.
The domain V appears strongly pronounced here.

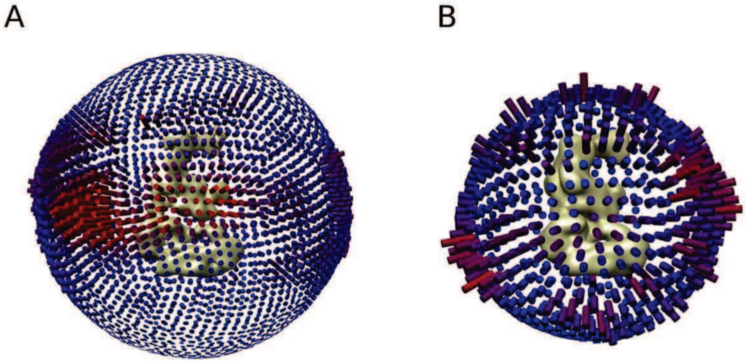

Fig. 77. Representation of an Euler sphere, showing the most abundant views in the 3D structures.
Each cylinder is related to the particular view, the height and the color of the cylinder are proportional to the number of particles contributing to a certain view. (a) – supported cryo-EM SAGA, (b) – unsupported cryo-EM SAGA. Supported carbon-adsorbed SAGA displays severe preferential orientation (one large red area), while the unsupported SAGA shows more even distribution of different orientations.

3.3. Interaction of *P. pastoris* SAGA complex with TBP.

To assess TBP binding, the SBP-tag purified *P. pastoris* SAGA complex was incubated with 5 molar excess of TBP for several hours, then the reconstituted complex was cross-linked and analyzed by negative staining EM. More than 100.000 particle images were recorded and analyzed by the means of 2D classification and alignment. The final class averages have shown the characteristic SAGA views as have been observed previously (fig. 78). Already at this level, the open and closed conformations of the molecular clamp formed by the lobe B could be identified.

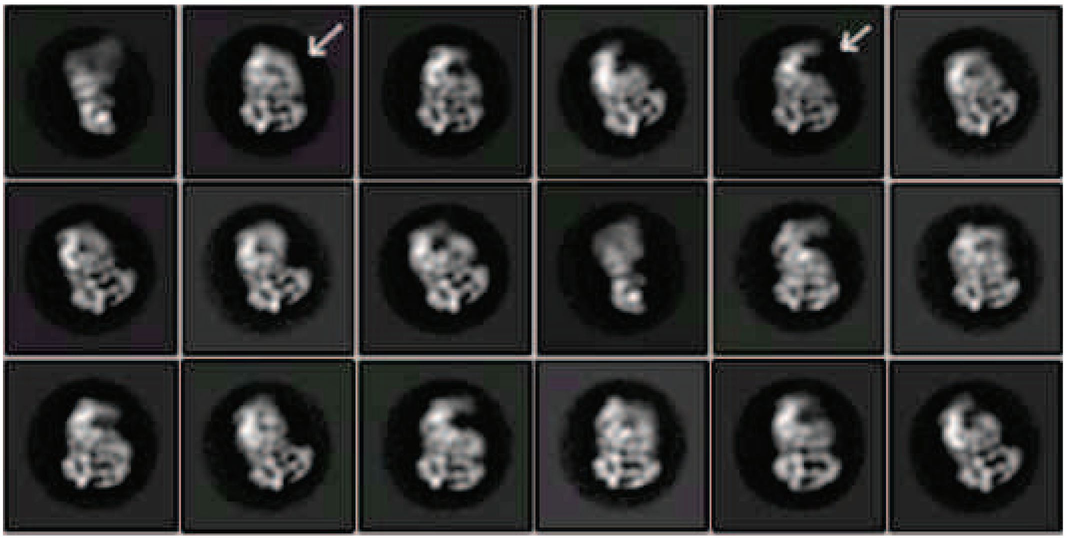

Fig. 78. 2D class averages of the SAGA complex with TBP.
Both open and closed conformations (marked by arrows) of the lobe B can be clearly distinguished. Box size is 50 nm.

After 3D reconstruction and 3D classification of the reconstituted complex we observed a similar set of conformations, as in case of SAGA alone: approx. 44% of particles clearly displayed a closed conformation of the lobe B, while 30% adopted an open conformation. Since the ratio of particles with closed versus open lobe B did not change significantly upon addition of TBP (in comparison with the negatively stain SAGA data set), we conclude that TBP does not stabilize a particular conformation of the molecular clamp. We have proceeded with a 3D refinement of the open and closed SAGA states, as for the previous data sets. The final models of the closed and open SAGA conformations have reached 25 Å resolution (fig. 79). Both models display excellent structural overlay with an exception of the domains III and V: the tails of both domains are more extended and brought together to form the closed conformation.

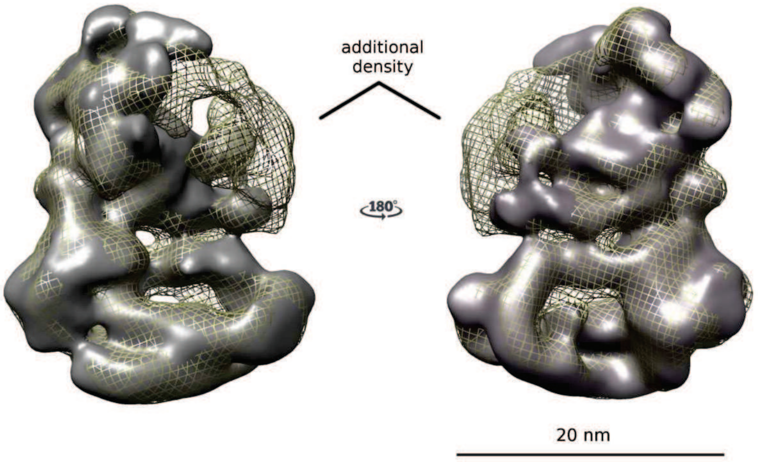

Fig. 79. Overlay between the open and closed conformations of the SAGA complex with TBP. *Opening and closing of the molecular clamp is mediated by the extensions of domains III and V.*

To locate the position of TBP we have proceeded as follows. First, we have addressed whether TBP can bind SAGA in the open conformation. After comparison of the 3D model in open conformation with TBP with previously obtained negative stain SAGA model we have unambiguously identified a small additional density that had the size and the characteristic saddle shape of TBP. This additional density was located at the interface between the two lobes, connecting the domains I and III (fig. 80). In this position the DNA binding surface of TBP is oriented towards the SAGA volume and is not exposed to the solvent. The position of TBP is consistent with the previously located Spt8 subunit, known to interact with TBP (Durand et al., 2014). In addition, we have performed a 3D classification of SAGA models with TBP, that displayed clearly open conformation, and has detected a single 3D class (10% of the data set), that was missing the identified TBP density (data not shown). This observation suggests that the data set contains two subpopulation of SAGA molecules: one which binds TBP and the other which does not.

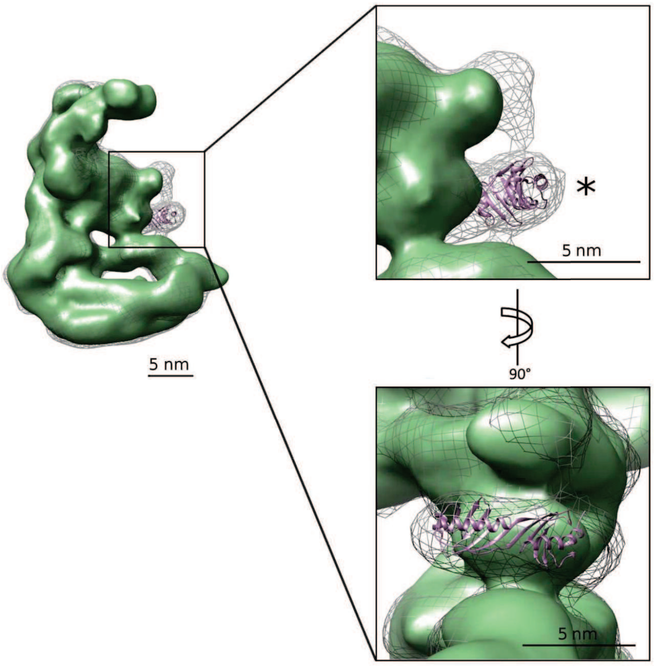

Fig. 80. Overlay of SAGA 3D models in open conformation without (green) and with (gray mesh) TBP.

Asterisk (*) marks the TBP position. TBP coordinates were taken from PDB 1YTF.

In the next step, we focused on the closed conformation and went through 3D classification of the particles that displayed a clearly closed conformation of the lobe B. Among the different 3D classes we have found one (10% of the total data set), that looked almost identical to the closed conformation of full SAGA without TBP (data not shown). This observation indicates that, as for the open conformation, a fraction of the closed SAGA complexes did not bind TBP. In addition, in the SAGA-TBP data set we were able to find a 3D class (20% of the total data set) in which an additional density was found near domain III besides the one, that we have assigned previously to TBP (fig. 81). Intriguingly, this additional density is much larger (~100 kDa) than the expected size of the TBP molecule. We speculate that domain V may fold back onto domain III and fill this density.

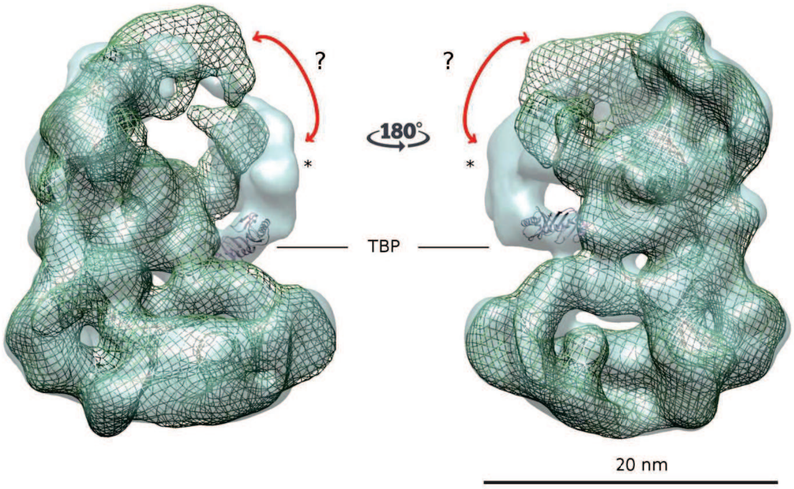

Fig. 81. Among the different 3D classes derived from the SAGA-TBP data set, part of the SAGA complex did not bind TBP.

Both models show a closed lobe B conformation, however the additional density, that does not correspond to the TBP, appears after direct comparison of the aligned 3D structures. We speculate that the domain V folds back onto the domain III (red arrows).

In conclusion, we show that both open and closed SAGA conformations can bind TBP and suggest that the SAGA in closed conformation undergoes a large conformational change upon TBP binding.

Chapter 4. Discussion.

The large number of subunits and elaborate composition of the SAGA complex impose many technical challenges to describe SAGA's molecular organization in the context of its functions in transcription regulation. The first structural description of SAGA complex (Wu et al., 2004) has provided initial insights into the intricate SAGA architecture. In that work, several subunits of the complex were immunolabeled and analyzed by single particle TEM. Lee et al. (Lee et al., 2011) used combinatorial depletion analysis and mass spectrometry to build a subunit interaction network, confirming SAGA modular organization. Han et al. (Han et al., 2014) applied CX/MS approach to characterize subunit arrangement and interaction of SAGA with TBP. Recently, Setiaputra et al. (Setiaputra et al., 2015) used immunolabeling, negative stain EM and CX/MS to map the subunit interconnectivity. Several discrepancies were found between the different proposed SAGA models, including positions of key subunits such as Spt3, Spt20 and the subunits of the HAT module.

4.1. Sample preparation has a significant impact on the quality of SAGA.

In this work we have determined the first cryo-EM structure of the SAGA complex at an intermediate resolution of 11-18 Å. Cryo-EM study of this complex was made possible by using high-quality purification protocol developed by A. Ben-Shem in the lab, that has allowed to produce weekly up to 70 ug of the homogenous and stable complex, giving a tremendous advantage over other groups in the field that use standard TAP-tag purification. All the previous EM studies of SAGA structure (Wu et al., 2004; Durand et al., 2014; Setiaputra et al., 2015) were performed using a negative staining technique on relatively small data sets. Large amount of the purified complex has allowed to find the optimal conditions for sample preparation for cryo-EM study. Importance of the sample quality for high-resolution cryo-electron microscopy is often underestimated. One cannot simply hope that bad sample will produce a correct high-quality structure. Therefore, good biochemistry still remains a key to successful structure determination by any structural biology method.

The new purification method has yielded SAGA devoid of SLIK complex, as evidenced by the complete lack of proteolysis in the Spt7 subunit which is prone to degradation. Avoiding the elution with EGTA during purification led to a better preservation of the dUb module in SAGA, that contains several Zn fingers in dUb subunits. Moreover, the streptavidin-binding peptide was placed onto Sgf73, since it was shown that dUb module may dissociate from SAGA

(Lim et al., 2013; A. Durand, unpublished data). Introduction of the tag onto Sgf73, that is linking dUb module to the rest of the complex (Lee et al., 2011), was a crucial step to ensure the retention of the dUb with the rest of the complex.

In this work we have analyzed large data sets of yeast SAGA complex by both negative staining and cryo-EM. Since the complex displayed high heterogeneity, we have tried to decrease it by several means: constant work at 4 °C or on ice, mild cross-linking by 0.1% glutaraldehyde, using fresh samples without freeze-thaw cycles. All these factors have improved SAGA preservation during sample preparation (homogeneity, particle distribution) and ensured good quality of the final EM maps.

4.2. Conformational flexibility of SAGA complex.

Our new purification protocol has allowed us to improve SAGA complex stability in comparison with the previous studies and more specifically address SAGA heterogeneity. Already at the stage of 2D analysis of our data sets we have noticed more pronounced appearance of the interlobe connection of the SAGA particles (see fig. 69). The lobe A of the complex, known to contain Tra1 and several TAFs, was found very stable and this has allowed us to improve its resolution up to 11 Å. The lobe B of the SAGA complex is composed of three domains: III, IV and V and it is thought to contain the dUb and the HAT modules, Spt subunits and the other TAF subunits. This lobe, on the contrary, was shown to be highly flexible and could adopt different conformations. We have discovered a complex rotation of the whole lobe B in respect to the lobe A, that changes the connection between the domains II-IV, and I-III. We describe this rotation as following: if the lobe A is fixed in space, lobe B can tilt in different directions. Extreme cases of such rotation were discovered during analysis of the SAGA mutant, in which the HAT module was removed. Apparently, the removal of this module switched SAGA to the open conformation (we suggest that the closed conformation cannot form without the domain V), leading to the extreme flexibility of the lobe B. In addition, the removal of the HAT module led to a large rearrangement of the protein density in the whole lobe, but this was not addressed further at that stage of analysis due to very low resolution of the obtained models.

Besides the rotation of the lobe B, the extended tails of the domains III and V form a molecular clamp, that can adopt different conformations from the open one to completely closed. This molecular clamp represents another source of heterogeneity in SAGA complex, obscuring the high resolution information of the lobe B. It is important to note, that the clamp is not restricted by two extreme conformations: fully open or fully closed, but can rather adopt many intermediate states. Therefore, analysis of the ratio between open and closed SAGA conformations cannot be very accurate. We have estimated that in negative stain 40% of SAGA particles are closed and about 50% are open. However, we did not perform further 3D refinement with a mask on the lobe B to further separate these states. In cryo conditions, we could identify only 15% of the SAGA particles in closed conformation. Several explanations can be proposed: (i) the lobe B appears very noisy in the cryo-EM data set and many more particles are required to improve the resolution and clearly identify the closed state; (ii) during the negative staining the pH is close to 4, which is not the case in cryo conditions – that might affect SAGA flexibility; (iii) in the case of cryo-EM, after cross-linking the sucrose is removed – this can also affect SAGA heterogeneity. We are now analyzing bigger unsupported cryo-EM data set, that already looks promising (in terms of less preferential SAGA orientations). We are convinced that this data set will allow to sort out different SAGA conformations without losing high resolution.

4.3. The HAT module is localized in the most flexible part of *S. cerevisiae* SAGA.

The first structural study of SAGA (Wu et al., 2004) has positioned the key subunit of the HAT module, Gcn5, into domain III. The position of the HAT module was recently challenged by a research group which localized the HAT subunits Gcn5, Ada2 and Sgf29 in domain V using GFP-tagged proteins to introduce a detectable density (Setiaputra et al., 2015). To obtain a more comprehensive location of the HAT module we have used a specific SAGA mutant, where the Ada2 subunit was deleted. This particular deletion was shown to dissociate the whole HAT module from SAGA, since the Ada2 is thought to be a linker subunit between the HAT and the rest of the SAGA complex (Lee et al., 2011). We have analyzed the TAP-tag purified *S. cerevisiae* SAGA mutant by negative stain and single particle analysis and discovered that the tail of the domain V is shortened in the mutant in comparison with the WT SAGA. At the same time, we have noticed that the domain III density was also slightly reduced. However, the large density in domain III that contains the DUB module was fully detected only when lobe B was analyzed separately (Durand et al., 2014), which was not possible for the Ada2 mutant. Therefore the deletion mutant strongly suggests that the HAT module is placed at the end of domain V, consistently with the GFP-labelling. More recently, we have obtained a mutant strain of SAGA that has only Gcn5 deleted. Such mutation was shown to affect only Gcn5 subunit leaving the rest of the SAGA complex intact. We plan to analyze this mutant by negative stain EM and localize the missing density corresponding to Gcn5. However, this might be complicated due to both small size of the Gcn5 protein (50 kDa) and high flexibility of the lobe B.

The previous location of Gcn5 in domain III using subunit-specific antibodies (Wu et al., 2004) was very convincing and our interpretation for this discrepancy is the following. In the closed conformation domain V contacts domain III and forms a closed protein ring. In this particular conformation and with the poor resolution of the previous map, we might have selected antibody-labeled complexes in the closed conformation and missed the more heterogeneous open complexes.

The localization of the HAT in the domain V, in the most flexible part of the complex, provides a tantalizing explanation of SAGA's ability to act on different chromatin substrates. The observed flexibility may have a functional role to adapt to different substrates. Multiple chromatin-binding domains have been identified within HAT subunits. The ability of SAGA to bind methylated and acetylated histones suggests a large plasticity in substrate recognition. The position of HAT and dUb modules at the tails of the corresponding SAGA domains, forming a molecular clamp, also reveals new possibilities of their functional interaction. Cross-linking results (Han et al., 2014) have positioned the two modules in proximity and demonstrated that C-terminus of Sgf73 can form numerous cross-links with Ada2 and Ada3. The same authors have analyzed the abundance of the HAT and dUb subunits in SAGA mutants, missing the dUb or the HAT module, respectively. They have concluded that the dUb module plays an important role to stabilize the interaction of the HAT module with SAGA. In the work of A. Durand, the dUb module was localized in the extended tail protruding from the domain III, suggesting that HAT subunits might interact with it only in closed conformation of the molecular clamp. However, more precise subunit localization is needed to further address this point.

4.3.1. SAGA HAT molecular organization, analyzed by different methods*.

* For additional discussion of the results, obtained by cross-linking/MS in combination with homology modeling, the reader can address to the "Discussion" part of the paper, presented at the page 131 of this work.

In this work, we have used several methods to study the HAT module of SAGA. First, we have analyzed the subunit interactions of the recombinant human HAT *in vitro* by pull-down different subunit combinations. Next, we have determined the stoichiometry and topology of the recombinant SAGA HAT subcomplex by cross-linking coupled to MS and subsequent homology modeling. We have also used negative staining EM and tried to obtain the information about the general shape of the subcomplex. Finally, we have studied the *S. cerevisiae* SAGA mutant, in which the HAT module was deleted in order to locate its position within the complete yeast SAGA complex.

Altogether, these experiments have provided a first information on the topological organization of the HAT module. Cross-linking and pull-down experiments have shown that SGF29 protein is most likely located at the periphery of the subcomplex, and linked to it only by ADA3, consistent with the previous studies (Table 14). ADA2 and ADA3 proteins form the core of the HAT module, that connects it to the rest of SAGA complex. The cross-linking pattern between these two subunits is conserved between yeast and human, even when we consider an isolated HAT module separately from the rest of the SAGA. This is consistent with the hypotheses of Ada2, forming the bridge between Ada3 and Gcn5 (Horiuchi et al., 1995; Candau et al., 1996; Candau and Berger, 1996), and Ada3 / Ada2, connecting the HAT to the rest of the SAGA complex (Lee et al., 2011; Han et al., 2014). However, we should indicate, that successive cross-linking does not necessarily mean interaction and only point to the proximity of the cross-linked residues. Intriguingly, in our pull-down experiments of recombinant human HAT we did not observe any binary interactions between Gcn5, Ada2 and Ada3 until all three of them were co-expressed together, indicating a complex pattern of interaction.

Subunit interaction	**Literature data**	**hHAT pull-down**	**hHAT CX/MS**	**ySAGA CX/MS**
Ada3 – Sgf29	+	+	+	+
Gcn5 – Sgf29	+	-	+	-
Ada2 – Sgf29	-	-	-	-
Ada3 – Gcn5	-	-	+	+
Ada2 – Gcn5	+	-	+	+
Ada2 – Ada3	+	N.D.	+	+

Table 14. Subunit interaction within the SAGA HAT module studied by different approaches.

** Literature data includes deletion mutant analysis and yeast two-hybrid assays. ✝ From Han et al., 2014 and Setiaputra et al., 2015. N.D. - no data was observed.*

In the present work, we have localized the HAT module in the most flexible part of SAGA complex – domain V of the lobe B, that was also confirmed by an independent study of Setiaputra et al. The EM study of SAGA mutant (with HAT removed) was significantly complicated by the flexibility of the lobe B of the SAGA, especially of the domain V. Our separate EM study of the purified HAT has revealed a high heterogeneity of the subcomplex: the size of the particles varied from very compact (9-10 nm in diameter) to elongated (12-14 nm). We assume that such variability has prohibited us to obtain a HAT structure at 20-25 Å resolution. In addition, SAGA HAT is a relatively small subcomplex (240 kDa) that makes it difficult to study by EM methods. At the same time, we cannot exclude that the purification or sample preparation of SAGA HAT led to partial dissociation or aggregation of the subcomplex. Another possible explanation for the high heterogeneity is the absence of interaction partners. While the group of J. Workman (Lee et al., 2011) has proposed the existence of a separate HAT module based on their deletion analysis, we cannot exclude the necessity of other SAGA proteins for the formation of a stable HAT module. In addition to that, the overall cross-linking map of the recombinant human HAT in comparison with the yeast HAT module within the SAGA complex was particularly different for the Ada3 subunit: in the whole complex yAda3 was shown to cross-link other SAGA proteins, while in absence of these proteins mADA3 displayed more cross-links with other HAT subunits.

Another interesting comparison between the cross-linking maps of hHAT alone vs yHAT within SAGA (besides different protein sequence in yeast and human) is the overall interaction network (fig. 82). In the isolated HAT we can see that all subunits (except ADA2B – SGF29) are tightly cross-linked to each other, which is also visible in our homology model of the HAT: all proteins are positioned rather close to each other, revealing a compact form of the HAT (fig. 82b). In contrast, the HAT module within the SAGA displays a somewhat linear cross-linking network (fig. 82a): Gcn5 – Ada2/Ada3 – Sgf29, where Gcn5 and Sgf29 are located at the two opposite sides and do not cross-link. This might reflect the possibility that the HAT adopts two different conformations: a compact one and an elongated one upon integration into the SAGA. Both conformations were observed in our EM analysis of HAT alone, however, they could not be not clearly separated probably because of a continuous conformational change. Furthermore, the shape of the domain V in our negative stain and cryo-EM maps is clearly elongated, as revealed in particular in the closed conformation. All HAT subunits, except for Ada3, contain several chromatin-binding domains (bromodomain in Gcn5; ZZ, SANT and SWIRM in Ada2; TUDOR in Sgf29). The possibility of an extended conformational space further extends the adaptation faculty of the SAGA HAT to sense chromatin modifications and fine tune variable distances between histone tails.

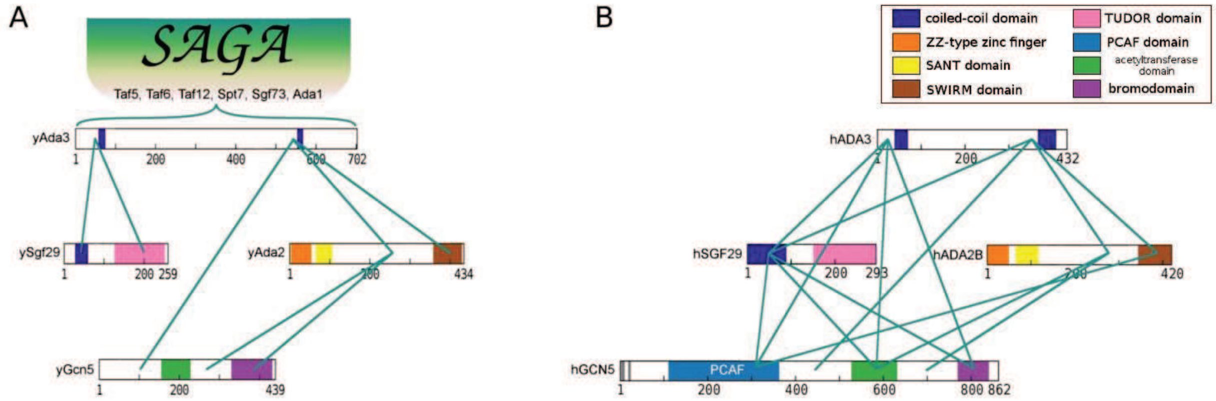

Fig. 82. Simplified cross-linking maps of the SAGA HAT module from yeast (a) and human (b).

Panel (a) is based on the data from Han et al., 2014.

The functional characterization of the hHAT module alone and incorporated into hSAGA has been performed by A. Riss (L. Tora group, unpublished results). The purified recombinant GCN5 alone, the HAT module or complete endogenous SAGA complex were tested for their ability to acetylate different substrates *in vitro* (Table 15). The enzymatic activity of GCN5 was significantly increased upon integration into the HAT module, however, it was not sufficient to acetylate nucleosomes *in vitro*. These results are consistent with a study of J. Workman group: they have also demonstrated an increase in acetylation activity (on H3 peptides) of recombinant yeast Gcn5 upon its integration into yeast SAGA complex (Eberharter et al., 1998). However, both studies have noticed a decrease in acetylation activity (in comparison with the peptides), when the endogenous SAGA (both yeast and human) was tested on the nucleosomes. These differences in activity may reflect the dynamic conformational changes of the HAT module in different environments.

	H3 (H4) N-terminal peptides	**histone octamers**
recombinant hGCN5	50 (20)	2000
recombinant hHAT	900 (55)	42.000
endogenous hSAGA	1500 (180)	73.000

Table 15. GCN5 acetylation activity on different substrates depends on the protein environment.
Numbers are given in arbitrary counts.

4.4. Location of the TAF subunits.

TAF subunits are key structural subunits of SAGA that are shared with TFIID. They are known to to be important for TFIID assembly and were also proposed to comprise a structural core of the SAGA complex.

In the present work we have obtained the first cryo-EM map of the *P. pastoris* SAGA complex that has confirmed a two-lobe architecture of the complex (Wu et al., 2004; Durand et al., 2014; Setiaputra et al., 2015). Lobe A is known to contain mostly Tra1 in the domain I, the largest protein in SAGA complex, and several TAF subunits in the domain II. This lobe did not display significant heterogeneity during image analysis and thus was refined to a higher resolution. Although the 11 Å structure of the lobe A does not allow to trace the alpha-helices in order to be able to identify individual proteins, we have identified a similarity between the shape of domain II and the structure of 3TAF subcomplex of TFIID (fig. 83). The SAGA complex shares several TAF subunits with TFIID complex: Taf5, Taf6, Taf9, Taf10, Taf12. Two copies of the first three proteins (Taf5, Taf6, Taf9) form a 3TAF subcomplex, representing the core of TFIID. SAGA is also thought to contain two copies of each of these proteins, however this stoichiometry was not confirmed experimentally yet. We speculate that one half of the symmetric 3TAF might be present in SAGA in a somewhat similar configuration. The second half of 3TAF is not visible in either lobes of SAGA. This might be due to the flexibility of the connection between the domains II and IV, completely different orientations of the second pair of TAF subunits in SAGA, or simply their absence. The proposed position of the TAF subunits is consistent with immunolabeling experiments, performed by two independent groups (Wu et al., 2004; Setiaputra et al., 2015. In both studies, numerous TAF subunits as well as their proposed interaction partners (Taf5, Taf6, Taf9, Spt7/Taf10 and Ada1/Taf12) we localized in the interface between the two lobes.

TAF subunits are known to play a structural role, providing integrity of the SAGA complex. Thus, the TFIID-like core positioned in the center of SAGA between domains II and IV seems very probable. Nevertheless, more investigation is required to address this question.

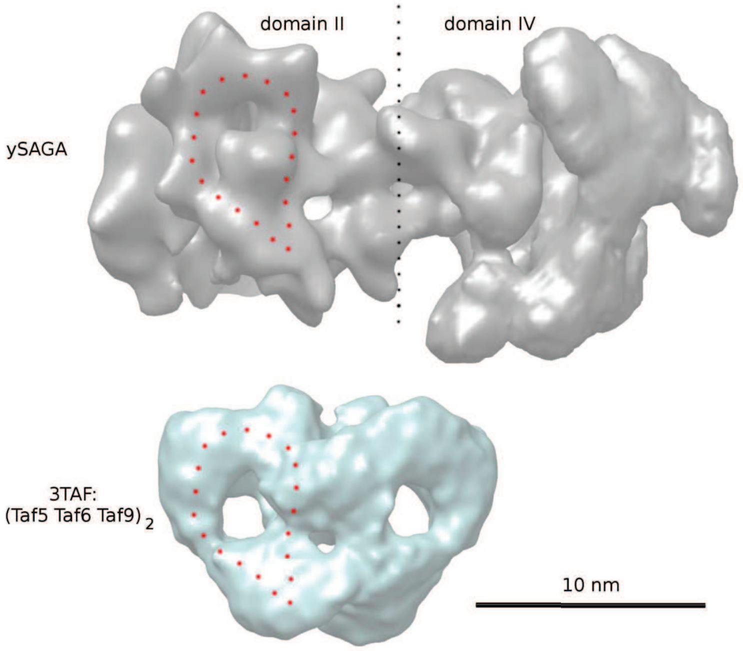

Fig. 83. Comparison between SAGA complex (gray) and 3TAF (green) subcomplex structures. *The domain II of the lobe A slightly resembles one of the halves of the 3TAF structure (EMD-2229).*

4.5. TBP binds *P. pastoris* SAGA in both open and closed conformations of the molecular clamp.

According to numerous biochemical and genetic studies, TBP is known to bind the Spt3 and Spt8 subunits of SAGA, that participate in the TBP recruitment to the promoters of SAGA-dependent genes. Both subunits play positive and negative roles in transcription regulation and were shown to bind TBP *in vivo* (Mohibullah and Hahn, 2008). Binding sites of two subunits were localized on different surfaces of the TBP. The Spt8 and Spt7 subunit was localized in the domain III of the lobe B of the SAGA by two independent EM immunolabeling studies (Durand et al., 2014; Setiaputra et al., 2015). However, for the positions of Spt3 and Spt20, contradictory results have been reported. Setiaputra et al. have placed both subunits in the domain III, while the former study by Wu et al. placed them in the domain V. We think that the domain V is mostly occupied by the HAT module, since it has a rather big molecular weight (210 kDa in yeast). In this respect, the positioning of both Spt3 and Spt20 subunits in the domain III seems reasonable but needs to be confirmed by higher resolution data. The latest results are also supported by the cross-linking MS study of Han et al., which place Spt3 and Spt20 in close proximity of the dUb module.

In the current work, we have shown that both open and closed conformations of the lobe B of SAGA can bind TBP, suggesting that such binding does not influence functioning of the molecular clamp. We have demonstrated that the TBP binding site is localized in the groove between domains I and III and it is positioned close to the dUb module and to Spt8 location.

In the open conformation the direct binding of TBP is best visualized and allows its precise positioning as well as the orientation of the characteristic saddle shape. The orientation of the N- and C-termini is however not possible at this resolution. Interestingly, when compared to the unbound SAGA state, a small density of SAGA protrudes out and interacts with one end of TBP (see fig. 80). The position of this additional density is consistent with the location of Spt8 but additional high resolution information is required to discriminate between Spt8 and the other known interaction partner of TBP – Spt3. The other end of TBP also seems to be connected to the SAGA density, however it's not clear at this level of the resolution.

In the closed conformation, when domain V comes close to domain III, the lobe B seems to be profoundly reorganized by the binding of TBP. At this stage of the study we do not fully understand the domain movements that occurred in the closed conformation between the TBP-bound and unbound states.

Using atomic coordinates of the TBP-TFIIA-DNA complex, we have built a model of SAGA with TBP and DNA, based on the position of TBP determined in the open state (fig. 84). Intriguingly, the SAGA structure presents a hole between the dUb module and TBP, through which the DNA might pass. However, higher resolution structure is needed to address this hypothesis.

Our work shows that TBP has a central position within SAGA, that the promoter DNA might be trapped by a topologically closed protein domain. Furthermore, in this central position TBP is close to the activator binding Tra1 subunit, located in the domain I of lobe A. Since activators are mainly positioned upstream of the TATA-containing proximal promoter, the lobe B is likely to interact with the downstream part of the promoter. Lobe B contains the histone modifying enzymes as well as the most of the protein domains recognizing histone modifications. This order, dictated by the position of key DNA recognition functions suggest that when bound to TBP and activator, SAGA is capable to sense histone modification and to acetylate/deubiquitinate the +1 nucleosome.

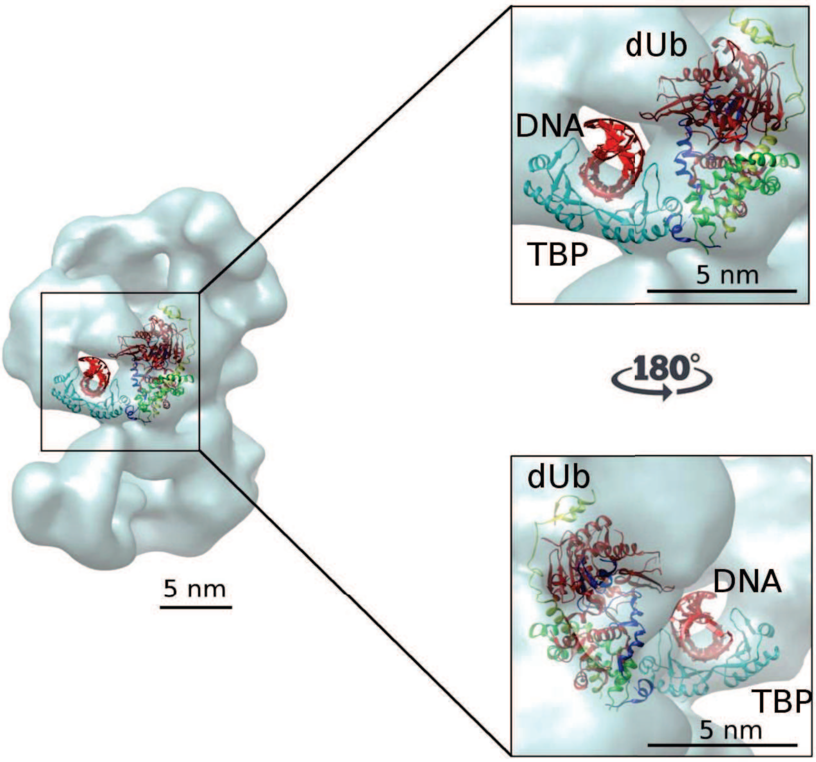

Fig. 84. A proposed model of SAGA interaction with TBP and DNA. DNA protrudes through the cleft formed by TBP and the dUb module.

Crystal structures used: TBP with DNA – 1YTF, the dUb module – 3M99.

Chapter 5. Conclusions and perspectives.

The SAGA coactivator complex represents one of the key players in eukaryotic transcription regulation. Since its discovery, after many years of biochemical, genetic and functional studies the complex was shown to integrate multiple functions provided by its composite modular architecture. However, a lack of structural information about the complex stays one of the major obstacles for full understanding of the mechanisms of its functioning. Several attempts were made to resolve the molecular organization of the SAGA complex by using different methods including chemical cross-linking and mass spectrometry, electron microscopy, deletion mutant analysis. Nevertheless, the lack of high resolution information about SAGA complex persists.

During this work, we have analyzed the organization of endogenous yeast SAGA by electron microscopy and obtained the first intermediate-resolution structure of the full complex by cryo-EM in near-native conditions. A new purification protocol developed in the lab has allowed to significantly improve the quality of purified complex, eliminating most of the contaminants and ensuring preservation of full complement of the subunits. We were able to more precisely analyze and visualize the open and closed conformational states of SAGA, suggesting that such flexibility of the complex allows it to quickly adapt to the interactions with different substrates and cofactors. In comparison with previous studies, we have significantly improved the spatial resolution of the existing SAGA models. Nevertheless, the flexibility remains one of the key problems to achieving high resolution structure of the complex. Large data sets are required to sort out different SAGA conformations with high enough resolution to unambiguously position all 19 subunits of the complex. We are currently processing such a data set of more than half a million single particles.

In the present work, the stoichiometry and subunit interaction network of the HAT module of SAGA was described using a recombinant human homolog of this subcomplex. Based on the cross-linking data and homology models we have determined a first low-resolution model of the HAT module and demonstrated that ADA2 and ADA3 subunits form the core of the HAT, most probably attaching the module to the full SAGA complex. We have compared our results with a cross-linking map that was published during the course of this work and have found that the network of domain interactions is highly conserved between yeast and human. We have also proposed that the HAT module is reorganized upon its integration into SAGA based on significant changes in cross-linking pattern of Ada3 subunit.

In this work, we have localized the HAT module in the most mobile area of the full complex by mutant analysis and demonstrated that its removal induces even larger conformational flexibility of SAGA. High flexibility of the HAT module might represent the ability of SAGA to interact with and modify numerous substrates, such as nucleosomal histones. Localization of the HAT module determined in this work is supported by recent study from other group. In addition, to more precisely localize all components of the module, we plan to analyze new mutants where one or several key subunits of the HAT module were removed.

The dUb module was shown to be very sensitive to the purification, and we were able to develop a new purification protocol, that ensured full retention of the dUb. The dUb module was localized in the full SAGA complex in the previous study by our group, at relatively low resolution. More high resolution analysis is required to precisely determine position of its components, that will allow to address the predicted cross-talk between dUb and HAT modules in SAGA.

We have defined the interaction site of SAGA with TBP within the proximity of dUb module and have observed that both open and closed SAGA conformations can bind TBP. This results further extend the previously proposed hypothesis that TBP might be bound by the molecular clamp itself, since we have discovered a large conformational rearrangement of the lobe B of SAGA upon TBP binding. At this moment, we do not fully understand these domain movements. The resolution of our model of the reconstituted SAGA-TBP complex does not allow an unambiguous identification of the orientation of TBP and thus, higher resolution structures are needed to address this issue. In the light of the discovery of the TBP binding site we plan to determine more precisely the location of the Spt20, Spt8 and Spt3 subunits to be able to understand the process of TBP binding and its consequences on transcription regulation. We have already produced yeast mutant strains where the Spt3 and/or Spt8 subunits were removed. These strains have been examined at the biochemical level and showed that both of them are viable and lose only the mutated subunit, thus giving us the opportunity to position each subunit within the full SAGA structure by comparing their molecular structure with that of the wild-type strain. These strains will also be analyzed for their ability to bind TBP and if so, the binding interface of TBP will be explored to detect a possible conformational variation in TBP interaction.

In conclusion, this work has tried to bring new insights about the molecular organization of the SAGA complex using mostly electron microscopy. Despite our efforts, we were not yet able to obtain high-resolution structure of SAGA, that would allow unambiguously position all its subunits. The main challenge remains high flexibility of the complex, precluding high-resolution analysis. Nevertheless, we are sure that analysis of larger high-quality data sets will allow to discern different conformations of SAGA with high resolution. Much efforts are put into stabilizing the most labile lobe of the SAGA complex, and we are exploring the possibility that nucleosomes or other cofactors could stabilize a single conformation.

Chapter 6. References

1. Allard, S., Utley, R.T., Savard, J., Clarke, A., Grant, P., Brandl, C.J., Pillus, L., Workman, J.L., and Côté, J. (1999). NuA4, an essential transcription adaptor/histone H4 acetyltransferase complex containing Esa1p and the ATM-related cofactor Tra1p. EMBO J. *18*, 5108–5119.

2. Allegretti, M., Mills, D.J., McMullan, G., Kühlbrandt, W., and Vonck, J. (2014). Atomic model of the F420-reducing [NiFe] hydrogenase by electron cryo-microscopy using a direct electron detector. eLife *3*, e01963.

3. Allfrey, V.G., Faulkner, R., and Mirsky, A.E. (1964). ACETYLATION AND METHYLATION OF HISTONES AND THEIR POSSIBLE ROLE IN THE REGULATION OF RNA SYNTHESIS*. Proc. Natl. Acad. Sci. U. S. A. *51*, 786–794.

4. Amerik, A.Y., Li, S.J., and Hochstrasser, M. (2000). Analysis of the deubiquitinating enzymes of the yeast Saccharomyces cerevisiae. Biol. Chem. *381*, 981–992.

5. Ansari, A.Z., Koh, S.S., Zaman, Z., Bongards, C., Lehming, N., Young, R.A., and Ptashne, M. (2002). Transcriptional activating regions target a cyclin-dependent kinase. Proc. Natl. Acad. Sci. U. S. A. *99*, 14706–14709.

6. Atanassov, B., Evrard, Y., Multani, A., Zhang, Z., Tora, L., Devys, D., Chang, S., and Dent, S. (2009). Gcn5 and SAGA regulate shelterin protein turnover and telomere maintenance. Mol. Cell *35*, 352–364.

7. Avvakumov, N., and Côté, J. (2007). The MYST family of histone acetyltransferases and their intimate links to cancer. Oncogene *26*, 5395–5407.

8. Bai, X., Fernandez, I.S., McMullan, G., and Scheres, S.H. (2013). Ribosome structures to near-atomic resolution from thirty thousand cryo-EM particles. eLife *2*.

9. Balasubramanian, R., Pray-Grant, M., Selleck, W., Grant, P., and Tan, S. (2002). Role of the Ada2 and Ada3 transcriptional coactivators in histone acetylation. J. Biol. Chem. *277*, 7989–7995.

10. Bartesaghi, A., Matthies, D., Banerjee, S., Merk, A., and Subramaniam, S. (2014). Structure of β-galactosidase at 3.2-Å resolution obtained by cryo-electron microscopy. Proc. Natl. Acad. Sci. *111*, 11709–11714.

11. Belotserkovskaya, R., Sterner, D.E., Deng, M., Sayre, M.H., Lieberman, P.M., and Berger, S.L. (2000). Inhibition of TATA-Binding Protein Function by SAGA Subunits Spt3 and Spt8 at Gcn4-Activated Promoters. Mol. Cell. Biol. *20*, 634–647.

12. Berger, I., Fitzgerald, D.J., and Richmond, T.J. (2004). Baculovirus expression system for heterologous multiprotein complexes. Nat. Biotechnol. *22*, 1583–1587.

13. Berger, S., Piña, B., Silverman, N., Marcus, G., Agapite, J., Regier, J., Triezenberg, S., and Guarente, L. (1992). Genetic isolation of ADA2: A potential transcriptional adaptor required for function of certain acidic activation domains. Cell *70*, 251–265.

14. Bhaumik, S.R., and Green, M.R. (2001). SAGA is an essential in vivo target of the yeast acidic activator Gal4p. Genes Dev. *15*, 1935–1945.

15. Bhaumik, S.R., and Green, M.R. (2002). Differential Requirement of SAGA Components for Recruitment of TATA-Box-Binding Protein to Promoters In Vivo. Mol. Cell. Biol. *22*, 7365–7371.

16. Bhaumik, S.R., Raha, T., Aiello, D.P., and Green, M.R. (2004). In vivo target of a transcriptional activator revealed by fluorescence resonance energy transfer. Genes Dev. *18*, 333–343.

17. Bian, C., Xu, C., Ruan, J., Lee, K.K., Burke, T.L., Tempel, W., Barsyte, D., Li, J., Wu, M., Zhou, B.O., et al. (2011). Sgf29 binds histone H3K4me2/3 and is required for SAGA complex recruitment and histone H3 acetylation. EMBO J. *30*, 2829–2842.

18. Bieniossek, C., Imasaki, T., Takagi, Y., and Berger, I. (2012). MultiBac: expanding the research toolbox for multiprotein complexes. Trends Biochem. Sci. *37*, 49–57.

19. Birck, C., Poch, O., Romier, C., Ruff, M., Mengus, G., Lavigne, A.-C., Davidson, I., and Moras, D. (1998). Human TAFII28 and TAFII18 Interact through a Histone Fold Encoded by Atypical Evolutionary Conserved Motifs Also Found in the SPT3 Family. Cell *94*, 239–249.

20. Bonnet, J., Wang, C.-Y., Baptista, T., Vincent, S.D., Hsiao, W.-C., Stierle, M., Kao, C.-F., Tora, L., and Devys, D. (2014). The SAGA coactivator complex acts on the whole transcribed genome and is required for RNA polymerase II transcription. Genes Dev. *28*, 1999–2012.

21. Boyer, L., Langer, M., Crowley, K., Tan, S., Denu, J., and Peterson, C. (2002). Essential role for the SANT domain in the functioning of multiple chromatin remodeling enzymes. Mol. Cell *10*, 935–942.

22. Brand, M., Leurent, C., Mallouh, V., Tora, L., and Schultz, P. (1999). Three-dimensional structures of the TAFII-containing complexes TFIID and TFTC. Science *286*, 2151–2153.

23. Brand, M., Moggs, J.G., Oulad-Abdelghani, M., Lejeune, F., Dilworth, F.J., Stevenin, J., Almouzni, G., and Tora, L. (2001). UV-damaged DNA-binding protein in the TFTC complex links DNA damage recognition to nucleosome acetylation. EMBO J. *20*, 3187–3196.

24. Brenner, S., and Horne, R.W. (1959). A negative staining method for high resolution electron microscopy of viruses. Biochim. Biophys. Acta *34*, 103–110.

25. 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.

26. Brown, C.E., Lechner, T., Howe, L., and Workman, J.L. (2000). The many HATs of transcription coactivators. Trends Biochem. Sci. *25*, 15–19.

27. Brown, C.E., Howe, L., Sousa, K., Alley, S.C., Carrozza, M.J., Tan, S., and Workman, J.L. (2001). Recruitment of HAT Complexes by Direct Activator Interactions with the ATM-Related Tra1 Subunit. Science *292*, 2333–2337.

28. Brownell, J., Zhou, J., Ranalli, T., Kobayashi, R., Edmondson, D., Roth, S., and Allis, C. (1996). Tetrahymena histone acetyltransferase A: a homolog to yeast Gcn5p linking histone acetylation to gene activation. Cell *84*, 843–851.

29. Burke, T.W., and Kadonaga, J.T. (1997). The downstream core promoter element, DPE, is conserved fromDrosophila to humans and is recognized by TAFII60 of Drosophila. Genes Dev. *11*, 3020–3031.

30. Burke, T.L., Miller, J.L., and Grant, P.A. (2013). Direct inhibition of Gcn5 protein catalytic activity by polyglutamine-expanded ataxin-7. J. Biol. Chem. *288*, 34266–34275.

31. Calonge, T.M., Eshaghi, M., Liu, J., Ronai, Z., and O'Connell, M.J. (2010). Transformation/Transcription Domain-Associated Protein (TRRAP)-Mediated Regulation of Wee1. Genetics *185*, 81–93.

32. Candau, R., and Berger, S.L. (1996). Structural and functional analysis of yeast putative adaptors. Evidence for an adaptor complex in vivo. J. Biol. Chem. *271*, 5237–5245.

33. Candau, R., Moore, P., Wang, L., Barlev, N., Ying, C., Rosen, C., and Berger, S. (1996). Identification of human proteins functionally conserved with the yeast putative adaptors ADA2 and GCN5. Mol. Cell. Biol. *16*, 593–602.

34. Carrozza, M., Utley, R., Workman, J., and Côté, J. (2003). The diverse functions of histone acetyltransferase complexes. Trends Genet. TIG *19*, 321–329.

35. Chalkley, G.E., and Verrijzer, C.P. (1999). DNA binding site selection by RNA polymerase II TAFs: a TAFII250–TAFII150 complex recognizes the Initiator. EMBO J. *18*, 4835–4845.

36. Chen, S., McMullan, G., Faruqi, A.R., Murshudov, G.N., Short, J.M., Scheres, S.H.W., and Henderson, R. (2013). High-resolution noise substitution to measure overfitting and validate resolution in 3D structure determination by single particle electron cryomicroscopy. Ultramicroscopy *135*, 24–35.

37. Clark-Adams, C.D., Norris, D., Osley, M.A., Fassler, J.S., and Winston, F. (1988). Changes in histone gene dosage alter transcription in yeast. Genes Dev. *2*, 150–159.

38. Cler, E., Papai, G., Schultz, P., and Davidson, I. (2009). Recent advances in understanding the structure and function of general transcription factor TFIID. Cell. Mol. Life Sci. *66*, 2123–2134.

39. Côté, J., Quinn, J., Workman, J.L., and Peterson, C.L. (1994). Stimulation of GAL4 derivative binding to nucleosomal DNA by the yeast SWI/SNF complex. Science *265*, 53–60.

40. Cowieson, N.P., Kobe, B., and Martin, J.L. (2008). United we stand: combining structural methods. Curr. Opin. Struct. Biol. *18*, 617–622.

41. Daniel, J.A., Torok, M.S., Sun, Z.-W., Schieltz, D., Allis, C.D., Yates, J.R., and Grant, P.A. (2004). Deubiquitination of Histone H2B by a Yeast Acetyltransferase Complex Regulates Transcription. J. Biol. Chem. *279*, 1867–1871.

42. Dhalluin, C., Carlson, J., Zeng, L., He, C., Aggarwal, A., Zhou, M.-M., and Zhou, M.-M. (1999). Structure and ligand of a histone acetyltransferase bromodomain. Nature *399*, 491–496.

43. Dubochet, J., Adrian, M., Chang, J.-J., Homo, J.-C., Lepault, J., McDowall, A.W., and Schultz, P. (1988). Cryo-electron microscopy of vitrified specimens. Q. Rev. Biophys. *21*, 129.

44. Durand, A., Bonnet, J., Fournier, M., Chavant, V., and Schultz, P. (2014). Mapping the Deubiquitination Module within the SAGA Complex. Structure *22*, 1553–1559.

45. Durso, R.J., Fisher, A.K., Albright-Frey, T.J., and Reese, J.C. (2001). Analysis of TAF90 mutants displaying allele-specific and broad defects in transcription. Mol. Cell. Biol. *21*, 7331–7344.

46. Dyda, F., Klein, D.C., and Hickman, A.B. (2000). GCN5-Related N-Acetyltransferases: A Structural Overview. Annu. Rev. Biophys. Biomol. Struct. *29*, 81–103.

47. Dynlacht, B.D., Hoey, T., and Tjian, R. (1991). Isolation of coactivators associated with the TATA-binding protein that mediate transcriptional activation. Cell *66*, 563–576.

48. Eberharter, A., John, S., Grant, P., Utley, R., and Workman, J. (1998). Identification and analysis of yeast nucleosomal histone acetyltransferase complexes. Methods San Diego Calif *15*, 315–321.

49. Eichner, J., Chen, H.-T., Warfield, L., and Hahn, S. (2010). Position of the general transcription factor TFIIF within the RNA polymerase II transcription preinitiation complex. EMBO J. *29*, 706–716.

50. Eisenmann, D.M., Dollard, C., and Winston, F. (1989). SPT15, the gene encoding the yeast TATA binding factor TFIID, is required for normal transcription initiation in vivo. Cell *58*, 1183–1191.

51. Eisenmann, D.M., Arndt, K.M., Ricupero, S.L., Rooney, J.W., and Winston, F. (1992). SPT3 interacts with TFIID to allow normal transcription in Saccharomyces cerevisiae. Genes Dev. *6*, 1319–1331.

52. Eisenmann, D.M., Chapon, C., Roberts, S.M., Dollard, C., and Winston, F. (1994). The Saccharomyces cerevisiae SPT8 gene encodes a very acidic protein that is functionally related to SPT3 and TATA-binding protein. Genetics *137*, 647–657.

53. Elmore, Z.C., Beckley, J.R., Chen, J.-S., and Gould, K.L. (2014). Histone H2B Ubiquitination Promotes the Function of the Anaphase-Promoting Complex/Cyclosome in Schizosaccharomyces pombe. G3 GenesGenomesGenetics *4*, 1529–1538.

54. Erickson, H.P., and Klug, A. (1971). Measurement and Compensation of Defocusing and Aberrations by Fourier Processing of Electron Micrographs. Philos. Trans. R. Soc. B Biol. Sci. *261*, 105–118.

55. Erni, R., Rossell, M.D., Kisielowski, C., and Dahmen, U. (2009). Atomic-Resolution Imaging with a Sub-50-pm Electron Probe. Phys. Rev. Lett. *102*, 096101.

56. Ezhkova, E., and Tansey, W.P. (2004). Proteasomal ATPases link ubiquitylation of histone H2B to methylation of histone H3. Mol. Cell *13*, 435–442.

57. Ferreiro, J.A., Powell, N.G., Karabetsou, N., Mellor, J., and Waters, R. (2006). Roles for Gcn5p and Ada2p in transcription and nucleotide excision repair at the Saccharomyces cerevisiae MET16 gene. Nucleic Acids Res. *34*, 976–985.

58. Fishburn, J., Mohibullah, N., and Hahn, S. (2005). Function of a Eukaryotic Transcription Activator during the Transcription Cycle. Mol. Cell *18*, 369–378.

59. Frank, J. (1996). Three-dimensional electron microscopy of macromolecular assemblies (San Diego: Academic Press).

60. Gangloff, Y.-G., Werten, S., Romier, C., Carré, L., Poch, O., Moras, D., and Davidson, I. (2000). The Human TFIID Components TAFII135 and TAFII20 and the Yeast SAGA Components ADA1 and TAFII68 Heterodimerize to Form Histone-Like Pairs. Mol. Cell. Biol. *20*, 340–351.

61. Gangloff, Y.-G., Sanders, S.L., Romier, C., Kirschner, D., Weil, P.A., Tora, L., and Davidson, I. (2001). Histone Folds Mediate Selective Heterodimerization of Yeast TAFII25 with TFIID Components yTAFII47 and yTAFII65 and with SAGA Component ySPT7. Mol. Cell. Biol. *21*, 1841–1853.

62. Gansheroff, L.J., Dollard, C., Tan, P., and Winston, F. (1995). The Saccharomyces cerevisiae SPT7 gene encodes a very acidic protein important for transcription in vivo. Genetics *139*, 523–536.

63. García-Oliver, E., García-Molinero, V., and Rodríguez-Navarro, S. (2012). mRNA export and gene expression: The SAGA–TREX-2 connection. Biochim. Biophys. Acta BBA - Gene Regul. Mech. *1819*, 555–565.

64. Gavin, A.-C., Bösche, M., Krause, R., Grandi, P., Marzioch, M., Bauer, A., Schultz, J., Rick, J.M., Michon, A.-M., Cruciat, C.-M., et al. (2002). Functional organization of the yeast proteome by systematic analysis of protein complexes. Nature *415*, 141–147.

65. Georgakopoulos, T., and Thireos, G. (1992). Two distinct yeast transcriptional activators require the function of the GCN5 protein to promote normal levels of transcription. EMBO J. *11*, 4145–4152.

66. Georgakopoulos, T., Gounalaki, N., and Thireos, G. (1995). Genetic evidence for the interaction of the yeast transcriptional co-activator proteins GCN5 and ADA2. Mol. Gen. Genet. MGG *246*, 723–728.

67. Giniger, E., Varnum, S.M., and Ptashne, M. (1985). Specific DNA binding of GAL4, a positive regulatory protein of yeast. Cell *40*, 767–774.

68. Ginsburg, D.S., Govind, C.K., and Hinnebusch, A.G. (2009). NuA4 lysine acetyltransferase Esa1 is targeted to coding regions and stimulates transcription elongation with Gcn5. Mol. Cell. Biol. *29*, 6473–6487.

69. Glozak, M.A., Sengupta, N., Zhang, X., and Seto, E. (2005). Acetylation and deacetylation of non-histone proteins. Gene *363*, 15–23.

70. Goodrich, J.A., and Tjian, R. (1994). Transcription factors IIE and IIH and ATP hydrolysis direct promoter clearance by RNA polymerase II. Cell *77*, 145–156.

71. Govind, C.K., Zhang, F., Qiu, H., Hofmeyer, K., and Hinnebusch, A.G. (2007). Gcn5 promotes acetylation, eviction, and methylation of nucleosomes in transcribed coding regions. Mol. Cell *25*, 31–42.

72. Grant, P., and Berger, S. (1999). Histone acetyltransferase complexes. Semin. Cell Dev. Biol. *10*, 169–177.

73. Grant, P.A., Duggan, L., Côté, J., Roberts, S.M., Brownell, J.E., Candau, R., Ohba, R., Owen-Hughes, T., Allis, C.D., Winston, F., et al. (1997). Yeast Gcn5 functions in two multisubunit complexes to acetylate nucleosomal histones: characterization of an Ada complex and the SAGA (Spt/Ada) complex. Genes Dev. *11*, 1640–1650.

74. Grant, P.A., Schieltz, D., Pray-Grant, M.G., Steger, D.J., Reese, J.C., Yates III, J.R., and Workman, J.L. (1998a). A Subset of TAFIIs Are Integral Components of the SAGA Complex Required for Nucleosome Acetylation and Transcriptional Stimulation. Cell *94*, 45–53.

75. Grant, P.A., Schieltz, D., Pray-Grant, M.G., Yates III, J.R., and Workman, J.L. (1998b). The ATM-Related Cofactor Tra1 Is a Component of the Purified SAGA Complex. Mol. Cell *2*, 863–867.

76. Greber, B.J., Bieri, P., Leibundgut, M., Leitner, A., Aebersold, R., Boehringer, D., and Ban, N. (2015). Ribosome. The complete structure of the 55S mammalian mitochondrial ribosome. Science *348*, 303–308.

77. Gregory, P.D., Schmid, A., Zavari, M., Münsterkötter, M., and Hörz, W. (1999). Chromatin remodelling at the PHO8 promoter requires SWI-SNF and SAGA at a step subsequent to activator binding. EMBO J. *18*, 6407–6414.

78. Guelman, S., Suganuma, T., Florens, L., Swanson, S., Kiesecker, C., Kusch, T., Anderson, S., Yates, J., Washburn, M., Abmayr, S., et al. (2006). Host cell factor and an uncharacterized SANT domain protein are stable components of ATAC, a novel dAda2A/dGcn5-containing histone acetyltransferase complex in Drosophila. Mol. Cell. Biol. *26*, 871–882.

79. Gunderson, F.Q., and Johnson, T.L. (2009). Acetylation by the Transcriptional Coactivator Gcn5 Plays a Novel Role in Co-Transcriptional Spliceosome Assembly. PLoS Genet. *5*.

80. Gurskiy, D., Orlova, A., Vorobyeva, N., Nabirochkina, E., Krasnov, A., Shidlovskii, Y., Georgieva, S., and Kopytova, D. (2012). The DUBm subunit Sgf11 is required for mRNA export and interacts with Cbp80 in Drosophila. Nucleic Acids Res. *40*, 10689–10700.

81. Ha, I., Roberts, S., Maldonado, E., Sun, X., Kim, L.U., Green, M., and Reinberg, D. (1993). Multiple functional domains of human transcription factor IIB: distinct interactions with two general transcription factors and RNA polymerase II. Genes Dev. *7*, 1021–1032.

82. Han, Y., Luo, J., Ranish, J., and Hahn, S. (2014). Architecture of the Saccharomyces

cerevisiae SAGA transcription coactivator complex. EMBO J. *33*, 2534–2546.

83. Hassan, A.H., Neely, K.E., and Workman, J.L. (2001). Histone acetyltransferase complexes stabilize swi/snf binding to promoter nucleosomes. Cell *104*, 817–827.

84. Hassan, A.H., Prochasson, P., Neely, K.E., Galasinski, S.C., Chandy, M., Carrozza, M.J., and Workman, J.L. (2002). Function and selectivity of bromodomains in anchoring chromatin-modifying complexes to promoter nucleosomes. Cell *111*, 369–379.

85. Hayward, S.B., and Glaeser, R.M. (1979). Radiation damage of purple membrane at low temperature. Ultramicroscopy *4*, 201–210.

86. Hebbes, T.R., Thorne, A.W., and Crane-Robinson, C. (1988). A direct link between core histone acetylation and transcriptionally active chromatin. EMBO J. *7*, 1395–1402.

87. Van Heel, M. (1987). Angular reconstitution: A posteriori assignment of projection directions for 3D reconstruction. Ultramicroscopy *21*, 111–123.

88. van Heel, M., and Keegstra, W. (1981). IMAGIC: A fast, flexible and friendly image analysis software system. Ultramicroscopy *7*, 113–129.

89. Helmlinger, D., Hardy, S., Sasorith, S., Klein, F., Robert, F., Weber, C., Miguet, L., Potier, N., Van-Dorsselaer, A., Wurtz, J.-M., et al. (2004). Ataxin-7 is a subunit of GCN5 histone acetyltransferase-containing complexes. Hum. Mol. Genet. *13*, 1257–1265.

90. Helmlinger, D., Hardy, S., Eberlin, A., Devys, D., and Tora, L. (2006). Both normal and polyglutamine- expanded ataxin-7 are components of TFTC-type GCN5 histone acetyltransferase- containing complexes. Biochem. Soc. Symp. 155–163.

91. Helmlinger, D., Marguerat, S., Villén, J., Swaney, D.L., Gygi, S.P., Bähler, J., and Winston, F. (2011). Tra1 has specific regulatory roles, rather than global functions, within the SAGA co-activator complex. EMBO J. *30*, 2843–2852.

92. Henderson, R. (1990). Cryo-Protection of Protein Crystals against Radiation Damage in Electron and X-Ray Diffraction. Proc. R. Soc. Lond. B Biol. Sci. *241*, 6–8.

93. Henderson, R. (1995). The potential and limitations of neutrons, electrons and X-rays for atomic resolution microscopy of unstained biological molecules. Q. Rev. Biophys. *28*, 171.

94. Henderson, R., Chen, S., Chen, J.Z., Grigorieff, N., Passmore, L.A., Ciccarelli, L., Rubinstein, J.L., Crowther, R.A., Stewart, P.L., and Rosenthal, P.B. (2011). Tilt-Pair Analysis of Images from a Range of Different Specimens in Single-Particle Electron Cryomicroscopy. J. Mol. Biol. *413*, 1028–1046.

95. Henderson, R., Sali, A., Baker, M.L., Carragher, B., Devkota, B., Downing, K.H., Egelman, E.H., Feng, Z., Frank, J., Grigorieff, N., et al. (2012). Outcome of the First Electron Microscopy Validation Task Force Meeting. Structure *20*, 205–214.

96. Henry, K.W., Wyce, A., Lo, W.-S., Duggan, L.J., Emre, N.C.T., Kao, C.-F., Pillus, L., Shilatifard, A., Osley, M.A., and Berger, S.L. (2003). Transcriptional activation via sequential histone H2B ubiquitylation and deubiquitylation, mediated by SAGA-

associated Ubp8. Genes Dev. *17*, 2648–2663.

97. Hirschhorn, J.N., and Winston, F. (1988). SPT3 is required for normal levels of a-factor and alpha-factor expression in Saccharomyces cerevisiae. Mol. Cell. Biol. *8*, 822–827.

98. Ho, Y., Gruhler, A., Heilbut, A., Bader, G.D., Moore, L., Adams, S.-L., Millar, A., Taylor, P., Bennett, K., Boutilier, K., et al. (2002). Systematic identification of protein complexes in Saccharomyces cerevisiae by mass spectrometry. Nature *415*, 180–183.

99. Hoang, T.V., Cavin, X., Schultz, P., and Ritchie, D.W. (2013). gEMpicker: a highly parallel GPU-accelerated particle picking tool for cryo-electron microscopy. BMC Struct. Biol. *13*, 25.

100. Holstege, F.C., van der Vliet, P.C., and Timmers, H.T. (1996). Opening of an RNA polymerase II promoter occurs in two distinct steps and requires the basal transcription factors IIE and IIH. EMBO J. *15*, 1666–1677.

101. Hong, L., Schroth, G.P., Matthews, H.R., Yau, P., and Bradbury, E.M. (1993). Studies of the DNA binding properties of histone H4 amino terminus. Thermal denaturation studies reveal that acetylation markedly reduces the binding constant of the H4 “tail” to DNA. J. Biol. Chem. *268*, 305–314.

102. Horiuchi, J., Silverman, N., Marcus, G., and Guarente, L. (1995). ADA3, a putative transcriptional adaptor, consists of two separable domains and interacts with ADA2 and GCN5 in a trimeric complex. Mol. Cell. Biol. *15*, 1203–1209.

103. Houle, M.J., Powell, R.L., and Fintschenko, P. (1967). Absorbance of UV light by various carbohydrates in strong sulfuric acid. Anal. Biochem. *21*, 462–466.

104. Huang, B., Wang, W., Bates, M., and Zhuang, X. (2008). Three-Dimensional Super-Resolution Imaging by Stochastic Optical Reconstruction Microscopy. Science *319*, 810–813.

105. Huisinga, K.L., and Pugh, B.F. (2004). A Genome-Wide Housekeeping Role for TFIID and a Highly Regulated Stress-Related Role for SAGA in Saccharomyces cerevisiae. Mol. Cell *13*, 573–585.

106. Ikeda, K., Steger, D.J., Eberharter, A., and Workman, J.L. (1999). Activation domain-specific and general transcription stimulation by native histone acetyltransferase complexes. Mol. Cell. Biol. *19*, 855–863.

107. Imbalzano, A.N., Zaret, K.S., and Kingston, R.E. (1994). Transcription factor (TF) IIB and TFIIA can independently increase the affinity of the TATA-binding protein for DNA. J. Biol. Chem. *269*, 8280–8286.

108. Ingvarsdottir, K., Krogan, N.J., Emre, N.C.T., Wyce, A., Thompson, N.J., Emili, A., Hughes, T.R., Greenblatt, J.F., and Berger, S.L. (2005). H2B ubiquitin protease Ubp8 and Sgf11 constitute a discrete functional module within the Saccharomyces cerevisiae SAGA complex. Mol. Cell. Biol. *25*, 1162–1172.

109. Jeong, C.J., Yang, S.H., Xie, Y., Zhang, L., Johnston, S.A., and Kodadek, T. (2001). Evidence that Gal11 protein is a target of the Gal4 activation domain in the

mediator. Biochemistry (Mosc.) *40*, 9421–9427.

110. Johnston, M., and Dover, J. (1987). Mutations that inactivate a yeast transcriptional regulatory protein cluster in an evolutionarily conserved DNA binding domain. Proc. Natl. Acad. Sci. U. S. A. *84*, 2401–2405.

111. Kastner, B., Fischer, N., Golas, M.M., Sander, B., Dube, P., Boehringer, D., Hartmuth, K., Deckert, J., Hauer, F., Wolf, E., et al. (2008). GraFix: sample preparation for single-particle electron cryomicroscopy. Nat. Methods *5*, 53–55.

112. Keefe, A.D., Wilson, D.S., Seelig, B., and Szostak, J.W. (2001). One-Step Purification of Recombinant Proteins Using a Nanomolar-Affinity Streptavidin-Binding Peptide, the SBP-Tag. Protein Expr. Purif. *23*, 440–446.

113. Khatter, H., Myasnikov, A.G., Natchiar, S.K., and Klaholz, B.P. (2015). Structure of the human 80S ribosome. Nature *520*, 640–645.

114. Kirschner, D.B., Baur, E. vom, Thibault, C., Sanders, S.L., Gangloff, Y.-G., Davidson, I., Weil, P.A., and Tora, L. (2002). Distinct Mutations in Yeast TAFII25 Differentially Affect the Composition of TFIID and SAGA Complexes as Well as Global Gene Expression Patterns. Mol. Cell. Biol. *22*, 3178–3193.

115. Klein, J., Nolden, M., Sanders, S.L., Kirchner, J., Weil, P.A., and Melcher, K. (2003). Use of a genetically introduced cross-linker to identify interaction sites of acidic activators within native transcription factor IID and SAGA. J. Biol. Chem. *278*, 6779–6786.

116. Knoll, M., and Ruska, E. (1932). Das Elektronenmikroskop. Z. Für Phys. *78*, 318–339.

117. Knutson, B.A., and Hahn, S. (2011). Domains of Tra1 Important for Activator Recruitment and Transcription Coactivator Functions of SAGA and NuA4 Complexes. Mol. Cell. Biol. *31*, 818–831.

118. Knutson, B.A., Luo, J., Ranish, J., and Hahn, S. (2014). Architecture of the Saccharomyces cerevisiae RNA polymerase I Core Factor complex. Nat. Struct. Mol. Biol. *21*, 810–816.

119. Koh, S.S., Ansari, A.Z., Ptashne, M., and Young, R.A. (1998). An Activator Target in the RNA Polymerase II Holoenzyme. Mol. Cell *1*, 895–904.

120. Köhler, A., Pascual-García, P., Llopis, A., Zapater, M., Posas, F., Hurt, E., and Rodríguez-Navarro, S. (2006). The mRNA export factor Sus1 is involved in Spt/Ada/Gcn5 acetyltransferase-mediated H2B deubiquitinylation through its interaction with Ubp8 and Sgf11. Mol. Biol. Cell *17*, 4228–4236.

121. Köhler, A., Schneider, M., Cabal, G.G., Nehrbass, U., and Hurt, E. (2008). Yeast Ataxin-7 links histone deubiquitination with gene gating and mRNA export. Nat. Cell Biol. *10*, 707–715.

122. Köhler, A., Zimmerman, E., Schneider, M., Hurt, E., and Zheng, N. (2010). Structural Basis for Assembly and Activation of the Hetero-tetrameric SAGA Histone

H2B Deubiquitinase Module. Cell *141*, 606–617.

123. Krebs, A., Demmers, J., Karmodiya, K., Chang, N.-C., Chang, A.C., and Tora, L. (2010). ATAC and Mediator coactivators form a stable complex and regulate a set of non-coding RNA genes. EMBO Rep. *11*, 541–547.

124. Krebs, A., Karmodiya, K., Lindahl-Allen, M., Struhl, K., and Tora, L. (2011). SAGA and ATAC histone acetyl transferase complexes regulate distinct sets of genes and ATAC defines a class of p300-independent enhancers. Mol. Cell *44*, 410–423.

125. Kucukelbir, A., Sigworth, F.J., and Tagare, H.D. (2014). Quantifying the local resolution of cryo-EM density maps. Nat. Methods *11*, 63–65.

126. Kuhlbrandt, W. (2014). The Resolution Revolution. Science *343*, 1443–1444.

127. Kulesza, C.A., Van Buskirk, H.A., Cole, M.D., Reese, J.C., Smith, M.M., and Engel, D.A. (2002). Adenovirus E1A requires the yeast SAGA histone acetyltransferase complex and associates with SAGA components Gcn5 and Tra1. Oncogene *21*, 1411–1422.

128. Kuo, M., Brownell, J., Sobel, R., Ranalli, T., Cook, R., Edmondson, D., Roth, S., and Allis, C. (1996). Transcription-linked acetylation by Gcn5p of histones H3 and H4 at specific lysines. Nature *383*, 269–272.

129. Kurabe, N., Katagiri, K., Komiya, Y., Ito, R., Sugiyama, A., Kawasaki, Y., and Tashiro, F. (2007). Deregulated expression of a novel component of TFTC/STAGA histone acetyltransferase complexes, rat SGF29, in hepatocellular carcinoma: possible implication for the oncogenic potential of c-Myc. Oncogene *26*, 5626–5634.

130. Kurshakova, M.M., Kopytova, D.V., Nabirochkina, E.N., Nikolenko, Y.V., Shidlovskii, Y.V., Georgieva, S.G., and Krasnov, A.N. (2009). Conservative E(y)2/Sus1 protein is the member of SAGA complex and new nuclear pore-associated complex in Drosophila. Russ. J. Genet. *45*, 1174–1181.

131. Lan, X., Koutelou, E., Schibler, A.C., Chen, Y.C., Grant, P.A., and Dent, S.Y.R. (2015). Poly(Q) Expansions in ATXN7 Affect Solubility but Not Activity of the SAGA Deubiquitinating Module. Mol. Cell. Biol. *35*, 1777–1787.

132. Längst, G., Bonte, E.J., Corona, D.F., and Becker, P.B. (1999). Nucleosome movement by CHRAC and ISWI without disruption or trans-displacement of the histone octamer. Cell *97*, 843–852.

133. Laprade, L., Rose, D., and Winston, F. (2007). Characterization of New Spt3 and TATA-Binding Protein Mutants of Saccharomyces cerevisiae: Spt3–TBP Allele-Specific Interactions and Bypass of Spt8. Genetics *177*, 2007–2017.

134. Larschan, E., and Winston, F. (2001). The S. cerevisiae SAGA complex functions in vivo as a coactivator for transcriptional activation by Gal4. Genes Dev. *15*, 1946–1956.

135. Lee, K., and Workman, J. (2007). Histone acetyltransferase complexes: one size doesn't fit all. Nat. Rev. Mol. Cell Biol. *8*, 284–295.

136. Lee, K.K., Florens, L., Swanson, S.K., Washburn, M.P., and Workman, J.L. (2005). The deubiquitylation activity of Ubp8 is dependent upon Sgf11 and its association with the SAGA complex. Mol. Cell. Biol. *25*, 1173–1182.

137. Lee, K.K., Sardiu, M.E., Swanson, S.K., Gilmore, J.M., Torok, M., Grant, P.A., Florens, L., Workman, J.L., and Washburn, M.P. (2011). Combinatorial depletion analysis to assemble the network architecture of the SAGA and ADA chromatin remodeling complexes. Mol. Syst. Biol. *7*.

138. Lee, T.I., Causton, H.C., Holstege, F.C., Shen, W.C., Hannett, N., Jennings, E.G., Winston, F., Green, M.R., and Young, R.A. (2000). Redundant roles for the TFIID and SAGA complexes in global transcription. Nature *405*, 701–704.

139. Leschziner, A. (2010). Chapter Nine - The Orthogonal Tilt Reconstruction Method. In Methods in Enzymology, G.J. Jensen, ed. (Academic Press), pp. 237–262.

140. 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.

141. Li, X., Mooney, P., Zheng, S., Booth, C.R., 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 cryo-EM. Nat. Methods *10*, 584–590.

142. Liao, M., Cao, E., Julius, D., and Cheng, Y. (2013). Structure of the TRPV1 ion channel determined by electron cryo-microscopy. Nature *504*, 107–112.

143. Lim, S., Kwak, J., Kim, M., and Lee, D. (2013). Separation of a functional deubiquitylating module from the SAGA complex by the proteasome regulatory particle. Nat. Commun. *4*.

144. Liu, X., Tesfai, J., Evrard, Y.A., Dent, S.Y.R., and Martinez, E. (2003). c-Myc transformation domain recruits the human STAGA complex and requires TRRAP and GCN5 acetylase activity for transcription activation. J. Biol. Chem. *278*, 20405–20412.

145. Liu, X., Bushnell, D.A., Wang, D., Calero, G., and Kornberg, R.D. (2010). Structure of an RNA Polymerase II–TFIIB Complex and the Transcription Initiation Mechanism. Science *327*, 206–209.

146. Lu, H., Zawel, L., Fisher, L., Egly, J.-M., and Reinberg, D. (1992). Human general transcription factor IIH phosphorylates the C-terminal domain of RNA polymerase II. Nature *358*, 641–645.

147. Lu, P., Bai, X., Ma, D., Xie, T., Yan, C., Sun, L., Yang, G., Zhao, Y., Zhou, R., Scheres, S.H.W., et al. (2014). Three-dimensional structure of human γ-secretase. Nature *512*, 166–170.

148. Marcus, G., Silverman, N., Berger, S., Horiuchi, J., and Guarente, L. (1994). Functional similarity and physical association between GCN5 and ADA2: putative transcriptional adaptors. EMBO J. *13*, 4807–4815.

149. Marmorstein, R., Carey, M., Ptashne, M., and Harrison, S.C. (1992). DNA

recognition by GAL4: structure of a protein-DNA complex. Nature *356*, 408–414.

150. Martinez, E., Kundu, T., Fu, J., and Roeder, R. (1998). A human SPT3-TAFII31-GCN5-L acetylase complex distinct from transcription factor IID. J. Biol. Chem. *273*, 23781–23785.

151. Martinez, E., Palhan, V.B., Tjernberg, A., Lymar, E.S., Gamper, A.M., Kundu, T.K., Chait, B.T., and Roeder, R.G. (2001). Human STAGA Complex Is a Chromatin-Acetylating Transcription Coactivator That Interacts with Pre-mRNA Splicing and DNA Damage-Binding Factors In Vivo. Mol. Cell. Biol. *21*, 6782–6795.

152. McMahon, S.B., Van Buskirk, H.A., Dugan, K.A., Copeland, T.D., and Cole, M.D. (1998). The Novel ATM-Related Protein TRRAP Is an Essential Cofactor for the c-Myc and E2F Oncoproteins. Cell *94*, 363–374.

153. McMahon, S.J., Pray-Grant, M.G., Schieltz, D., Yates, J.R., and Grant, P.A. (2005). Polyglutamine-expanded spinocerebellar ataxia-7 protein disrupts normal SAGA and SLIK histone acetyltransferase activity. Proc. Natl. Acad. Sci. U. S. A. *102*, 8478–8482.

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

155. Melcher, K., and Johnston, S.A. (1995). GAL4 interacts with TATA-binding protein and coactivators. Mol. Cell. Biol. *15*, 2839–2848.

156. Mengus, G., May, M., Jacq, X., Staub, A., Tora, L., Chambon, P., and Davidson, I. (1995). Cloning and characterization of hTAFII18, hTAFII20 and hTAFII28: three subunits of the human transcription factor TFIID. EMBO J. *14*, 1520–1531.

157. Merkhofer, E., and Johnson, T. (2014). SAGA-mediated acetylation interacts with Prp5 ATPase activity in pre-mRNA splicing (940.1). FASEB J. *28*, 940.1.

158. Michel, B., Komarnitsky, P., and Buratowski, S. (1998). Histone-like TAFs Are Essential for Transcription In Vivo. Mol. Cell *2*, 663–673.

159. Mirza, S., Rakha, E., Alshareeda, A., Mohibi, S., Zhao, X., Katafiasz, B., Wang, J., Gurumurthy, C.B., Bele, A., Ellis, I., et al. (2013). Cytoplasmic localization of alteration/deficiency in activation 3 (ADA3) predicts poor clinical outcome in breast cancer patients. Breast Cancer Res. Treat. *137*, 721–731.

160. Mischerikow, N., Spedale, G., Altelaar, A.F.M., Timmers, H.T.M., Pijnappel, W.W.M.P., and Heck, A.J.R. (2009). In-Depth Profiling of Post-Translational Modifications on the Related Transcription Factor Complexes TFIID and SAGA. J. Proteome Res. *8*, 5020–5030.

161. Mohibi, S., Gurumurthy, C.B., Nag, A., Wang, J., Mirza, S., Mian, Y., Quinn, M., Katafiasz, B., Eudy, J., Pandey, S., et al. (2012). Mammalian alteration/deficiency in activation 3 (Ada3) is essential for embryonic development and cell cycle progression. J. Biol. Chem. *287*, 29442–29456.

162. Mohibullah, N., and Hahn, S. (2008). Site-specific cross-linking of TBP in vivo and in vitro reveals a direct functional interaction with the SAGA subunit Spt3. Genes Dev. *22*, 2994–3006.

163. Van Mullem, V., Wery, M., Werner, M., Vandenhaute, J., and Thuriaux, P. (2002). The Rpb9 subunit of RNA polymerase II binds transcription factor TFIIE and interferes with the SAGA and elongator histone acetyltransferases. J. Biol. Chem. *277*, 10220–10225.

164. Müller, M.Q., and Sinz, A. (2012). Chemical cross-linking and high-resolution mass spectrometry to study protein-drug interactions. Methods Mol. Biol. Clifton NJ *803*, 205–218.

165. Müller, D.R., Schindler, P., Towbin, H., Wirth, U., Voshol, H., Hoving, S., and Steinmetz, M.O. (2001). Isotope-Tagged Cross-Linking Reagents. A New Tool in Mass Spectrometric Protein Interaction Analysis. Anal. Chem. *73*, 1927–1934.

166. Murakami, K., Elmlund, H., Kalisman, N., Bushnell, D.A., Adams, C.M., Azubel, M., Elmlund, D., Levi-Kalisman, Y., Liu, X., Gibbons, B.J., et al. (2013). Architecture of an RNA polymerase II transcription pre-initiation complex. Science *342*, 1238724.

167. Nagy, Z., and Tora, L. (2007). Distinct GCN5/PCAF-containing complexes function as co-activators and are involved in transcription factor and global histone acetylation. Oncogene *26*, 5341–5357.

168. Nagy, Z., Riss, A., Romier, C., le Guezennec, X., Dongre, A.R., Orpinell, M., Han, J., Stunnenberg, H., and Tora, L. (2009). The Human SPT20-Containing SAGA Complex Plays a Direct Role in the Regulation of Endoplasmic Reticulum Stress-Induced Genes. Mol. Cell. Biol. *29*, 1649–1660.

169. Nakayama, J., Rice, J.C., Strahl, B.D., Allis, C.D., and Grewal, S.I.S. (2001). Role of Histone H3 Lysine 9 Methylation in Epigenetic Control of Heterochromatin Assembly. Science *292*, 110–113.

170. Narlikar, G.J., Sundaramoorthy, R., and Owen-Hughes, T. (2013). Mechanisms and Functions of ATP-Dependent Chromatin-Remodeling Enzymes. Cell *154*, 490–503.

171. Natarajan, K., Jackson, B.M., Rhee, E., and Hinnebusch, A.G. (1998). yTAFII61 Has a General Role in RNA Polymerase II Transcription and Is Required by Gcn4p to Recruit the SAGA Coactivator Complex. Mol. Cell *2*, 683–692.

172. Neuwald, A.F., and Landsman, D. (1997). GCN5-related histone N-acetyltransferases belong to a diverse superfamily that includes the yeast SPT10 protein. Trends Biochem. Sci. *22*, 154–155.

173. Nikolov, D.B., Chen, H., Halay, E.D., Usheva, A.A., Hisatake, K., Lee, D.K., Roeder, R.G., and Burley, S.K. (1995). Crystal structure of a TFIIB–TBP–TATA-element ternary complex. Nature *377*, 119–128.

174. Nourani, A., Utley, R.T., Allard, S., and Côté, J. (2004). Recruitment of the NuA4 complex poises the PHO5 promoter for chromatin remodeling and activation. EMBO J. *23*, 2597–2607.

175. Ogryzko, V., Kotani, T., Zhang, X., Schiltz, R., Howard, T., Yang, X., Howard, B., Qin, J., and Nakatani, Y. (1998). Histone-like TAFs within the PCAF histone acetylase complex. Cell *94*, 35–44.

176. Orlova, E.V., and Saibil, H.R. (2011). Structural Analysis of Macromolecular Assemblies by Electron Microscopy. Chem. Rev. *111*, 7710–7748.

177. Owen-Hughes, T., Utley, R.T., Côté, J., Peterson, C.L., and Workman, J.L. (1996). Persistent site-specific remodeling of a nucleosome array by transient action of the SWI/SNF complex. Science *273*, 513–516.

178. Palhan, V.B., Chen, S., Peng, G.-H., Tjernberg, A., Gamper, A.M., Fan, Y., Chait, B.T., La Spada, A.R., and Roeder, R.G. (2005). Polyglutamine-expanded ataxin-7 inhibits STAGA histone acetyltransferase activity to produce retinal degeneration. Proc. Natl. Acad. Sci. U. S. A. *102*, 8472–8477.

179. Pascual-García, P., and Rodríguez-Navarro, S. (2009). A tale of coupling, Sus1 function in transcription and mRNA export. RNA Biol. *6*, 141–144.

180. Pascual-García, P., Govind, C.K., Queralt, E., Cuenca-Bono, B., Llopis, A., Chavez, S., Hinnebusch, A.G., and Rodríguez-Navarro, S. (2008). Sus1 is recruited to coding regions and functions during transcription elongation in association with SAGA and TREX2. Genes Dev. *22*, 2811–2822.

181. Paulsen, C.E., Armache, J.-P., Gao, Y., Cheng, Y., and Julius, D. (2015). Structure of the TRPA1 ion channel suggests regulatory mechanisms. Nature *520*, 511–517.

182. Penn, M.D., Galgoci, B., and Greer, H. (1983). Identification of AAS genes and their regulatory role in general control of amino acid biosynthesis in yeast. Proc. Natl. Acad. Sci. U. S. A. *80*, 2704–2708.

183. Pennacchio, L.A., Bickmore, W., Dean, A., Nobrega, M.A., and Bejerano, G. (2013). Enhancers: five essential questions. Nat. Rev. Genet. *14*, 288–295.

184. Perrakis, A., Musacchio, A., Cusack, S., and Petosa, C. (2011). Investigating a macromolecular complex: The toolkit of methods. J. Struct. Biol. *175*, 106–112.

185. Pijnappel, W.W.M.P., and Timmers, H.T.M. (2008). Dubbing SAGA Unveils New Epigenetic Crosstalk. Mol. Cell *29*, 152–154.

186. Pollard, K., and Peterson, C. (1997). Role for ADA/GCN5 products in antagonizing chromatin-mediated transcriptional repression. Mol. Cell. Biol. *17*, 6212–6222.

187. Powell, D.W., Weaver, C.M., Jennings, J.L., McAfee, K.J., He, Y., Weil, P.A., and Link, A.J. (2004). Cluster analysis of mass spectrometry data reveals a novel component of SAGA. Mol. Cell. Biol. *24*, 7249–7259.

188. Pray-Grant, M.G., Schieltz, D., McMahon, S.J., Wood, J.M., Kennedy, E.L., Cook, R.G., Workman, J.L., Yates III, J.R., and Grant, P.A. (2002). The Novel SLIK Histone Acetyltransferase Complex Functions in the Yeast Retrograde Response Pathway. Mol. Cell. Biol. *22*, 8774–8786.

189. Puig, O., Caspary, F., Rigaut, G., Rutz, B., Bouveret, E., Bragado-Nilsson, E., Wilm, M., and Séraphin, B. (2001). The Tandem Affinity Purification (TAP) Method: A General Procedure of Protein Complex Purification. Methods *24*, 218–229.

190. Qian, C., Zhang, Q., Li, S., Zeng, L., Walsh, M., and Zhou, M.-M. (2005). Structure and chromosomal DNA binding of the SWIRM domain. Nat. Struct. Mol. Biol. *12*, 1078–1085.

191. Radermacher, M. (1988). Three-Dimensional reconstruction of single particles from random and nonrandom tilt series. J. Electron Microsc. Tech. *9*, 359–394.

192. Rando, O.J., and Winston, F. (2012). Chromatin and Transcription in Yeast. Genetics *190*, 351–387.

193. Rappsilber, J. (2011). The beginning of a beautiful friendship: Cross-linking/mass spectrometry and modelling of proteins and multi-protein complexes. J. Struct. Biol. *173*, 530–540.

194. Reeves, W.M., and Hahn, S. (2005). Targets of the Gal4 Transcription Activator in Functional Transcription Complexes. Mol. Cell. Biol. *25*, 9092–9102.

195. Rhee, H.S., and Pugh, F. (2012). Genome-wide structure and organization of eukaryotic pre-initiation complexes. Nature *483*, 295–301.

196. Rigaut, G., Shevchenko, A., Rutz, B., Wilm, M., Mann, M., and Séraphin, B. (1999). A generic protein purification method for protein complex characterization and proteome exploration. Nat. Biotechnol. *17*, 1030–1032.

197. Roberts, S.M., and Winston, F. (1996). SPT20/ADA5 encodes a novel protein functionally related to the TATA-binding protein and important for transcription in Saccharomyces cerevisiae. Mol. Cell. Biol. *16*, 3206–3213.

198. Roberts, S., and Winston, F. (1997). Essential functional interactions of SAGA, a Saccharomyces cerevisiae complex of Spt, Ada, and Gcn5 proteins, with the Snf/Swi and Srb/mediator complexes. Genetics *147*, 451–465.

199. Rodríguez-Navarro, S., Fischer, T., Luo, M.-J., Antúnez, O., Brettschneider, S., Lechner, J., Pérez-Ortín, J.E., Reed, R., and Hurt, E. (2004). Sus1, a Functional Component of the SAGA Histone Acetylase Complex and the Nuclear Pore-Associated mRNA Export Machinery. Cell *116*, 75–86.

200. Rosenthal, P.B., and Henderson, R. (2003). Optimal Determination of Particle Orientation, Absolute Hand, and Contrast Loss in Single-particle Electron Cryomicroscopy. J. Mol. Biol. *333*, 721–745.

201. Sainsbury, S., Bernecky, C., and Cramer, P. (2015). Structural basis of transcription initiation by RNA polymerase II. Nat. Rev. Mol. Cell Biol. *16*, 129–143.

202. Samara, N.L., Datta, A.B., Berndsen, C.E., Zhang, X., Yao, T., Cohen, R.E., and Wolberger, C. (2010). Structural insights into the assembly and function of the SAGA deubiquitinating module. Science *328*, 1025–1029.

203. Sanders, S.L., and Weil, P.A. (2000). Identification of two novel TAF subunits of the yeast Saccharomyces cerevisiae TFIID complex. J. Biol. Chem. *275*, 13895–13900.

204. Sanders, S., Jennings, J., Canutescu, A., Link, A., and Weil, A. (2002). Proteomics of the eukaryotic transcription machinery: identification of proteins associated with components of yeast TFIID by multidimensional mass spectrometry. Mol. Cell. Biol. *22*, 4723–4738.

205. Scheres, S.H.W. (2012). RELION: Implementation of a Bayesian approach to cryo-EM structure determination. J. Struct. Biol. *180*, 519–530.

206. Scheres, S.H.W. (2015). Semi-automated selection of cryo-EM particles in RELION-1.3. J. Struct. Biol. *189*, 114–122.

207. Scheres, S.H.W., and Chen, S. (2012). Prevention of overfitting in cryo-EM structure determination. Nat. Methods *9*, 853–854.

208. Scheres, S.H.W., Núñez-Ramírez, R., Sorzano, C.O.S., Carazo, J.M., and Marabini, R. (2008). Image processing for electron microscopy single-particle analysis using XMIPP. Nat. Protoc. *3*, 977–990.

209. Scheres, S.H.W., Melero, R., Valle, M., and Carazo, J.-M. (2009). Averaging of Electron Subtomograms and Random Conical Tilt Reconstructions through Likelihood Optimization. Structure *17*, 1563–1572.

210. Schnitzler, G.R., Cheung, C.L., Hafner, J.H., Saurin, A.J., Kingston, R.E., and Lieber, C.M. (2001). Direct Imaging of Human SWI/SNF-Remodeled Mono- and Polynucleosomes by Atomic Force Microscopy Employing Carbon Nanotube Tips. Mol. Cell. Biol. *21*, 8504–8511.

211. Schwarz, R., Tänzler, D., Ihling, C.H., Müller, M.Q., Kölbel, K., and Sinz, A. (2013). Monitoring conformational changes in peroxisome proliferator-activated receptor α by a genetically encoded photoamino acid, cross-linking, and mass spectrometry. J. Med. Chem. *56*, 4252–4263.

212. Sealy, L., and Chalkley, R. (1978). DNA associated with hyperacetylated histone is preferentially digested by DNase I. Nucleic Acids Res. *5*, 1863–1876.

213. Selleck, W., Howley, R., Fang, Q., Podolny, V., Fried, M.G., Buratowski, S., and Tan, S. (2001). A histone fold TAF octamer within the yeast TFIID transcriptional coactivator. Nat. Struct. Mol. Biol. *8*, 695–700.

214. Sermwittayawong, D., and Tan, S. (2006). SAGA binds TBP via its Spt8 subunit in competition with DNA: implications for TBP recruitment. EMBO J. *25*, 3791–3800.

215. Setiaputra, D., Ross, J.D., Lu, S., Cheng, D.T., Dong, M.-Q., and Yip, C.K. (2015). Conformational Flexibility and Subunit Arrangement of the Modular Yeast Spt-Ada-Gcn5 Acetyltransferase Complex. J. Biol. Chem. *290*, 10057–10070.

216. Shilatifard, A. (2006). Chromatin modifications by methylation and ubiquitination: implications in the regulation of gene expression. Annu. Rev. Biochem. *75*, 243–269.

217. Shlyueva, D., Stampfel, G., and Stark, A. (2014). Transcriptional enhancers: from properties to genome-wide predictions. Nat. Rev. Genet. *15*, 272–286.

218. Sigworth, F.J. (1998). A Maximum-Likelihood Approach to Single-Particle Image Refinement. J. Struct. Biol. *122*, 328–339.

219. Smale, S.T., and Kadonaga, J.T. (2003). The Rna Polymerase Ii Core Promoter. Annu. Rev. Biochem. *72*, 449–479.

220. Smith, E., Belote, J., Schiltz, R., Yang, X., Moore, P., Berger, S., Nakatani, Y., and Allis, C. (1998). Cloning of Drosophila GCN5: conserved features among metazoan GCN5 family members. Nucleic Acids Res. *26*, 2948–2954.

221. Spedale, G., Mischerikow, N., Heck, A.J.R., Timmers, H.T.M., and Pijnappel, W.W.M.P. (2010). Identification of Pep4p as the Protease Responsible for Formation of the SAGA-related SLIK Protein Complex. J. Biol. Chem. *285*, 22793–22799.

222. Spedale, G., Timmers, H.T.M., and Pijnappel, P.W. (2012). ATAC-king the complexity of SAGA during evolution. Genes Dev. *26*, 527–541.

223. Spitz, F., and Furlong, E.E.M. (2012). Transcription factors: from enhancer binding to developmental control. Nat. Rev. Genet. *13*, 613–626.

224. Stark, H., Zemlin, F., and Boettcher, C. (1996). Electron radiation damage to protein crystals of bacteriorhodopsin at different temperatures. Ultramicroscopy *63*, 75–79.

225. Sterner, D., Wang, X., Bloom, M., Simon, G., and Berger, S. (2002a). The SANT domain of Ada2 is required for normal acetylation of histones by the yeast SAGA complex. J. Biol. Chem. *277*, 8178–8186.

226. Sterner, D.E., Grant, P.A., Roberts, S.M., Duggan, L.J., Belotserkovskaya, R., Pacella, L.A., Winston, F., Workman, J.L., and Berger, S.L. (1999). Functional organization of the yeast SAGA complex: distinct components involved in structural integrity, nucleosome acetylation, and TATA-binding protein interaction. Mol. Cell. Biol. *19*, 86–98.

227. Sterner, D.E., Belotserkovskaya, R., and Berger, S.L. (2002b). SALSA, a variant of yeast SAGA, contains truncated Spt7, which correlates with activated transcription. Proc. Natl. Acad. Sci. *99*, 11622–11627.

228. Suganuma, T., Gutiérrez, J., Li, B., Florens, L., Swanson, S., Washburn, M., Abmayr, S., and Workman, J. (2008). ATAC is a double histone acetyltransferase complex that stimulates nucleosome sliding. Nat. Struct. Mol. Biol. *15*, 364–372.

229. Syntichaki, P., Topalidou, I., and Thireos, G. (2000). The Gcn5 bromodomain co-ordinates nucleosome remodelling. Nature *404*, 414–417.

230. Thomas, M.C., and Chiang, C.-M. (2006). The General Transcription Machinery and General Cofactors. Crit. Rev. Biochem. Mol. Biol. *41*, 105–178.

231. Thon, F. (1966). Zur Defokussierungsäbhängigkeit des Phasenkontrastes bei der

elektronenmikroskopischen Abbildung. Z Naturforsch *21a*, 476–478.

232. Todd, T.W., and Lim, J. (2013). Aggregation formation in the polyglutamine diseases: Protection at a cost? Mol. Cells *36*, 185–194.

233. Tollervey, J.R., and Lunyak, V.V. (2012). Epigenetics. Epigenetics *7*, 823–840.

234. Traven, A., Jelicic, B., and Sopta, M. (2006). Yeast Gal4: a transcriptional paradigm revisited. EMBO Rep. *7*, 496–499.

235. Triezenberg, S.J., Kingsbury, R.C., and McKnight, S.L. (1988). Functional dissection of VP16, the trans-activator of herpes simplex virus immediate early gene expression. Genes Dev. *2*, 718–729.

236. Trowitzsch, S., Bieniossek, C., Nie, Y., Garzoni, F., and Berger, I. (2010). New baculovirus expression tools for recombinant protein complex production. J. Struct. Biol. *172*, 45–54.

237. Unwin, P.N., and Henderson, R. (1975). Molecular structure determination by electron microscopy of unstained crystalline specimens. J. Mol. Biol. *94*, 425–440.

238. Utley, R., Ikeda, K., Grant, P., Côté, J., Steger, D., Eberharter, A., John, S., and Workman, J. (1998). Transcriptional activators direct histone acetyltransferase complexes to nucleosomes. Nature *394*, 498–502.

239. Venters, B.J., Wachi, S., Mavrich, T.N., Andersen, B.E., Jena, P., Sinnamon, A.J., Jain, P., Rolleri, N.S., Jiang, C., Hemeryck-Walsh, C., et al. (2011). A Comprehensive Genomic Binding Map of Gene and Chromatin Regulatory Proteins in Saccharomyces. Mol. Cell *41*, 480–492.

240. Vermeulen, M., Eberl, H.C., Matarese, F., Marks, H., Denissov, S., Butter, F., Lee, K.K., Olsen, J.V., Hyman, A.A., Stunnenberg, H.G., et al. (2010). Quantitative interaction proteomics and genome-wide profiling of epigenetic histone marks and their readers. Cell *142*, 967–980.

241. Vidali, G., Boffa, L.C., Bradbury, E.M., and Allfrey, V.G. (1978). Butyrate suppression of histone deacetylation leads to accumulation of multiacetylated forms of histones H3 and H4 and increased DNase I sensitivity of the associated DNA sequences. Proc. Natl. Acad. Sci. *75*, 2239–2243.

242. Vignali, M., Steger, D.J., Neely, K.E., and Workman, J.L. (2000). Distribution of acetylated histones resulting from Gal4-VP16 recruitment of SAGA and NuA4 complexes. EMBO J. *19*, 2629–2640.

243. Vogelauer, M., Wu, J., Suka, N., and Grunstein, M. (2000). Global histone acetylation and deacetylation in yeast. Nature *408*, 495–498.

244. Voorhees, R.M., Fernández, I.S., Scheres, S.H.W., and Hegde, R.S. (2014). Structure of the Mammalian Ribosome-Sec61 Complex to 3.4 Å Resolution. Cell *157*, 1632–1643.

245. Walker, A.K., Rothman, J.H., Shi, Y., and Blackwell, T.K. (2001). Distinct

requirements for C.elegans TAF(II)s in early embryonic transcription. EMBO J. *20*, 5269–5279.

246. Wallberg, A.E., Flinn, E.M., Gustafsson, J.A., and Wright, A.P. (2000). Recruitment of chromatin remodelling factors during gene activation via the glucocorticoid receptor N-terminal domain. Biochem. Soc. Trans. *28*, 410–414.

247. Walz, J., Typke, D., Nitsch, M., Koster, A.J., Hegerl, R., and Baumeister, W. (1997). Electron Tomography of Single Ice-Embedded Macromolecules: Three-Dimensional Alignment and Classification. J. Struct. Biol. *120*, 387–395.

248. Wang, Y.-L., Faiola, F., Xu, M., Pan, S., and Martinez, E. (2008). Human ATAC Is a GCN5/PCAF-containing acetylase complex with a novel NC2-like histone fold module that interacts with the TATA-binding protein. J. Biol. Chem. *283*, 33808–33815.

249. Warfield, L., Ranish, J.A., and Hahn, S. (2004). Positive and negative functions of the SAGA complex mediated through interaction of Spt8 with TBP and the N-terminal domain of TFIIA. Genes Dev. *18*, 1022–1034.

250. Watson, J.D. (2014). Molecular biology of the gene.

251. Watson, J.D., and Crick, F.H.C. (1953). Molecular Structure of Nucleic Acids: A Structure for Deoxyribose Nucleic Acid. Nature *171*, 737–738.

252. Weake, V.M., and Workman, J.L. (2008). Histone Ubiquitination: Triggering Gene Activity. Mol. Cell *29*, 653–663.

253. Weake, V.M., Lee, K.K., Guelman, S., Lin, C.-H., Seidel, C., Abmayr, S.M., and Workman, J.L. (2008). SAGA-mediated H2B deubiquitination controls the development of neuronal connectivity in the Drosophila visual system. EMBO J. *27*, 394–405.

254. Weake, V.M., Dyer, J.O., Seidel, C., Box, A., Swanson, S.K., Peak, A., Florens, L., Washburn, M.P., Abmayr, S.M., and Workman, J.L. (2011). Post-transcription initiation function of the ubiquitous SAGA complex in tissue-specific gene activation. Genes Dev. *25*, 1499–1509.

255. Wery, M., Shematorova, E., Van Driessche, B., Vandenhaute, J., Thuriaux, P., and Van Mullem, V. (2004). Members of the SAGA and Mediator complexes are partners of the transcription elongation factor TFIIS. EMBO J. *23*, 4232–4242.

256. Whitehouse, I., Flaus, A., Cairns, B.R., White, M.F., Workman, J.L., and Owen-Hughes, T. (1999). Nucleosome mobilization catalysed by the yeast SWI/SNF complex. Nature *400*, 784–787.

257. Wieczorek, E., Brand, M., Jacq, X., and Tora, L. (1998). Function of TAFII-containing complex without TBP in transcription by RNA polymerase II. Nature *393*, 187–191.

258. Williams, D.B., and Carter, C.B. (2009). Transmission electron microscopy a textbook for materials science (New York: Springer).

259. Wilson, M.A., Koutelou, E., Hirsch, C., Akdemir, K., Schibler, A., Barton, M.C.,

and Dent, S.Y.R. (2011). Ubp8 and SAGA Regulate Snf1 AMP Kinase Activity ▿. Mol. Cell. Biol. *31*, 3126–3135.

260. Winston, F., Durbin, K.J., and Fink, G.R. (1984). The SPT3 gene is required for normal transcription of Ty elements in S. cerevisiae. Cell *39*, 675–682.

261. Winston, F., Dollard, C., Malone, E.A., Clare, J., Kapakos, J.G., Farabaugh, P., and Minehart, P.L. (1987). Three genes are required for trans-activation of Ty transcription in yeast. Genetics *115*, 649–656.

262. Wong, W., Bai, X., Brown, A., Fernandez, I.S., Hanssen, E., Condron, M., Tan, Y.H., Baum, J., and Scheres, S.H. (2014). Cryo-EM structure of the Plasmodium falciparum 80S ribosome bound to the anti-protozoan drug emetine. eLife *3*, e03080.

263. Wu, P.-Y.J., and Winston, F. (2002). Analysis of Spt7 function in the Saccharomyces cerevisiae SAGA coactivator complex. Mol. Cell. Biol. *22*, 5367–5379.

264. Wu, P.-Y.J.Y., Ruhlmann, C., Winston, F., and Schultz, P. (2004). Molecular architecture of the S. cerevisiae SAGA complex. Mol. Cell *15*, 199–208.

265. Wu, Y., Reece, R.J., and Ptashne, M. (1996). Quantitation of putative activator-target affinities predicts transcriptional activating potentials. EMBO J. *15*, 3951–3963.

266. Wyce, A., Henry, K.W., and Berger, S.L. (2004). H2B ubiquitylation and de-ubiquitylation in gene activation. Novartis Found. Symp. *259*, 63–73; discussion 73–77, 163–169.

267. Wyce, A., Xiao, T., Whelan, K.A., Kosman, C., Walter, W., Eick, D., Hughes, T.R., Krogan, N.J., Strahl, B.D., and Berger, S.L. (2007). H2B ubiquitylation acts as a barrier to Ctk1 nucleosomal recruitment prior to removal by Ubp8 within a SAGA-related complex. Mol. Cell *27*, 275–288.

268. Xie, X., Kokubo, T., Cohen, S.L., Mirza, U.A., Hoffmann, A., Chait, B.T., Roeder, R.G., Nakatani, Y., and Burley, S.K. (1996). Structural similarity between TAFs and the heterotetrameric core of the histone octamer. Nature *380*, 316–322.

269. Xu, W., Edmondson, D., and Roth, S. (1998). Mammalian GCN5 and P/CAF acetyltransferases have homologous amino-terminal domains important for recognition of nucleosomal substrates. Mol. Cell. Biol. *18*, 5659–5669.

270. Xu, W., Edmondson, D.G., Evrard, Y.A., Wakamiya, M., Behringer, R.R., and Roth, S.Y. (2000). Loss of Gcn5l2 leads to increased apoptosis and mesodermal defects during mouse development. Nat. Genet. *26*, 229–232.

271. Yamauchi, T., Yamauchi, J., Kuwata, T., Tamura, T., Yamashita, T., Bae, N., Westphal, H., Ozato, K., and Nakatani, Y. (2000). Distinct but overlapping roles of histone acetylase PCAF and of the closely related PCAF-B/GCN5 in mouse embryogenesis. Proc. Natl. Acad. Sci. U. S. A. *97*, 11303–11306.

272. Yan, Z., Bai, X., Yan, C., Wu, J., Li, Z., Xie, T., Peng, W., Yin, C., Li, X., Scheres, S.H.W., et al. (2015). Structure of the rabbit ryanodine receptor RyR1 at near-atomic resolution. Nature *517*, 50–55.

273. Yang, X., Ogryzko, V., Nishikawa, J., Howard, B., and Nakatani, Y. (1996). A p300/CBP-associated factor that competes with the adenoviral oncoprotein E1A. Nature *382*, 319–324.

274. Yonekura, K., Braunfeld, M.B., Maki-Yonekura, S., and Agard, D.A. (2006). Electron energy filtering significantly improves amplitude contrast of frozen-hydrated protein at 300 kV. J. Struct. Biol. *156*, 524–536.

275. Yu, Y., Teng, Y., Liu, H., Reed, S.H., and Waters, R. (2005). UV irradiation stimulates histone acetylation and chromatin remodeling at a repressed yeast locus. Proc. Natl. Acad. Sci. U. S. A. *102*, 8650–8655.

276. Yudkovsky, N., Logie, C., Hahn, S., and Peterson, C.L. (1999). Recruitment of the SWI/SNF chromatin remodeling complex by transcriptional activators. Genes Dev. *13*, 2369–2374.

277. Zenke, F.T., Engles, R., Vollenbroich, V., Meyer, J., Hollenberg, C.P., and Breunig, K.D. (1996). Activation of Gal4p by galactose-dependent interaction of galactokinase and Gal80p. Science *272*, 1662–1665.

278. Zhang, X.-Y., Varthi, M., Sykes, S.M., Phillips, C., Warzecha, C., Zhu, W., Wyce, A., Thorne, A.W., Berger, S.L., and McMahon, S.B. (2008). The Putative Cancer Stem Cell Marker USP22 Is a Subunit of the Human SAGA Complex Required for Activated Transcription and Cell-Cycle Progression. Mol. Cell *29*, 102–111.

279. Zhao, Y., Lang, G., Ito, S., Bonnet, J., Metzger, E., Sawatsubashi, S., Suzuki, E., Le Guezennec, X., Stunnenberg, H.G., Krasnov, A., et al. (2008). A TFTC/STAGA Module Mediates Histone H2A and H2B Deubiquitination, Coactivates Nuclear Receptors, and Counteracts Heterochromatin Silencing. Mol. Cell *29*, 92–101.

Chapter 7. Annexes.

7.1. Protocol for recombinant human SAGA HAT purification.

7.1.1. Buffers.

Buffers for IMAC	0.2 M $NiSO_4$, 0.1 M EDTA
Buffer A (lysis, washing)	20 mM HEPES-KOH pH 8, 500 mM NaCl, 10% glycerol, 5 mM $MgCl_2$
Buffer B (elution)	20 mM HEPES-KOH pH 8, 500 mM NaCl, 10% glycerol, 250 mM Imidazole
Buffer C (gel filtration)	20 mM HEPES-KOH pH 8, 500 mM NaCl, 10% glycerol, 1 mM DTT, 0.1 mM EDTA

Table 16. List of buffers used for SAGA HAT purification.

	Buffer A	**Buffer B**	**Buffer C**
HEPES-KOH 1M pH 8	10 ml	10 ml	20 ml
NaCl 5M	50 ml	50 ml	100 ml
Glycerol 100%	63 g	63 g	126 g
Imidazole 1M	-	125 ml	-
$MgCl_2$ 1M	2.5 ml	-	-
EDTA 0.5 M	-	-	200 µl
DTT 1M	-	-	1 ml
Adjust pH to 8 at 4°C			
H_2O Milli-Q up to	500 ml	500 ml	1000 ml

Table 17. Preparation of purification buffers.

7.1.2. Harvesting insect cell culture 72h after infection.

1. Centrifuge cells for 20 min at 400 g and discard the supernatant (media).
2. Wash the pellets with 25 ml of PBS 1X with slow agitation (try not to use pipettes, cells are very fragile), then transfer the solution in Falcons 50ml.
3. Centrifuge for 20 min at 400 g, discard the supernatant.
4. Freeze pellets in liquid nitrogen. You will have about 5-7 ml of yellow pellet (due to YFP) with thin gray layer on the top.

7.1.3. Cell lysis.

1. Thaw the cell pellets on ice and add buffer A up to 50 ml, also put 1 tablet of PIC (complete, EDTA-free, Roche) and 500 µl of 100 mM PMSF (final 1 mM), 7 µl of 14 M β-mercaptoethanol (final 2 mM).
2. Lyse the cells by sonication on CV33 (big tip, amplitude 40%, 6 min total, 1 s ON / 2 s OFF, Tmax 15°C) on ice, take aliquot 15 µl for the gel (SNP).
3. Add 200 µl of DNAse I (40U; stock is 400U/mg = 5 g/l), incubate for 30 min at 4°C on a rotating shaker.
4. Centrifuge extracts for 60 min at 50.000 g, at 4°C.
5. Filter the supernatant with 0.2 or 0.5 µm filter, take 15 µl aliquot for the gel (SN).

7.1.4. IMAC protocol for ÄKTA purifier system.

1. Load the HiTrap Chelating HP column with nickel at 3 ml/min: 3 CV with H_2O, 3 CV with $NiSO_4$, 3 CV with H_2O again.
2. Equilibrate the column with buffer A (5 CV), 3ml/min.
3. Load the sample at 3ml/min in buffer A.
4. Wash (also collect flow-through [FT]) and run a step-by-step gradient elution in 96-well plate (2 ml per fraction) according to the table.

Imidazole concentration	% of the buffer B	CV
0 mM	0 %	10
10 mM	4 %	10
30 mM	12 %	12
60 mM	24 %	7
100 mM	40 %	7
250 mM	100 %	10

Table 18. Elution with a step-by-step imidazole gradient.

5. After the end of elution, wash the column with EDTA (3 ml/min; 3 CV with H_2O, 3 CV with EDTA, 3 CV with H_2O again.
6. Wash the column with 20% ethanol (2 ml/min, 5 CV), and disconnect the column.
7. Load the fractions (5 µl for SNP, SN, FT; 10 µl for elutions) on a 10% SDS-PAGE gel.
8. Measure the concentration of all aliquots by Bradford in 3 replicas: 1 ml Bradford 1x + 0.5-2 µl sample.
9. Choose the best fractions containing your subcomplex (4 fractions max) for gel filtration. Pool them and check OD, take 15 µl aliquot GF INP.

7.1.5. Gel filtration protocol for ÄKTA purifier system.

1. Equilibrate the Superdex 200 16/60 GL column for 1.2 CV with increasing concentration of the buffer B.
2. Load 4 ml of pooled sample from the previous step into the loop.
3. Manually inject your sample and empty the loop with 5 ml.
4. Run the elution at 1 ml/min for 1.2 CV in 96-well plates.
5. At the end, run the equilibration with 20% ethanol.
6. Load GF INP and 10 µl of each elution fraction on the 10% SDS-PAGE gel.

Grigory SHAROV

Étude structurale du co-activateur transcriptionnel SAGA et de son module d'acétylation des histones

Résumé

L'initiation de la transcription chez les eucaryotes nécessite le recrutement de l'ARN polymérase II (Pol II) et des facteurs de transcription généraux sur les promoteurs de gènes formant le complexe de pré-initiation (PIC). Des activateurs se lient en amont du promoteur et stimulent l'ouverture de la chromatine et la formation du PIC en recrutant des complexes co-activateurs. SAGA est un tel co-activateur, conservé chez les eucaryotes, connu pour modifier les histones de tous les gènes et impliqué dans la transcription par Pol II. Dans ce travail, j'ai analysé l'organisation moléculaire de SAGA par microscopie électronique. J'ai (i) étudié l'architecture et les interactions des sous-unités du module d'acétylation des histones et l'ai localisé dans SAGA; (ii) obtenu la première carte cryo-EM du complexe SAGA chez la levure et analysé sa flexibilité; (iii) défini le site d'interaction entre TBP et SAGA et montré que le complexe subit un changement conformationnel lors de cette liaison.

Mots-clés : co-activateur de transcription, SAGA, microscopie électronique, biologie structurale, acétylation des histones

Résumé en anglais

Transcription initiation in eukaryotes requires the recruitment of RNA polymerase II (Pol II) and general transcription factors to the promoters of protein coding genes in order to form a Pre-Initiation Complex (PIC). Sequence-specific activators bind upstream of the promoter, stimulating chromatin opening and PIC formation via recruitment of coactivator complexes. SAGA is such a coactivator, conserved in all eukaryotes, known to modify the histones on all expressed genes in yeast and human and involved in Pol II transcription. In this work I have analyzed SAGA's molecular organization mostly by electron microscopy. I have (i) studied the architecture and subunit interactions of SAGA histone acetylation (HAT) module and localized it in the full SAGA complex; (ii) obtained the first cryo-EM map of yeast SAGA and analyzed its flexibility; (iii) defined the interaction site of SAGA with TBP protein and shown that the complex undergoes a large conformational change upon TBP binding.

Keywords : transcriptional coactivator, SAGA, electron microscopy, structural biology, histone acetylation