Universitat Autònoma de Barcelona

STRUCTURAL CHARACTERIZATION OF A MITOCHONDRIAL DNA MAINTENANCE PROTEIN OF BIOMEDICAL INTEREST

ALEIX TARRÉS SOLÉ

Universitat Autònoma de Barcelona

FACULTAT DE BIOCIÈNCIES

DOCTORAT EN BIOQUÍMICA BIOLOGIA MOLECULAR I BIOMEDICINA

STRUCTURAL CHARACTERIZATION OF A MITOCHONDRIAL DNA MAINTENANCE PROTEIN OF BIOMEDICAL INTEREST

Analysis of protein binding, protein dynamics in solution and DNA compacting properties of a mitochondrial DNA maintenance factor from 'Candida albicans'.

ibmb
Institut de Biologia Molecular de Barcelona
CSIC

SBU
Structural Biology Unit
CSIC

ALEIX TARRÉS SOLÉ, 2020

DIRECTORA DE TESI: MARIA SOLÀ I VILARRUBIAS

TUTORA DE TESI: SANDRA VILLEGAS HERNÁNDEZ

Universitat Autònoma de Barcelona

FACULTAT DE BIOCIÈNCIES

DOCTORAT EN BIOQUÍMICA BIOLOGIA MOLECULAR I BIOMEDICINA

STRUCTURAL CHARACTERIZATION OF MITOCHONDRIAL DNA MAINTENANCE FACTOR GCF1P

Aleix Tarrés Solé

Dra. Maria Solà i Vilarrubias

Dra. Sandra Villegas Hernández

Memòria presentada per Aleix Tarrés Solé per optar al títol de doctor, amb menció de doctorat internacional, per la Universitat Autònoma de Barcelona dins del programa de doctorat en Bioquímica, Biologia molecular i Biomedicina de la Facultat de Biociències. El treball experimental contingut en aquest llibre s'ha realitzat majoritàriament en l'Institut de Biologia Molecular de Barcelona (IBMB-CSIC) dins de la unitat de biologia estructural (IBMB-CSIC). A més del treball realitzat a l'IBMB-CSIC s'han realitzat dues estades sumant un total de tres mesos i tres setmanes en l'Institut Gustave Roussy (Villejuif, França) sota la supervisió del Dr. Olivier Pietrement i el Dr. Éric le Cam. La tesi doctoral s'ha realitzat en el marc de l'ajuda a la contractació de personal universitari FPU/06021. **ALEIX TARRÉS SOLÉ, 2020**

ibmb
Institut de Biologia Molecular de Barcelona
CSIC

SBU
Structural Biology Unit
CSIC

“Everyone knows that in research there are no final answers, only insights that allow one to formulate new questions”

“Significant advances in science often have a peculiarity: they contradict obvious, commonsense opinions.”

“What matters in science is the body of findings and generalizations available today: a time-defined cross-section of the process of scientific discovery. I see the advance of science as self-erasing in the sense that only those elements that have become part of the active body of knowledge survive.”

SALVADOR LURIA

A tothom que em coneix i a tothom que és i ha sigut part d’aquest viatge

Acknowledgements

De totes les persones que poden estar llegint aquesta tesi, tinc molt clar que hi ha dos tipus molt definits. Les que només llegiran l'apartat d'agraïments i les que llegiran la resta de pàgines. Ja que les últimes tenen unes 200 pàgines per davant, crec que és just començar per les primeres.

Vull començar agraint a la meva família, per ser-hi. Al meu pare, el Sisco, que ha tingut un any complicat però és un màquina i se n'ha sortit de tot amb nota, com ha fet sempre. A la meva, la Mercè per tenir la paciència d'aguantar-nos a tots. A la meva germana, l'Alba per ser-hi sempre i per poder comptar amb ella pel que sigui tot i la distància. A ella i al David els hi desitjo tota la sort del món en el futur. Seguir parlant de família es parlar dels cerveceros, tots ells estaran contents de que hagi defensat la tesi finalment. Està lleig destacar però no puedo menos que parlar del Ruben, que segurament és el que més m'ha aguantat les queixes i moments baixos d'aquests anys. De l'Ivan, que no tinc cap dubte que seguirà sent-hi com fins ara, el pavo, i amb el qual tinc ganes de prendre uns cachopos quan això acabi. Per últim i no per això menys important, l'Hernan que m'ha obert les portes de casa aquests últims dies i està veient com finalment estic acabant la tesi. I també a la Cèlia que sap que encara que m'en vagi seguirem fent cafès per Skype. Tota la resta del grup, sap que també els estic immensament agraït per tot el que m'han aportat aquests anys. Li vull agrair també a Olaya, per tot el que m'ha aportat aquests últims mesos, la paciència, la confiança i la certesa de que posant-li ganes i força podem arribar a fer tot el que ens plantegem. Tant ella com Celia m'han fet sentir com a casa durant aquests mesos de confinament. Finalment vull també agrair a la gent del grau de bioquímica (Cris, Vicente, Laura, Marta, Manu, Guillem i Clara) el fer aquests estius molt més agradables encara, aviat serem uns quants doctors més al mon!!.

Vull agrair a tota la gent de Cri-7 començant per la Maria Solà la meva supervisora, qui m'ha permès entrar en aquest projecte i realitzar-lo amb tota la llibertat i la confiança de que al final treuríem resultats rellevants. Vull seguir agraint a la gent que ha estat directament implicada en el projecte de Gcf1p en diferents punts i que ben aviat rebran noticies de que les coses finalment han arribat a bon port. He de començar amb Neus, qui va aportar moltíssima ajuda dins del laboratori realitzant els experiments d'unió a l'ADN i que fora del laboratori em recordava que som capaços de fer tot allò que ens apassiona, per difícil que sembli, si deixem que ens neixi de dins. També de Nico, que va ajudar en la obtenció de mutants amb el seu treball de màster i que em recorda que no tota la ciència es fa des de l'acadèmia. Por último Sabina, ella ha estado en la última parte de la tesis, ha visto lo más estresante de la investigación però también la más gratificante y estoy seguro que en el futuro grandes aportacions a nuestro campo. Siguiendo con compañeros implicados en este trabajo quiero agradecer a Javier y a Elena, los dos técnicos de laboratorio con quienes he tenido el placer de trabajar. Ellos han llevado la carga de la producción de Gcf1p cuando estaba produciendo cristales y sin ellos esta tesis tampoco hubiera sido possible. A Javi le deseo que él, Joana y los ninos encuentren su lugar en Alemania y a Elena, Chus y Aitana todo lo mejor que les puedan traer estos años. I want to thank to my co-supervisor in the first years of my thesis Sébastien who taught me the basics of the experimental work in the bench and continuously reminded me that experiments are do only way to get somewhere trustable in science. I hope all the best for him in the next years. Same applies for Arka, which with whom I am proud to say I could work with and share some of the most interesting talks during my years of PhD. I am sure he will continue contributing greatly to science with his unnegotiable personal style. También agradezco Cuppi por todo lo que me ha enseñado, por toda la alegría que me ha transmitido en el lab y en los viajes a Grenoble, por su apoyo en los momentos de duda y por ser algo así como una hermana mayor dentro del laboratorio. I want to thank Claus for always enriching our conversations and debates with his vision of the things. He is also finishing the thesis and I know he will present and publish some outstanding science very soon. I want to thank my Arkadiusz for all this year of craziness and jokes, and for being a bro inside and outside the lab. Por último, el joven padawan Ramiro, el último en incorporarse al laboratorio con quien es un gran placer charlar de lo que sea (dentro o fuera del laboratorio) y que además es un gran currante del laboratorio desde el dia 1 en que empezó la tesis (y por lo que sé antes también). Le deseo toda la fuerza y la suerte del mundo, aunque va a necesitar poca, y le agradezco todas las veces que ha bajado conmigo a fumar, aunque fuera solo para pedirme tabaco. Quiero agradecer también a toda la gente de Cri-6, en especial a los compis de pre-doc Víctor, María y Omar, así como a Mónica por todas las pausas de media mañana. Quiero agradecer a Diego todos los 'tips' recibidos.

También a la gente de Cri 3, en especial a Tibi por tener siempre una sonrisa y llamarme doctor Tarrés desde hace cuatro años. También a Arturo por la ayuda y por los viajes al sincrotrón. And also I want to thank Ulrich, although he incorporated recently I could share useful scientific talks with him and together with Arturo they are just the perfect comic duo. Vull agrair també al Xavi Gomis per totes les converses sobre cristal·lografia que finalmente han ajudat a resoldre l'estructura.

A la gente de Cri 4, en especial Iracema, por todas las birras y los cotilleos. A Nico, Rafa y Claudia por tener paciencia explicando y ser fuentes de conocimiento inagotables, y sobretodo Massimo, quien además se implicó en el proceso de resolución de la estructura. Quiero agradecer también a Isabel Usón por todos los conocimientos de cristalografía que me ha transmitido así como por lo bien acogido que me he sentido siempre que he estado con la gente de su grupo.

A la gent del laboratorio de Joaquim Roca, a Joana, Ofelia, Toni i Sílvia per lo bé que m'ho he passat sempre que els he pujat a veure, o a fer servir el Nanodrop. A Belén especialmente por su buen rollo y por realizar los ensayos topológicos que no están incluídos en este libro pero que nos han ayudado a entender como el ADN es compactado por Gcf1p. Per últim a Joaquim Roca per totes les llargues discusión sobre Gcf1p.

No puc posar tota la resta de gent de l'hèlix a la que estic agrait per que sino això s'allargarà més que la tesi en sí, pero vull agrair especialment a tota la gent que estaba al PhD committee per la paciencia amb la que han aguantat les xapes sindicals. Especialment la gent dels labos de Martí Aldea i Carme Gallego.

Respecte gent del cluster vull agrair a Joana i Ekaitz el quedar-se fins tard amb mi a la línea per ensenyar-me que és això de la difracció i pels moments ultra-divertits en els primers viatges a Grenoble. Vull agrair de nou al Manu per les short talks camí de la plataforma de purificació. A la gent de la PAC: Román, Xandra, Laura i Joan per tota l'ajuda, la implicació i la feina ben fet. També a la gent de Cri 1 i Cri 2, especialmente al David Aparicio i l'Ignasi Fita per tota l'ajuda oferida alhora de resoldre l'estructura cristal·logràfica.

A la gent de XALOC: Fernando, Barbara, Roeland i Xavi per tota l'ajuda rebuda a la línea cada vegada que la he visitat.

I want to thank all the people from IGR (Pauline, Yann, Ali) but especially to Sonia, Olivier and Éric, who made me comfortable during my stays in Paris and were of great help in providing an experimental point of view outside crystallography.

També vull agrair al Pau Bernadó per les classses de processat de dades de SAXS així com tots els skypes fets durant el confinement que tant ens han ajudat a treure tot el suc possible a les nostres dades.

Finally, I also want to thank life all the nice people I could meet attending to different courses around the world and with whom I could share short albeit intense moments. Regarding this last point I want to express my tremendous gratitude to the organizers and participants of the last CCP4-DLS course, where I could finally solve my structure thanks to the lessons and the immense amount of helps provided by the tutors. I want to thank especially to Airlie McCoy, Misha Isupov and Clement Vornheim. Also to Prof Rany Read form having me a couple of days in Cambridge.

Table of contents

GLOSSARY OF TERMS

List of nitrogenous bases

A (Ade): Adenine

C (Cyt): Cytosine

G (Gua): Guanine

T (Thy): Thymine

U (Ura): Uracyl

Proteinogenic L-aminoacids list

A (Ala): Alanine

C (Cys): Cysteine

D (Asp): Aspartate

E (Glu): Glutamate

F (Phe): Phenylalanine

G (Gly): Glycine

H (His): Histidine

I (Ile): Isoleucine

K (Lys): Lysine

L (Leu): Leucine

M (Met): Methionine

N (Asn): Asparagine

P (Pro): Proline

Q (Gln): Glutamine

R (Arg): Arginine

S (Ser): Serine

SeMet: Seleno-Methionine

T (Thr): Threonine

V (Val): Valine

W (Trp): Tryptophan

Y (Tyr): Tyrosine

Other macromolecules, biological and experimental terms

a.u.: Asymmetric unit

Abf2p: ARS-1 binding factor 2

ADP: Adenosine Di-Phosphate

AEC: Anionic Exchange Chromatography

ARS-1: Autonomously Replicating Sequence 1

ATP: Adenosine Tri-Phosphate

CEC: Cationic Exchange Chromatography

DLS: Dynamic Light Scattering

DNA: Deoxyribonucleic Acid

EMSA: Electrophoretic Mobility Shift Assay

FAD^+/$FADH_2$: Oxidized and Reduced Flavin-Adenine Dinucleotide

GSSG/GSH: Oxidized and Reduced Glutathione

GST: Glutathione S-transferase

HMG: High Mobility Group

MAD: Multiple-wavelength Anomalous Diffraction

MALLS: Multiple Angle Laser Light Scattering

MR: Molecular Replacement

MW: Molecular Weight

MX: Macromolecular crystallography

pI: Isoelectric point

pH: *pondus hydrogenii* (latin for quantity of hydrogens, the universal indicator of acidity in aqueous solution)

$NAD(P)^+$/NAD(P)H: Oxidized and Reduced Nicotinamide Dinucleotide (Phosphate)

R_g: Radius of gyration

RNA: Ribonucleic Acid

SAXS: Small-Angle X-ray Scattering

SAD: Single-wavelength Anomalous Diffraction

SEC: Size-Exclusion Chromatography

TFAM: Mitochondrial Transcription Factor A

Culture media and bacterial strains

BL21 strain: Standard *Escherichia coli* strain for protein expression. It is defective in LON protease for enhanced expression yields.

DE3 strains: DE3 strains are *Escherichia coli* strains infected by λDE3 phage lysogen. Hence the lysogen, DE3 strains possess in its genome the RNA polymerase gene under the control of an *Escherichia coli* endogenous *lac* promoter. T7 polymerase shows an increase expression yield as compared to the endogenous *Escherichia coli* RNA polymerase. When a strain is positive for the λDE3 lysogen, it is indicated following the original name of the strain *e.g.* BL21(DE3), Rosetta 2 (DE3) etc.

DH5α strain: Escherichia coli strains defective in the *RecA* gene encoding for recombinase A. Hence this deletion this strain shows an enhanced genome stability and it is the strain of choice for plasmid amplification.

Escherichia coli: Escherichia coli or *E. coli* is a gram-negative bacterium naturally present as a part of the microbiota from the gastrointestinal tract of homeothermic animals. It is also a profusely used system for heterologous expression of proteins.

LB: Lysogeny Broth or LB, developed by Giuseppe Bertani and Salvador E Luria, medium for the growing of bacteria. It is profusely used as the media of choice for plasmid amplification and protein expression. Its composition is 10g of peptone, 5g of yeast extract and 5 g of NaCl per 1L of media.

Origami^{TM} strains: K-12 derivative *E. coli* strains developed by Novagen ®. This strains contain mutations in the *trxB* and *gor* genes encoding for thioredoxin reductase and glutathione reductase respectively, which favours the proper formation of disulphide bonds within *E.coli* cytoplasm. Mutations are selected by Streptomycin and Tetracycline resistance.

plysS plasmid: A chloramphenicol resistant plasmid carrying T7 lysozyme gene. T7 lysozyme happens to be a strong inhibitor for the T7 RNA polymerase. Therefore, *plysS* is profusely used in DE3 cells in order to avoid expression leakage.

Rosetta^{TM} 2 strains: Modified BL21 strains developed by Novagen ® for the expression of eukaryotic genes containing codon rarely used in *E. coli*. Rosetta^{TM} 2 strains are transformed with a compatible chloramphenicol-resistant plasmid carrying 7 tRNA genes for codons AGA, AGG, AUA, CUA, GGA, CCC and CGG.

TB: Terrific Broth or TB media for the growing of bacteria. It is used to enhance expression yields and to achieve later induction due to the presence of a pH buffer system. Due to its increased nutrient composition its use may result in promoter leakage as an undesired side effect. Its composition is 20g of tryptone, 24g of yeast extract and 4% glycerol per 900mL of media, complemented with 100mL of autoclaved 0.017M KH_2PO_4, 0.072M K_2HPO_4 buffering solution.

Buffers and reagents of common usage:

BME: 2-beta-mercaptoethanol, a volatile reducing agent used for the disruption of both intra and intermolecular disulphide bonds.

DTT: Dithiothreitol a volatile reducing agent used for the disruption of both intra and intermolecular disulphide bonds.

LSB: Laemmli Sample Buffer, named after Ulrich K Laemmli, for the preparation of samples for denaturing gel electrophoresis. LSB 1X composition is 1%SDS, 10mM BME, 10% Glycerol, 50mM Tris pH6.8 and 500µg/mL Bromophenol Blue.

PMSF: PhenylMethylSulphonyl Fluoride, is an irreversible serin-protease inhibitor commonly used to neutralize proteases naturally present in a cell lysate.

SDS: Sodium dodecyl sulphate

TAE: Tris-Acetate-EDTA running buffer for native gel electrophoresis of protein or nucleic-acids both in agarose and polyacrylamide gels. TAE 1X buffer composition is 40mM Tris-Acetate pH8.2, 1mM EDTA.

TBE: Tris-Borate-EDTA running buffer for native of protein or nucleic-acids both in agarose and polyacrylamide gels. TBE 1X buffer composition is 89mM Tris-Borate, 89mM Boric Acid pH8.3, 1mM EDTA. Borate is an inhibitor for many enzymes, such as DNA ligase. Hence to this, TAE should always be used for molecular biology purposes.

TCEP: Tris-(2-carboxyethyl)-phosphine, a non-volatile reducing agent used for the disruption of both intra and intermolecular disulphide bonds. It shows an enhanced stability as compared to DTT and BME.

TEV protease: Tobacco Etch Virus protease, is a serine-protease that shows high specificity for its target ENLYFQ GS. It is a common prote↓ase for the specific removal of affinity tags.

ABSTRACT

Resum de la tesi en català

En aquest treball de tesi presentem la caracterització estructural de la proteïna de *Candida albicans* (*C.albicans*) Gcf1p. Gcf1p és una proteïna d'unió a l'ADN, que localitza al mitocondri de *C.albicans* i que és essencial per al manteniment de l'ADN mitocondrial (ADNmt) així com per a la supervivència del microorganisme. *C.albicans* és un fong dimòrfic amb capacitat de créixer en hifes invasives i causar patologia en humans. *C.albicans* forma part de la flora microbiana en individus sans, no obstant, pot esdevenir patògen oportunista causant de infeccions superficials (candidiasi) així com d'infeccions invasives (candidèmia). El creixent nombre de casos de candidèmia, derivats principalment d'infeccions d'origen nosocomial en hospitals, així com la seva elevada mortalitat associada , al voltant del 50%, fa de *C.albicans* una microorganisme d'interès biomèdic. Per altra banda, l'adquisició de resistència als tractaments antifúngics convencionals per part de *C.albicans* i d'espècies properes (*Candida auris*) fa urgent la necessitat d'una descripció dels mecanismes replicatius i invasius d'aquestes espècies. En aquesta línia, una millor descripció dels mecanismes fonamentals en el manteniment de l'ADNmt és un potencial punt de partida per a la recerca de nous fàrmacs específics. Aquest treball de tesi aporta informació novedosa sobre com la proteïna de *C.albicans* Gcf1p uneix i compacta l'ADNmt.

Per altra banda, aquest treball de tesi aporta informació novedosa que pot ésser d’interès en el camp de la biologia evolutiva. L'origen i evolució dels mitocondris des de organismes independents a orgànuls cel·lulars (teoria endosimbiòtica) és acceptada per la comunitat científica com a punt clau en l'aparició d'organismes eucariotes (protists, fongs, plantes i animals). Els mecanismes de compactació i replicació de l'ADNmt presenta alta variabilitat inter i intra-regnes. La proteïna Gcf1p està relacionada amb la replicació depenent de recombinació que està present en l'ADNmt *de C.albicans* i que és completament diferent al model de replicació de l'ADNmt en mamífers. La informació aportada per aquest treball respecte a la unió i compactació de l'ADN per part de Gcf1p obre un punt de partida per a comparar la variabilitat en la organització de l'ADNmt de *C.albicans* amb la d'altres llevats (principalment *Saccharomyces cerevisiae* (*S.cerevisiae*) així com les diferències amb organismes llunyans en l'evolució (*Homo sapiens* (*H.sapiens*)). Aquest treball és també un potencial punt de partida per a entendre millor la divergència evolutiva observada en eucariotes.

En aquest treball s'han usat tècniques de biologia molecular per al clonatge en vectors plasmídics i l'expressió heteròloga de proteïnes de llevat en *Escherichia coli*. Així mateix, s'han usat tècniques de purificació de proteïnes i s'ha aconseguit optimitzar un protocol de producció compatible en termes de rendiment i puresa amb l'anàlisi per mètodes de biologia estructural, principalment cristal·lografia de macromolècules (MX), dispersió de rajos X a angle petit (SAXS) i microscopia electrònica (EM). Els resultats obtinguts per aquestes tres tècniques aporten observacions complementaries sobre la interacció de Gcf1p amb l'ADN.

En suma, aquest treball de tesi aporta evidències sòlides que indiquen que el mecanisme de reconeixement de l'ADNmt en *C.albicans* presenta diferències fonamentals amb els descrits per *S.cerevisiae* i *H.sapiens*. Els nostres resultats suposen un important avenç en la descripció d'un procès essencial per als organismes eucariotes com és la organització i manteniment de l'ADN mitocondrial.

Resumen de la tesis en castellano

En este trabajo de tesis se presenta la caracterización estructural de la proteína de *Candida albicans* (*C.albicans*) Gcf1p. Gcf1p es una proteína de unión al ADN, que transloca a la mitocondria de *C.albicans* y que es esencial para el mantenimiento del ADN mitocondrial (ADNmt) así como para la supervivencia del organismo. *C.albicans* es un hongo dimórfico con capacidad de formar hifas invasivas y causar patología en humanos. *C.albicans* forma parte de la flora microbiana en individuos sanos, no obstante, puede devenir patógeno oportunista causante de infecciones superficiales (candidiasis) así como de infecciones invasivas (candidémia). El crecimiento del número de casos de candidémia, derivados principalmente de infecciones nosocomiales, así como la elevada tasa de mortalidad asociada (alrededor del 50 %), hace de *C.albicans* un organismo de elevado interés biomédico. Por otro lado, la adquisición de resistencia a los tratamientos antifúngicos más comunes por parte de *C.albicans* y especies relacionadas (*Candida auris*) hace urgente la necesidad de una descripción de los mecanismos replicativos e invasivos de estas especies. En la misma línea, una mejor descripción de los mecanismos fundamentales en el mantenimiento del ADNmt es un potencial punto de partida para la investigación de nuevos fármacos específicos. En este trabajo de tesis se aporta información novedosa sobre como la proteína de *C.albicans* Gcfp une y compacta el ADNmt.

Por otro lado, este trabajo de tesis aporta información novedosa que puede ser de interés en el campo de la biología evolutiva. El origen y evolución de las mitocondrias desde organismos independientes a orgánulos celulares (teoría endosimbiótica) es aceptada por la comunidad científica como punto de partida en la aparición de organismos eucariotas (protistas, hongos, plantas y animales). Los mecanismos de compactación y replicación del ADNmt presentan alta variabilidad inter e intra-reinos. La proteína Gcf1p está relacionada con la replicación dependiente de recombinación presente en *C.albicans* y que es completamente diferente al modelo de replicación del ADNmt en mamíferos. La información aportada en este trabajo de tesis acerca de la unión y compactación del ADN por parte de Gcf1p supone un punto de partida para comparar la variabilidad en la organización del ADNmt en *C.albicans* respecto otras levaduras (principalmente *Saccharomyces cerevisiae* (*S.cerevisiae*) así como respecto de otros organismos evolutivamente distantes (*Homo sapiens* (*H.sapiens*)). Asimismo, este trabajo supone también un potencial punto de partida para entender la divergencia evolutiva observada entre eucariotas.

En este trabajo se han usado técnicas de biología molecular para el clonaje en vectores plasmídicos y la expresión heteróloga de proteínas de levadura en *Escherichia coli*. Asimismo, se han usado técnicas de purificación de proteínas logrando optimizar un protocolo de producción compatible en términos de rendimiento y pureza con el análisis por métodos de biología estructural, principalmente cristalografía de macromoléculas (MX), dispersión de rayos X en ángulo pequeño (SAXS) y microscopia electrónica (EM). Los resultados obtenidos por estas tres técnicas aportan observaciones complementarias sobre la interacción de Gcf1p con el ADN.

En su conjunto, este trabajo de tesis aporta evidencias que indican que el mecanismo de reconocimiento del ADNmt en *C.albicans* presenta diferencias fundamentales respecto de los descritos para *S.cerevisiae* y *H.sapiens*. Nuestros resultados suponen un importante avance en la descripción de un proceso esencial para los organismos eucariotas como es la organización y mantenimiento del ADN mitocondrial.

Abstract in English

This thesis work is centred in the structural characterisation of *Candida albicans* (*C.albicans*) Gcf1p. Gcf1p is a DNA-binding protein located in *C.albicans* mitochondria that is essential for mitochondrial DNA (mtDNA) maintenance as well as for the viability of such organism. *C.albicans* is a dimorphic yeast with the capability to form invasive hyphal structures causing pathology in humans. *C.albicans* is a part of the mycobiota in healthy individuals. Nevertheless, it can be causing both superficial infections (candidiasis) as well as invasive infections (candidemia). Growing prevalence of candidemia, mainly caused by nosocomial transmission in hospital, together with the its high lethality (about 50% mortality rate) makes *C.albicans* a potential biomedical target of high interest. In addition, *C.albicans* and related species (*Candida auris*) display increasing resistance to the conventional antifungal treatments. Regarding this point, a better understanding of the fundamental mechanisms involved in mtDNA maintenance is a potential starting point for drug discovery. This thesis work provides novel information regarding how Gcf1p binds and compacts mtDNA.

This thesis work also provides with novel information that can be of high interest in evolutionary biology. Origin and evolution of mitochondria from independent organisms to cell organelles (endosymbiotic theory) is broadly regarded by the scientific community as a key point in the apparition of eukaryotes (protists, fungi, plants and animals). The mechanisms behind mtDNA compaction and replication shows a high diversity both inter-reigns and intra-reigns. Gcf1p is also related with recombination-driven-replication mechanism present in *C.albicans,* which is completely different to the current model for mtDNA replication in mammals. The results provided by this thesis work in regard the union and compaction of DNA by Gcf1p suppose also a starting point for comparing the mtDNA in *Candida albicans* in regard of other yeast (mainly *Saccharomyces cerevisiae* (*S.cerevisiae*)) as well as in regard of distant organisms (*Homo sapiens* (*H.sapiens*)). It is also a starting point to understand evolutionary divergence amongst eukaryotes.

This work has made use of molecular biology techniques for the cloning in plasmid vector, as well as, for the heterologous expression of yeast proteins in *Escherichia coli*. In addition, protein purification techniques have been applied obtaining an optimized production protocol compatible with the experimental analysis by means of structural biology methods, mainly macromolecular crystallography (MX), Small-Angle X-ray Scattering (SAXS) and Electron Microscopy (EM). Results from these three techniques provide complementary evidences about the interaction of Gcf1p with DNA.

In summary, this thesis work provides evidences that indicate that the mtDNA recognition mechanism in *C.albicans* presents fundamental differences regarding those described for *S.cerevisiae* and *H.sapiens*. Our results suppose an important advance in the description of spatial organization of mitochondrial DNA, an essential process eukaryotic organism.

INTRODUCTION

Candida albicans, yeasts on the tightrope between commensalism and pathogenicity

The *Candida* genus is composed of dimorphic yeast species, i.e. species that can grow both in cellular and filamentous (hyphal) forms. *Candida* spp. form part of the varied fungal microbiota, i.e. mycobiota, present in mammalian hosts, including *Homo sapiens*. Amongst species belonging to the *Candida* genus, *Candida albicans* arises as the most common species found in the mucosal surfaces of warm-blooded animals [1]. In mammals, *Candida* spp. are particularly prominent in the gastrointestinal (GI) tract, they are also present in the oral microbiome [2], in the reproductive tracts [3] and at the external epithelium [4]. *Candida* spp. are early-colonizers of the host mucosa, as they are acquired at perinatal stages by the neonate through mucosa-mucosa and skin-mucosa physical contact [5] *Candida* spp. are found in the gut microbiota of healthy infants and young children [6]. As common commensals, *Candida* spp. are thus part of the healthy mycobiota, and carry out beneficiary functions for the host derived from its colonization and commensalism lifestyle. Among other functions, they regulate the GI-tract mucosal immune response by increasing the relative presence of Th17 lymphocytes, involved in pathogen clearance, in the colon *lamina propria* of murine models [7]. In addition, *Candida*-colonized mice have a reduced vulnerability to infection by the major nosocomial pathogen *Clostridioides difficile* [8]. In *Homo* sapiens, *Candida albicans* is related to the triggering of the trained immunity pathways. It has been proposed to mainly influence the STAT-1 pathway, via adhesion and recognition of *Candida albicans* extracellular components such as chitin and β-Glucan through host Dectin-1 fungal receptor [9].

Despite its beneficial effects to the host related to its commensalism, *Candida* genus have a notorious ability in infecting and causing a broad range of diseases in the host. This ranges from superficial (skin and mucosa) to invasive (inner organs) infections that sometimes end in life-threatening systemic infections. Amongst *Candida* spp., the beneficial commensal *C. albicans* emerges as a prototypic opportunistic pathogen. It is cause of an astonishing high number of infections worldwide. For example, vulvovaginal candidiasis (VVC) affects 138 million women annually (within a range of 103-172 million) with a global annual prevalence of 3871 per 100000 women and with a maximum prevalence of 9% in the 25-34 age range [10]. Furthermore, a myriad of mouth, throat and oesophagus infections caused by *C. albicans* should be added to these statistics, although no usual national surveillance exists for such infections [11].

Candida spp. are not only cause of localized infections in patients, but their potential to cause invasive infections (invasive candidiasis) in hospitals is of increasing concern. A notorious infection of nosocomial origin, is *Candida spp.* bloodstream invasion or candidemia [12], which is the main form of invasive candidiasis and it has been one of the most common causes of bloodstream infections in hospitalized patients in the United States along the last two decades [13], [14]. Such bloodstream infections cause longer hospital stays and increase the healthcare-associated costs and mortality [15]. Nosocomial candidemia can reach mortality rates of 49%, as revealed by clinical studies in affected patients in intensive care unit [16]. Bloodstream-invasive candidiasis, or candidemia, usually affects people with a weakened immune system including chemotherapy patients, organ-transplanted patients and patients affected by neutropenia [17]. Other risk groups are intensive care unit long-term patients, pre-term neonates, post-surgery patients, multiple-antibiotic-treated patients, parenteral or intra-venous catheterized patients, drug-abusive patients and patients with previous pathologies such as diabetes or kidney failure [11]. Increasing numbers of habit-related diseases, specially obesity and diabetes, over-use of antibiotics in intensive care unit patients and saturation of the public hospitals are behind the dramatic increase of nosocomial candidiasis and candidemia within the last years. Treatment of invasive candidiasis usually consists in a combination of antifungal medication including echinocandins (caspofungin, micafungin, or anidulafungin), as well as fluconazole and amphotericin B. However, due to the high mortality rate of invasive candidiasis and drug-resistance acquired by *Candida* spp., antifungal prophylaxis is a preferred method and it is normally administered to risk patients in intensive care unit patients [11].

Over a hundred *Candida* species are part of our commensal flora but only a few are known to cause infections. The most commonly causing invasive systemic infections are *C. albicans, C. glabrata, C. parapsilosis, C. tropicalis* and *C. krusei* [11]. *C. albicans* causes most of the cases of invasive candidiasis but it has reduced drug-resistance. Other minor species, such as *C. glabrata*, are frequently drug-resistant and show a higher mortality rate [18]. An emerging pathogen species, *Candida auris*, is a cause of major concern for public healthcare systems. *C. auris* shows drug-resistance to the three classes of antifungal drugs used for the treatment of invasive candidiasis: the azoles (including fluconazole), the echinocandins and amphotericin. It is usually misidentified as a less dangerous *Candida* species and it has already caused outbreaks in healthcare settings [19]. Stablished and emerging *Candida* spp. are therefore a major health concern, expected to affect an increasing number of hospitalized patients and to display also an increased drug-resistance ([11]. *C. albicans*, as the best characterized Candida species, is of high interest as biomedical target. A better understanding of the mechanisms underlying *C. albicans* viability is key for the development of new treatments that may scape the drug resistance mechanisms developed by *Candida* invasive species.

Mitochondria, the powerhouse of the cell

Mitochondria are eukaryotic cell organelles that play a central role in metabolism, hosting the pathways involved in the degradation of lipids, carbohydrates and aminoacids, as well as the Electron Transport Chain (ETC) which is coupled to the oxidative phosphorylation, the main source of ATP production [20]. Furthermore, more recent discoveries highlight the central role of mitochondria in calcium homeostasis, cell survival, senescence and tumorigenesis. Therefore, mitochondria are a prominent research target in biochemistry, to understand the pathways and their interdependence; in biomedicine, to address the mitochondrial malfunctioning-related diseases; and in evolutionary biology, that explores the potential origins of mitochondria.

The origins of mitochondria

The evolutionary mechanism by which mitochondria were originated is still undetermined but is intimately linked to the origin of eukaryotic cells and three main models of this evolutionary process are currently available (**Figure I1**). As stated in [21]: '*It is now known that the last eukaryotic common ancestor was complex and that endosymbiosis played a crucial role in eukaryogenesis at least via the acquisition of the alphaproteobacterial ancestor of mitochondria'*. According to the endosymbiotic theory, mitochondria derived from independent aerobic bacteria (α-proteobacteria) that were engulfed by a proto-eukaryote. Both entities evolved together in a symbiotic relationship, the α-proteobacteria became a membrane-enclosed subcellular organelle that provided the ability to the host to survive as an aerobe, while receiving nutrients and shelter. The establishment of such an interrelation was tremendously beneficious for both parts and was selected by evolution as it enabled eukaryotes to thrive in an oxidative environment [22]. Throughout evolution, mitochondria evolved from symbiotic partners to cell organelles that exert a central role in cell metabolism. Nevertheless, mitochondria maintained its structures that evidence its independent origin, as it will be explained below (*see Mitochondria structure and compartments section,* **Introduction**).

The endosymbiotic origin of eukaryotic cells were first suggested in 1905 by Konstantin Mereschkowsky [23] [24], who proposed that plant chloroplasts were originally independent cyanophytes, a photosynthetic cyanobacteria species. This hypothesis was rescued in 1966 by Lynn Margulis, who revived the idea that chloroplasts evolved from incorporated cyanobacteria and proposed that mitochondria evolved from an ancestral purple bacteria [25]. Notwithstanding, the first organelle phylogenetic analysis in 1978 placed chloroplasts among cyanobacteria and mitochondria among α-proteobacteria, a finding that was highly controversial. Evidences of gene transfer from organellar to nuclear genome [26] [27] further supported endosymbiosis. Deep-sequencing analysis confirmed the monophyly of mitochondria within α-proteobacteria, i.e. all contemporary mitochondria derive from the same prokaryotic ancestor [21]. Likewise, eukaryotic lineage monophyly from a Last Eukaryotic Common Ancestor (LECA) is well stablished. The gene transfer from the mitochondrial to the nuclear genome has been proven by phylogenetic studies, and genes from mitochondrial origin can be tracked back in nuclear genomes amongst eukaryotes, even to parasitic protists lacking mitochondria, which suggested that LECA already possessed mitochondria [21].

Gene transfer from mitochondria to nucleus is an example of a general trend observed in obligatory symbioses, in which genes of the endosymbiont are transferred to the host leading to a strong genome reduction at the endosymbiont, and thus increasing inter-dependence and host-shaped evolution of the endosymbiont [26]. Genome reduction is a significant trait of mitochondria, since its current-living α-proteobacterial relatives possess genomes of ~1.3Mb (in free-living *Pelagibacter* spp. and *Rickettsia* spp.) [21], whereas the size of human mtDNA is of 16.5 Kb, or a bit more than 85 kb in *S. cerevisiae*. The largest mitochondrial genomes count with ~100 genes, whereas the smallest ones codify for only 13 or even 3 proteins, in animal and apicomplexan protists, respectively. Comparison of mitochondrial genomes of *Homo sapiens* (16.5 kb, 37 genes), *Saccharomyces cerevisiae* (85 kb, 35 genes) and *Candida albicans* (40 kb, 40 genes) evidences such a variability. mtDNA genomes encode not only proteins, but also rRNAs and tRNAs to equip mitochondria for protein synthesis.

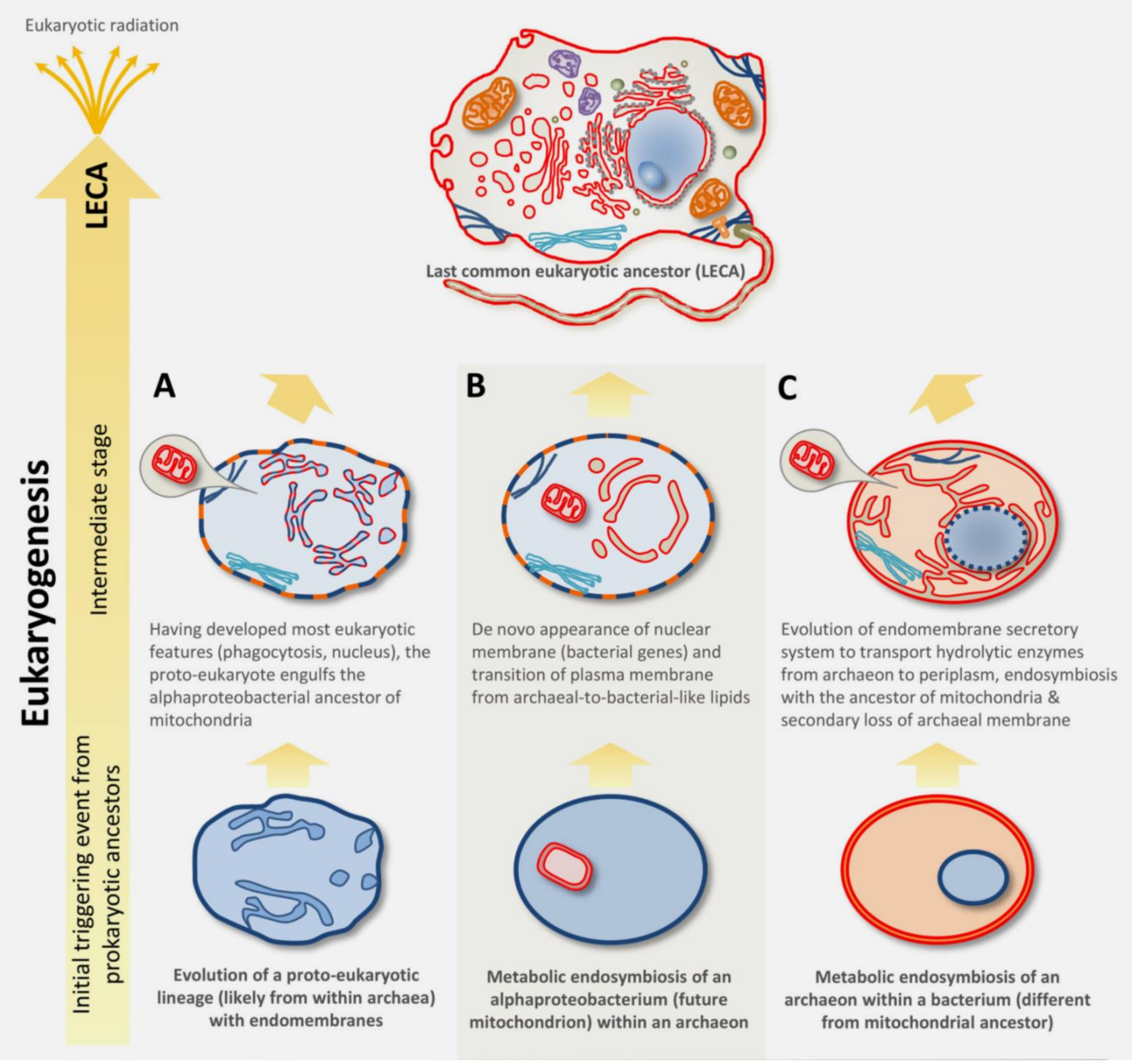

Eukaryotic radiation
LECA
Last common eukaryotic ancestor (LECA)
Eukaryogenesis
Intermediate stage
Initial triggering event from prokaryotic ancestors
A
B
C
Having developed most eukaryotic features (phagocytosis, nucleus), the proto-eukaryote engulfs the alphaproteobacterial ancestor of mitochondria
De novo appearance of nuclear membrane (bacterial genes) and transition of plasma membrane from archaeal-to-bacterial-like lipids
Evolution of endomembrane secretory system to transport hydrolytic enzymes from archaeon to periplasm, endosymbiosis with the ancestor of mitochondria & secondary loss of archaeal membrane
Evolution of a proto-eukaryotic lineage (likely from within archaea) with endomembranes
Metabolic endosymbiosis of an alphaproteobacterium (future mitochondrion) within an archaeon
Metabolic endosymbiosis of an archaeon within a bacterium (different from mitochondrial ancestor)

Figure I1, three hypothetical origins of the eukaryotic cell: *Three hypothesis on the endosymbiotic theory as the origin of eukaryotic lineage. They all have in common the monophyly of mitochondria within alpha proteobacteria and differ on the origin and apparition time of the nucleus and endomembranous system along evolution. In A, mitochondria are incorporated by a proto-eukaryotic cell with already developed eukaryotic features. In B, on the other side, such eukaryotic features as endomembranous system and nucleus appear after mitochondria incorporation, which appear as a 'triggering factor for eukaryogenesis'. Finally, in C, nucleus and the endomembranous system appear as a result of a first endosymbiosis event after the incorporation of an* archaeon within bacteria *and mitochondria would later be incorporated in a second event. The monophyletic origin of* eukaryotic *traces back to the last eukaryotic ancestor, from which protozoa, fungi, algae, plants and animals have radiated.* Extracted and adapted from [21]. López-García et al. 2015, Trends ecology evolution.

Mitochondria structure and compartments

Human mitochondria are 1-2 µm long and 0.5 µm wide organelles with mitochondrial outer and inner membranes (MOM and MIM, respectively) separated by an intermembrane space [28] (**Figure I2**). In addition, MIM expands towards the matrix forming specialized structures, the cristae. Mitochondria are highly dynamic and have the ability to change their shape and size in function of the metabolic state of the cell via mechanisms of fission and fusion [29]. Despite the overall structural changes, the different intra-mitochondrial compartments maintain their unique structural characteristics and associated functions.

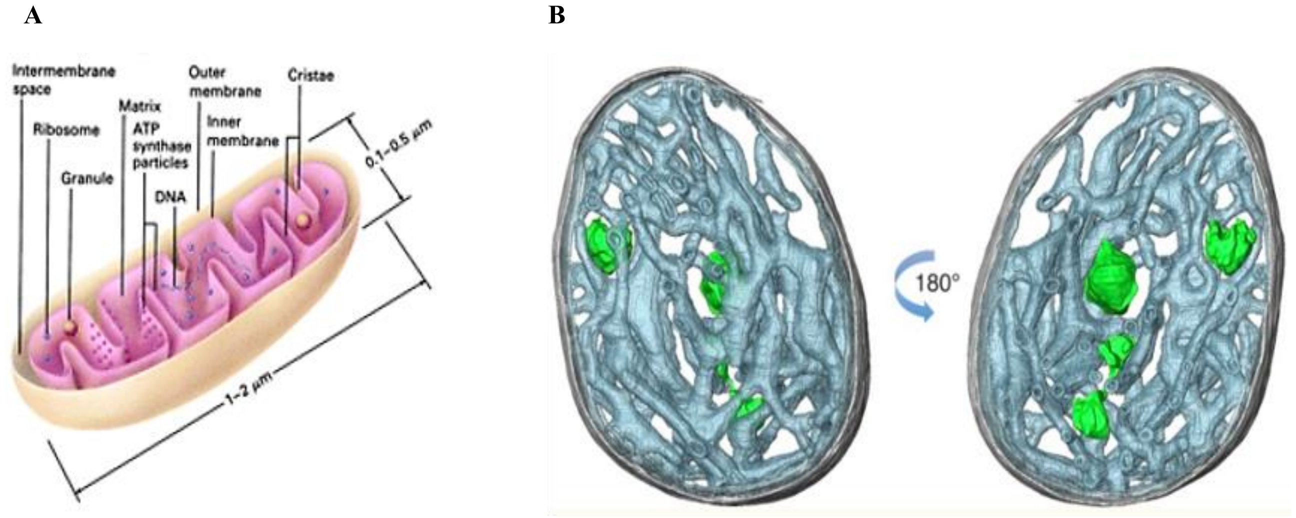


Figure I2, mitochondrial compartments. *(A) A diagram of the mitochondria indicating each of its inner compartments* (extracted from: *[30]* Logan, David C. 2006, Journal of experimental botanics). ***(B)*** *Cryo-electron tomography 3D reconstruction of a transversal thin section of bovine mitochondria. In gray the* ***mitochondrial outer membrane****, in blue the* ***mitochondrial inner membrane*** *and the* ***mitochondrial cristae****. Mitochondrial nucleoids (in green) appear attached to the cristae* (extracted from: *[31]* Kukat, Christian, et al. 2015, PNAS).

Mitochondrial outer membrane (MOM), is a phospholipid bilayer rich in mitochondrial porins, which are 30-35 kDa proteins also known as Voltage-Depending Anionic Conducts (VDAC) [32]. Due to the presence of such channels, the outer membrane is porous to certain metabolites (mainly anions) such as phosphates, chlorides, organic anions and adenine nucleotides that enter the intermembrane space by VDAC-mediated passive transport.

Mitochondrial inner membrane (MIM), contrary to the former, this membrane is highly impermeable to most of the ions contained in the intermembrane space. A broad range of specialized transporters populate MIM and transfer metabolites such as pyruvate, ATP and citrate to the mitochondrial matrix [32]. Importantly, MIM contains four complexes, Complex I to IV, which are close to each other and altogether form the Electron Transport Chain (ETC), or respiratory chain. The ETC captures electrons from products of the last step of catabolism (NADH and $FADH_2$), which are transferred from complex I to IV throughout the membrane until they reach a ½ diatomic oxygen molecule (O_2) and two protons (H^+) from the matrix and form a water molecule. The transport of electrons by ETC is coupled to the translocation of H^+ from the matrix to the intermembrane space, where they accumulate so that an electrochemical gradient with associated membrane potential ($\Delta\Psi$), is generated. The last complex that participates in this pathway is the F_1-F_0 ATPase or complex V [28]. By the F_1-F_0 ATPase, H^+ accumulated at the intermembrane space traverse the inner membrane back to the matrix, following the electrochemical gradient. When a proton enters the F_1-F_0 ATPase complex and is released to the matrix, 1/3 ATP molecule is synthesised by Complex V from ADP+Pi (or, three H^+ generate one ATP). Due to the ETC and this last phosphorylation step, the whole pathway is termed oxidative phosphorylation (OXPHOS). Note that in most eukaryotes, the ATP synthesized by OXPHOS is the main source of energy for the organism and thus is key for the metabolism of eukaryotic cells (this pathway is explained in detail in *mitochondria and metabolism section,* **Introduction**).

Mitochondrial cristae, the inner membrane propagates towards the matrix forming specific structures, the cristae, which are indeed invaginations of MIM. Note that the OXPHOS pathway is mainly localized at the cristae, and cristae change their overall shape, size and available surface depending on the metabolic state and ATP requirements of the cell, a process regulated by cristae-shaping proteins [33].

Mitochondrial intermembrane space: Due to the permeability of MOM, the intermembrane space has a chemical composition similar to that of the cytoplasm, except for those particles bigger than 5 kDa, which depend on specific active transport through the Translocase Outer Membrane complex (TOM) [20]. The Translocase Inner Membrane complex (TIM) together with the soluble Hsp70 chaperones translocate the proteins through MIM, from the intermembrane space to the mitochondrial matrix. This is an ATP-dependent process [20]. It could be argued that this energetic cost, associated to protein transfer, is the toll to pay by the cell for the endosymbiosis-associated gene transfer responsible for the mitochondria transition to cell organelles.

Mitochondrial matrix corresponds to the lumen of mitochondria, the space delimited by the inner face of MIM. Due to the continuous electron transport through the ETC complexes and the associated proton transport and accumulation at the intermembrane space, the mitochondrial matrix has a pH = 8, higher than that of the intermembrane space, with a pH = 7. As it will further be discussed in the following section, integrity of the mitochondrial inner membrane and maintenance of the electrochemical gradient is essential for the mitochondrial energy production and organelle integrity (*see mitochondrial and metabolism section,* **Introduction**). Due to the restrictive permeability of MIM, the composition of the mitochondrial matrix is different to that of the intermembrane space and the cytoplasm not only for the proton concentration but also for other metabolites. A case worth noting is the concentration of Ca^{2+} ions. Since the 1960's decade, it is known that mitochondria can rapidly uptake Ca^{2+} ions to the mitochondrial matrix as reviewed in [34]. However, the exact mechanism by which Ca^{2+} ions enter the mitochondria remained a mystery for many decades. It is now known that the calcium uptake by mitochondria is buffered by calcium-binding proteins present in the mitochondrial matrix and that such binding/release of the ions modulate the activity of the ETC [35]. Furthermore, the capability of the mitochondrial matrix to store calcium ions places mitochondria in a central position in cell signal transduction. Mitochondrial matrix hosts several metabolic routes which are essential for the survival of cells and pluricellular organisms. Among them, it is worth noting the urea cycle, which regulates ammonia elimination in mammals; the β-oxidation pathway, involved in lipid degradation; or the Krebs or Tri-Carboxylic Acid (TCA) cycle, which produces substrates for the ETC that are from the Acetyl-CoA delivered by the degradation of aminoacids, lipids and carbohydrates). Amongst the approximately 1000 mitochondrial matrix proteins, caspase-2 is a pro-apoptotic protein that induces programmed cell death upon several stimuli directly from the mitochondrial compartment. Thus, caspase-2 links proper mitochondria functioning with cell viability and tumorigenesis [36]

At the mitochondrial matrix, there are membrane-less compartments formed by nucleoprotein complexes. Amongst them, the most recently discovered ones are the **mitochondrial RNA granules**. Such granules are formed by the newly synthesized mitochondrial RNA and have been implicated in mitochondrial RNA processing and ribosome biogenesis. In a recent work, which we had the opportunity to collaborate with, it has been shown that such mitochondrial granules contain classical human mtDNA replicative factors, such as the replicative helicase Twinkle and the mitochondrial Single-Stranded DNA binding protein (SSBP-1) [37].

The best characterized membrane-less compartment within mitochondrial matrix and the one that is most related with the contents of this work is the **mitochondrial nucleoid**. Mitochondrial nucleoids are nucleoprotein complexes formed by the mitochondrial DNA (mtDNA) and mtDNA-regulatory proteins, including packaging factors. In eukarya, the proteins involved in mtDNA compaction belong to the HMG-box family [31] [38]. Amongst the HMG-box proteins known to package mtDNA, the only structurally-characterized ones are the human Transcription Factor A (TFAM) [39], [40], [41] and the *Saccharomyces cerevisiae* Ars-Binding factor 2 (Abf2p) [42]. Both proteins have been extensively characterized in our laboratory by X-ray crystallography and biophysical methods (see next sections for structural and functional description).

Significant advances were done on human mitochondrial nucleoid morphology and dynamics by biochemical and microscopy methods [43]. According to these works, human mitochondrial nucleoids are densely packed nucleoprotein complexes formed by mtDNA and TFAM together with other proteins such as the replicative and transcription machineries (Single-Stranded DNA Binding Protein 1, SSBP-1 and Twinkle replicative helicase), the ATPase ATAD3 or the protease LON. One model of nucleoid organization proposed a central core of mtDNA regulatory proteins that is covered by a layer of more loosely bound, transient proteins [43]. The advent of super-resolution optic microscopy [44] allowed for the visualization of human mitochondrial nucleoids in mitochondria [45], [31]. They form irregular elongated spherical particles of ~100 nm that contain 1.4 human mtDNA molecules per nucleoid on average [45]. Nucleoids are attached to the matrix-facing side of the inner mitochondrial membrane, preferentially located in cristae-free regions and coordinated in space with RNA granules and mitochondrial ribosomes [46]. Gene expression of mtDNA is correlated by the overall nucleoid compaction *in vivo*, where transcription have been proposed to occur in less dense, expanded nucleoids containing relaxed DNA [46]. In vitro experiments showed that high TFAM levels result in high DNA compaction and transcription and replication repression [47]. It has been proposed that mitochondrial nucleoids act as a switch that controls mtDNA expression, shield and preserve mtDNA from oxidative damage, and provide the required DNA metabolism machinery to ensure proper DNA replication and segregation during mitochondrial division [48]. Mutations in mtDNA and/or mtDNA-regulatory proteins have been linked to the onset of disabling mitochondrial diseases [49]. Mutations in SSBP1 cause optic atrophy and foveopathy [50]. Similarly, mutations on the adenine transporter ANT1, which interacts with mtDNA [51], is linked to multiple mtDNA deletions and progressive ophthalmoplegia [52]. In addition, subunits of the ETC complex I and the E2 subunits of ATP synthase and 2-oxo-acid dehydrogenase have been identified in nucleoids and related to the onset of mitochondrial diseases and aging [51]. More generally, mutations in mtDNA as well as nucleoid aberrant morphologies might lead to a broad range of conditions, including rare diseases, neurodegenerative disorders and cancer [48], [53], [54].

Mitochondria and cell metabolism

From the broadest point of view, metabolism refers to the sum of all the chemical transformations that may take place within a living organism. Metabolism is a highly coordinated cellular activity in which multienzyme systems, the so-called metabolic pathways, interact and cooperate. Functions are disparate and include pathways involved in the generation of chemical energy from light (photosynthesis) or are related to the degradation of energy-rich nutrients in so-called catabolic pathways. On the contrary, anabolic pathways convert nutrients to macromolecule precursors, which are polymerized to form cell components such as proteins, carbohydrates, lipids and nucleic acids. Synthesis and degradation of macromolecules is required to sustain specialized functions, exerted be signalling factors, pigments, pheromones, alkaloids or venoms (secondary metabolism) [55]. The term **catabolism** refers to the degradative phase of the metabolism, in which organic nutrient molecules of different chemical nature (carbohydrates, lipids and amino acids) are converted into smaller end products (CO_2, NH_3, Lactic acid…) in a process that releases energy, generally used for the production of Adenosine Triphosphate (ATP). In **anabolism** or biosynthesis, small simple precursors (CO_2, Acetyl-CoA…) are transformed into complex organic molecules of higher energy content (carbohydrates, lipids, aminoacids and nucleotides) that may then be used for the formation of even more complex molecules and cell structures (proteins, DNA, RNA, membranes, ribosomes…). Anabolic reactions often require an energy input provided by the highly energetic phosphoryl group transfer from ATP and the reducing power provided by the co-factors NADH, NADPH and $FADH_2$.

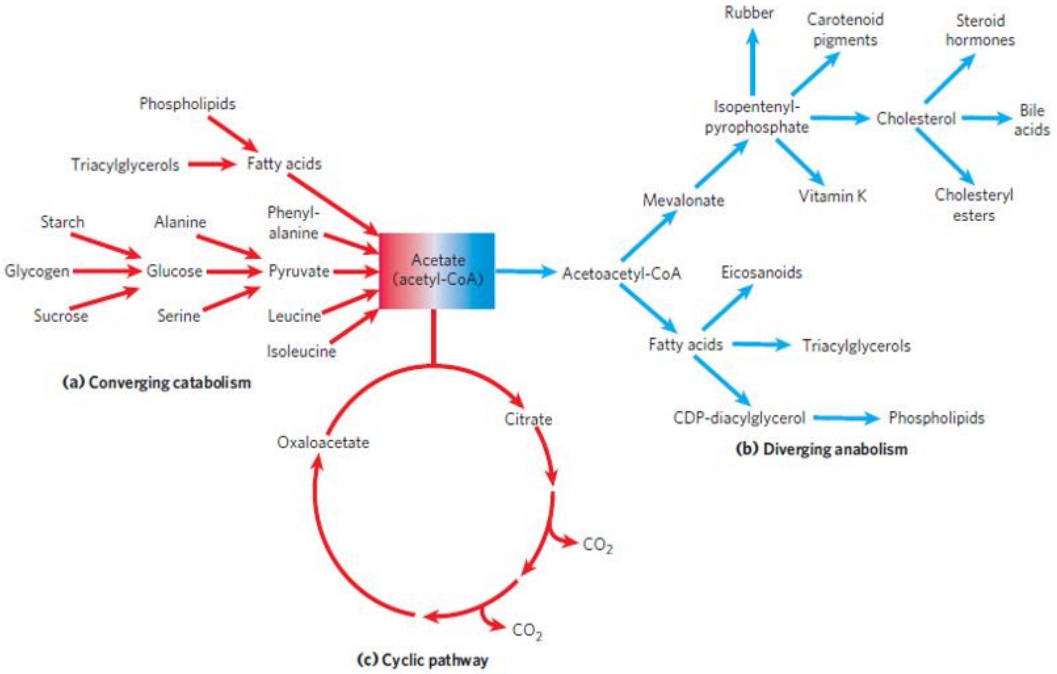


Figure I3: *Schematic view of the catabolism (**a**) of carbohydrates, lipids and amino acids converging in the acetate (acetyl-CoA). The anabolism (**b**) diverges from acetate to a broad variety of complex molecules of lipidic nature (cholesterol, fatty acids, Vitamin K…). In between an example of cyclic pathway, the TCA cycle (**c**) which takes place in mitochondria and interconnects catabolism with anabolism.* Extracted from: [55] Nelson, David L and Cox, Michael. Bioenergetics and metabolism. Lehninger Principles of Biochemistry. 2008, pp. 505-543.

The so-called **intermediary metabolism** consists on the processes that interconvert precursors and metabolites of less than 1000 Da. Such processes are found in both anabolism and catabolism and are usually connected by **cyclic metabolic pathways** in which the end-product is recycled and re-enters at the first step of the pathway (see **Figure I3**) [55]. Cyclic metabolic pathways allow for the continuous interplay between anabolism and catabolism and for the self-sustainability of the system. Due to its nature, such cycles may have **amphibolic nature**, i.e. they may serve as a platform for both anabolic and catabolic reactions (**Figure I4**). The most prominent example of a cyclic metabolic pathway with amphibolic properties is the **Citric Acid Cycle**. Also known as tricarboxylic acid cycle (**TCA cycle**) or **Krebs cycle**, it takes place within the mitochondrial matrix. Along this cycle, activated acetate (acetyl-CoA) produced from glucose [56, pp. 543-587] and fatty acid degradation [57, pp. 667-695] is subsequently oxidized in order to reduce the cofactors NAD^+ and FAD^+ to NADH and $FADH_2$, respectively. In addition, one molecule of GTP is generated per each turn of Krebs cycle [58, pp. 633-667]. The production of reduced cofactors is an efficient way to preserve the energy liberated from oxidation reactions in the catabolic steps of the Krebs cycle, and thus feed the electron transport chain (ETC). As a by-product of the catabolic reactions of the TCA cycle, 2 CO_2 molecules are released per each turn of cycle. Despite the central role of the TCA cycle in the energy-yielding metabolism, Despite the central role of the TCA cycle in the last steps of catabolism, its role does not only limit to the production of ATP, GTP and reduced NADH and $FADH_2$. Four and five carbon intermediate molecules of the cycle serve as precursors for a plethora of biomolecules as depicted in **Figure I4**. It is worth highlighting that both α-ketoglutarate and oxaloacetate serve as precursors of aspartate and glutamate respectively, which can be precursors of purines and pyrimidines synthesis pathways. Whilst in aerobic organisms the TCA cycle shows such amphibolic properties, in anaerobic organisms (in which the ETC is not functional) the accumulation of succinate and α-ketoglutarate due to the incompletion of the cycle serve as precursors for the synthesis of amino acids, nucleotides and heme groups [58, pp. 633-667]. This highlights the TCA cycle not only as the last pathway for acetate oxidation and overall catabolism but as a hub of intermediary metabolism.

All oxidative steps in the degradation pathways of carbohydrates, fats and aminoacids -including the TCA cycle- converge and culminates in the **mitochondrial Electron Transport (respiratory) Chain** (ETC) coupled to the **oxidative phosphorylation** synthesis of ATP (OXPHOS) [59, pp. 732-788]. Five protein complexes located in the mitochondrial inner membrane participate in this process, as briefly introduced above. Four of these complexes (I, II, III and IV) transfer electrons from complex I to IV while transferring H+ to the intermembrane space and thus creating a membrane potential (or H+ gradient across the membrane) in small steps that, eventually, is used for ATP synthesis by the last complex V, the ATP synthase.

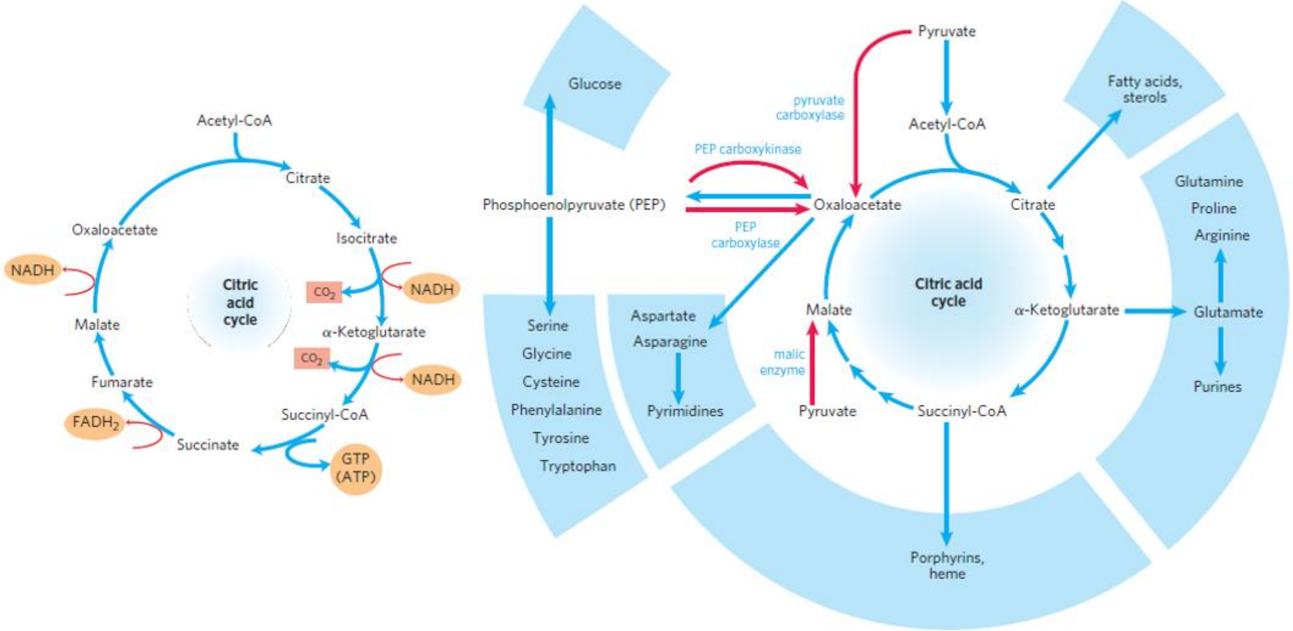


Figure I4: *Mitochondrial TCA cycle as a source of ATP and reducing power in the form of reduced cofactors (NADH and FADH$_2$) (left) and as a source of intermediary metabolites to serve as precursors of several anabolic pathways (right).* Extracted from *[58]* Nelson, David L and Cox, Michael. The citric acid cycle. Lehninger Principles of Biochemistry Sixth Edition. 2008, p. 633-667.

The sequential order of the ETC was determined experimentally by using specific inhibitors for each complex [59, pp. 732-788]. Electrons enter to the ETC through universal electron acceptors, i.e. (NADH/NADPH)-dependent dehydrogenases and flavoproteins linked FADH$_2$. Complexes of the electron transport chain contain diverse prosthetic groups that act as electron shuttles along the ETC, i.e. the ubiquinone/ubiquinol pair (**coenzyme Q**), iron-containing prosthetic heme groups of **cytochromes** (**a**, **b**, **c**), and **iron-sulphur** (**Fe-S**) cluster in which the reduced iron atom is associated with inorganic sulphur, or with sulphur atoms from cysteine residues, or both [59, pp. 732-788]. ETC complex I, also known as NADH dehydrogenase, catalyses two obligate coupled processes: spontaneous transfer of two hydride (H^-) ion to ubiquinone and the pumping of 4 protons (H^+) from the mitochondrial matrix to the intermembrane space. Complex II or succinate dehydrogenase, is also the only membrane-bound enzyme of the TCA cycle. Complex II contains a substrate binding site for succinate that is oxidized to fumarate with the help of the covalently bound FAD cofactor. Thus, the electrons are transferred from FADH$_2$ to different Fe-S centres, and from those to membrane-embedded ubiquinone. Complex III or ubiquinone-cytochrome c oxidoreductase catalyses the electron transfer from reduced ubiquinone (ubiquinol) to cytochrome c coupled to the transport of H^+ from the mitochondrial matrix to the intermembrane space. Finally, the complex IV or cytochrome c oxidase carries electrons from cytochrome c to molecular O_2 reducing it to H_2O. Such an energetically favoured redox reaction is coupled to the pumping of a H^+ per every electron transferred by cytochrome c [59, pp. 732-788]. Continuous flow of electrons through the ETC supposes a continuous pumping of H^+ to the intermembrane space, generating an electrochemical gradient of H^+ across the mitochondrial inner membrane. Such a proton gradient is used as a proton-motive force by the F-type ATPase F_0-F_1 ATP synthase or Complex V, which couples the transfer of protons back to the mitochondrial matrix with the synthesis of ATP from ADP and inorganic phosphate (P_i) (**Figure I5**). Therefore, by means of the OXPHOS system, the energy obtained from nutrient oxidation is stored with the generation of an electrochemical gradient and finally by the formation of the highly energetic phospho-anhydrous bonds of ATP. This molecule, which stores the energy in a chemical form, is then used by the cell to perform a broad range of reactions involved in anabolism, protein synthesis, DNA replication, gene expression, intracellular transport and contraction in the case of skeletal muscle cells [55, pp. 732-788] [60].

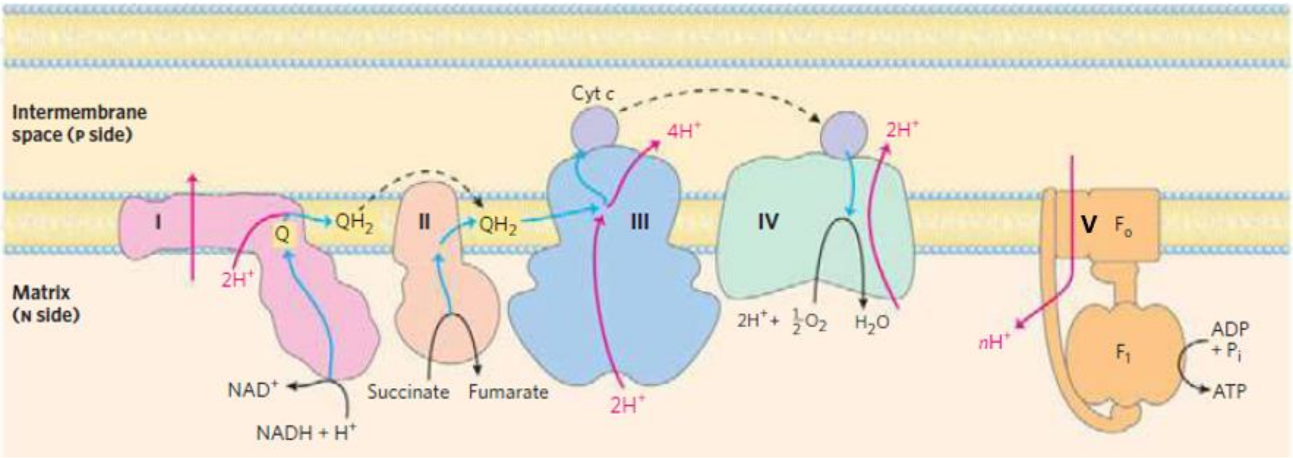


Figure I5: *Schematic representations of the oxidative phosphorylation pathway in which proton pumping by the Electron Transport Chain (Complexes I to IV) is coupled to the proton gradient-driven ATP synthesis (Complex V).* Extracted from: *[59]* Nelson, David L and Cox, Michael. Oxidative phosphorylation and photophosphorylation. Lehninger Principles of Biochemistry Sixth Edition. 2008, pp. 732-788.

The OXPHOS system, located in MIM, provides further indications about the bacterial origin of mitochondria, thus supporting the endosymbiosis theory. The membranes of current bacteria possess the pathways to transfer electrons from substrates (such as NADH) to O_2, which is coupled to the phosphorylation of cytoplasmic ADP. Furthermore, rotational molecular motors of bacterial flagella are fuelled by an electrochemical gradient between the periplasm and cytoplasm, generated by an ETC that pumps H$^+$ and which is similar to the OXPHOS pathway found in mitochondria [59, pp. 732-788].

Candida albicans mitochondrial DNA, a linear genome without telomeres

The mitochondrial genome of *Candida albicans* is a linear molecule of 40.4 kb bp without telomeres. Recent high-throughput RNA sequencing studies revealed the mtDNA transcriptome of this microorganism [61]. Gene transfer occurred also in C. albicans during the mitochondria endosymbiotic process [26], but a portion of essential genes were maintained in *Candida albicans* mitochondrial genome.

Candida albicans mtDNA contains two independent coding regions: the **short coding region** (**SCR**, coordinates from kb ~1 to 6) and the **long coding region** (**LCR**, kb ~12-34) (Figure 6). Altogether, both LCR and SCR code for genes of 14 components of the OXPHOS pathway that belong to Complex I, Complex IV and Complex V. In addition, LCR and SCR code for 2 rRNA (rRNAs) that belong to the two mitochondrial ribosome subunits, and 24 tRNAs codified in polycistronic Transcription Units (TUs). Two inverted repeats regions IRa and IRb with scarce genetic content spans in the ranges of kb ~6-12 and ~34-40, IRa separate both SCR and LCR (**Figure I6**).

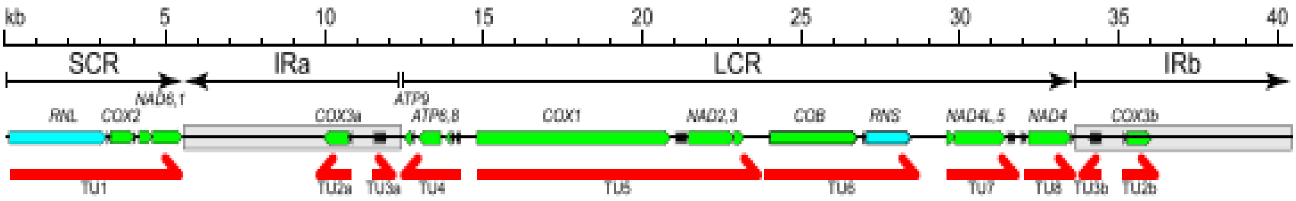


Figure I6, mitochondrial DNA coding sequences in Candida albicans. *On top, schematic view of the coding regions within C. albicans mtDNA. Blue rectangles represent rRNA coding regions, green rectangles represent the ETC subunits codified in mtDNA and red arrows represent the transcription units (TU) containing tRNAs.* Extracted from: *[61]*. Kolondra, Adam, et al. 2015, BMC Genomics

Linear yeast mtDNA (40-80 kb depending on the species) shows a different organization than the circular, intron-less and overlapping mammalian mitochondrial genome (16.5 kb in human) [62]. Within yeast kingdom, remarkable variability of mtDNA organization is displayed. In this context, research in *C. albicans* mtDNA is of great interest, not only because C. albicans is a human pathogen but because is a representative clade that is distant from the model yeasts *Saccharomyces cerevisiae* and *Schizosaccharomyces pombe*. Unlike them, *C. albicans* mtDNA codifies for Complex I subunits. More striking differences appear when *C. albicans* mtDNA structure and topology is compared to that of the yeast models. Whilst the mitochondrial genome of model yeast organisms was earlier identified as a linear genome as opposed to that of mammalian [62], in *C. albicans* circular mtDNA molecules were inferred as a result of first endonuclease mapping experiments [63], [64] and [65]. Such a circular arrangement in *C. albicans* mtDNA was proved to be incorrect when 2D agarose gel analysis (**Figure I7A** and **I7B**) showed that no circular replicative intermediates were found in *C. albicans* mtDNA. Instead, strand invasion structures were detected in the inverted repeat of *C. albicans* mtDNA [66]. Using specific restriction enzymes, the regions involved in such homologous recombination events were localized (**Figure I7C** and **I7D**). Such results suggested a recombination-driven replication (RDR) mechanism for *C. albicans* mtDNA. According to the RDR replication model, *C. albicans* mtDNA would be primed by the 3' end of an invading strand resulting from homology recombination events in the inverted repeat regions (**Figure I7E**). Such a model provides a reasonable explanation of how primers for the DNA polymerase can be created in this telomere-less linear genome. Similar findings on *C. parapsilosis* replication suggest a common recombination-based mechanism of replication for yeast mtDNA, or at least within the *Candida* genus [67].

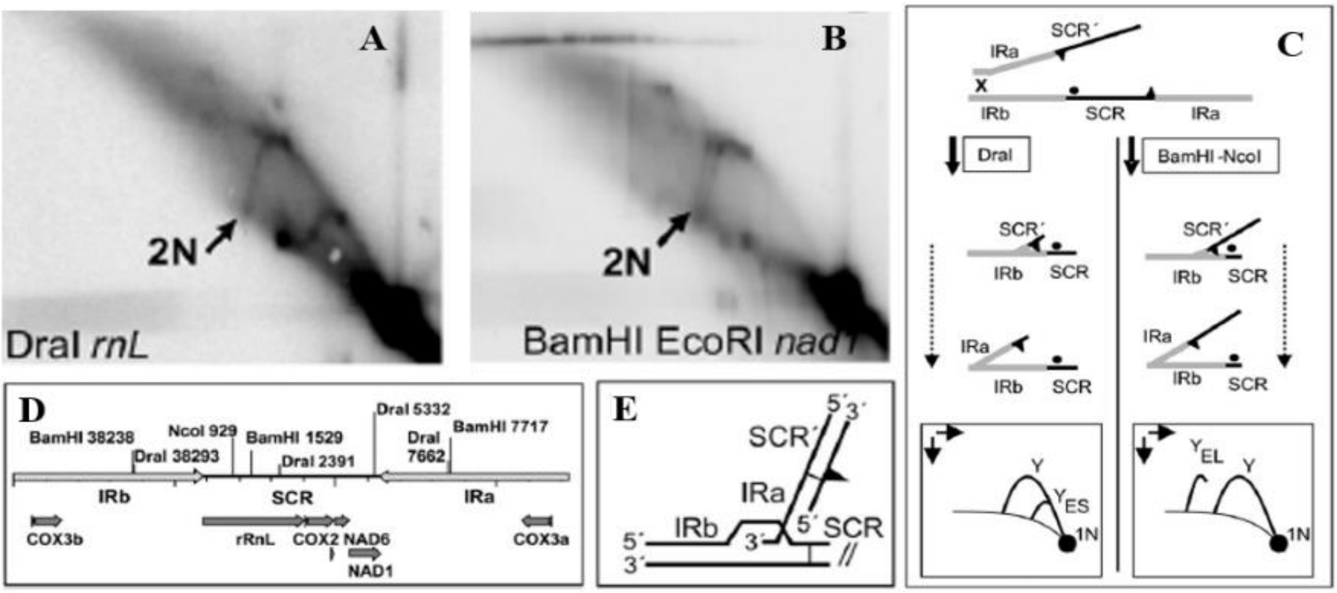


Figure I7. ***Model of strand invasion in C albicans.*** *(**A** and **B**) 2D gel analyses of C. albicans mtDNA reveal the presence of an arc (2N) corresponding to the regular Y structures indicative of recombination intermediates. (**C**) Interpretation of the arcs observed in the 2D gels. (**C** and **D**) Restriction enzyme digestion at different points of the SCR and hybridization with a fluorescent probe (solid dot) reveal the presence of Extra-long Y structures (Y_{EL}) and extra-short Y structures (Y_{ES}). (**E**) RDR model in which both Y_{EL} and Y_{ES} structures are formed via strand invasion events.* Extracted from: [66] Gerhold, Joachim M, et al. 2010, Molecular Cell,.

HMG proteins and DNA organization

High-Mobility Group (HMG) DNA binding domains are ubiquitous, present in both the nucleus and organelles and with a remarkable range of functions [68] [69]. The name High-Mobility Group stands for the high electrophoretic mobility of the protein in SDS-PAGE. HMG proteins are classified in three families, namely HMG-A, HMG-N and HMG-B. HMG-A, formerly referred to as HMG-I/Y, are highly flexible proteins that harbour an AT-hook domain that contacts DNA through the minor groove, see **Figure I8A** [70]. They are proposed to act as interaction hubs in molecular pathways related to transcriptional regulation, DNA repair, chromatin remodelling and RNA processing [71]. Proteins belonging to the HMG-N family possess a nucleosome binding domain through which they contact chromatin with higher affinity than naked DNA, see **Figure I8B** [72]. HMG-N proteins have been related to chromatin decompaction and transcription activation only in chromatin and not in naked DNA. Such an activity is mediated by a negatively charged C-terminal tail [68], [73], [74].

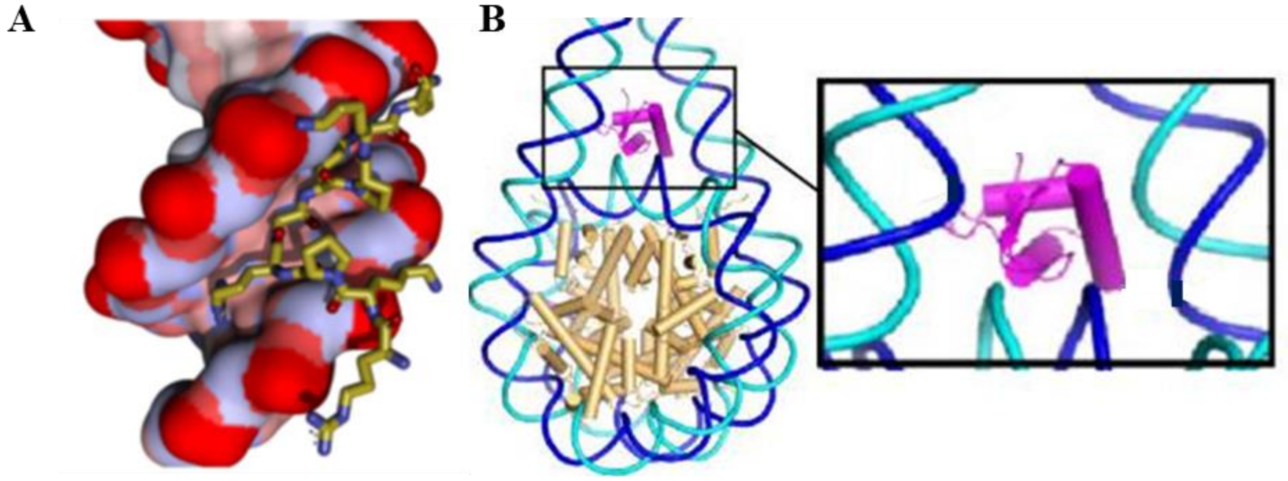
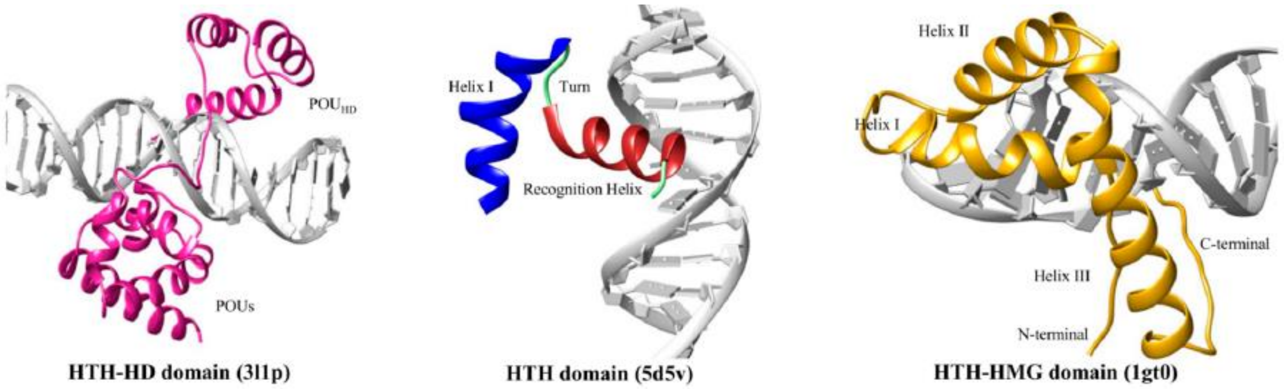


Figure I8. structures of HMGA and HMGN. *(A) Crystal structure of an HMGA, showing the AT hook interacting with DNA. Residues PRGP from HMGA1 show intimate association with the minor groove of the DNA substrate. Two water molecules (not shown in this figure) contribute to establish the structure via hydrogen bonding between atoms belonging to the peptide chain.* Extracted from: [70] (Fonfría-Subirós, et al., 2012*). (**B**)Model of an HMGN protein. These proteins, containing nucleosome binding domains interact with DNA when this latter is already bound and distorted by the binding of a histone octet. Note that this is not a crytal structure, HMGN position was inferred from cross-linking results. Tue tubes symbolize the α helices, which are connected by thin loops.* Extracted from: [72] Zhu, Nan and Hansen, Ulla. 2010, Biochimica et Biophysica Acta (BBA) Gene Regulatory Mechanisms.

HMGB protein family

The HMGB proteins constitute the largest family within the HMG group [69]. The HMGB proteins possess a HMG-box DNA binding domain, and together with homeodomains (HD-HTH), and the initially defined HTH binding domains, they have all been proposed to belong to the Helix-Turn-Helix (HTH) motif superfamily, which represent one of the various forms of DNA recognition by proteins [75] (see **Figure I9**).



HTH-HD domain (3l1p) **HTH domain (5d5v)** **HTH-HMG domain (1gt0)**

Figure I9. DNA recognition mechanisms by HTH domains. *Crystal structures of protein:DNA complexes formed with HTH-Homeodomain (HTH-HD left), HTH domain (middle) and HTH-HMGB domain. The different PDB entry codes correspond to: Mus musculus POU transcription factor domain (3l1p) [191], Human heat shock transcription factor 1 (Hsf1) (5d5v) [190], Sox2 from Homo sapiens (1gt0) [189].* Figure extracted and adapted from: *[75].* Yesudhas, Dhanusha, et al. 2017, Genes.

Specifically, HMG-box domains are characterized by a conserved sequence of approximately 75 amino acids [69]. The HMG-box domains are found throughout eukaryotes, both in nuclear and organellar proteins, and are present in transcription factors and chromosomal proteins, and have been classified in two broad subfamilies based on phylogenetic analysis, i.e. the MATA/TCF/SOX and the UBF/HMG families [76]. The name of the subfamilies are related to the following proteins: yeast mating genes (MATA), T-cell transcription factors (TCF) and Sry-related genes (SOX) for MATA/TCF/SOX; and nucleolar transcription factor 1 (UBTF-1) and high mobility group box 1 (HMGB1) for UBF/HMG subfamily.

The first structure of an HMG-box domain was determined by NMR from the major chromatin-associated non-histone protein HMGB1 [77]. This type of domains consists of three characteristic α-helices (helix 1, helix 2 and helix 3) that, together with an N-terminal extended region, fold in an L-shape (**Figure I10A**). Specifically, the extended N-terminal region packs against helix 3, forming the long arm of the L shape (note that in the NMR structure the N-terminal extended region and helix 3 show considerable concerted mobility, Figure **I10A**). Between these two segments, helix 1 and helix 2 form a small antiparallel coiled coil which corresponds to the short L-arm. The HMG-box domain structure is stable, as it is found in the different HMGB proteins that have been crystallized both in the absence and presence of DNA [78], [79], [80], [81] [78], [82], [83], [84], [39], [40], [42], [41]. The crystal structures of these proteins show that the characteristic L-shape of an HMG-box domain recognizes the DNA minor groove and binds to it by its concave surface. Helix 1 and 2 separate the strands like a wedge and imposes a bend to the contacted DNA region of by approximately 90º (see **Figure I9**, right panel). This strong bend is facilitated by the insertion of a residue, which is generally hydrophobic such as Met, Leu or Phe, between two base pairs that, therefore, lose their stacking interactions. With the widening of the minor groove and the insertion of the residue, the DNA becomes less rigid and more deformable, and a local unwinding

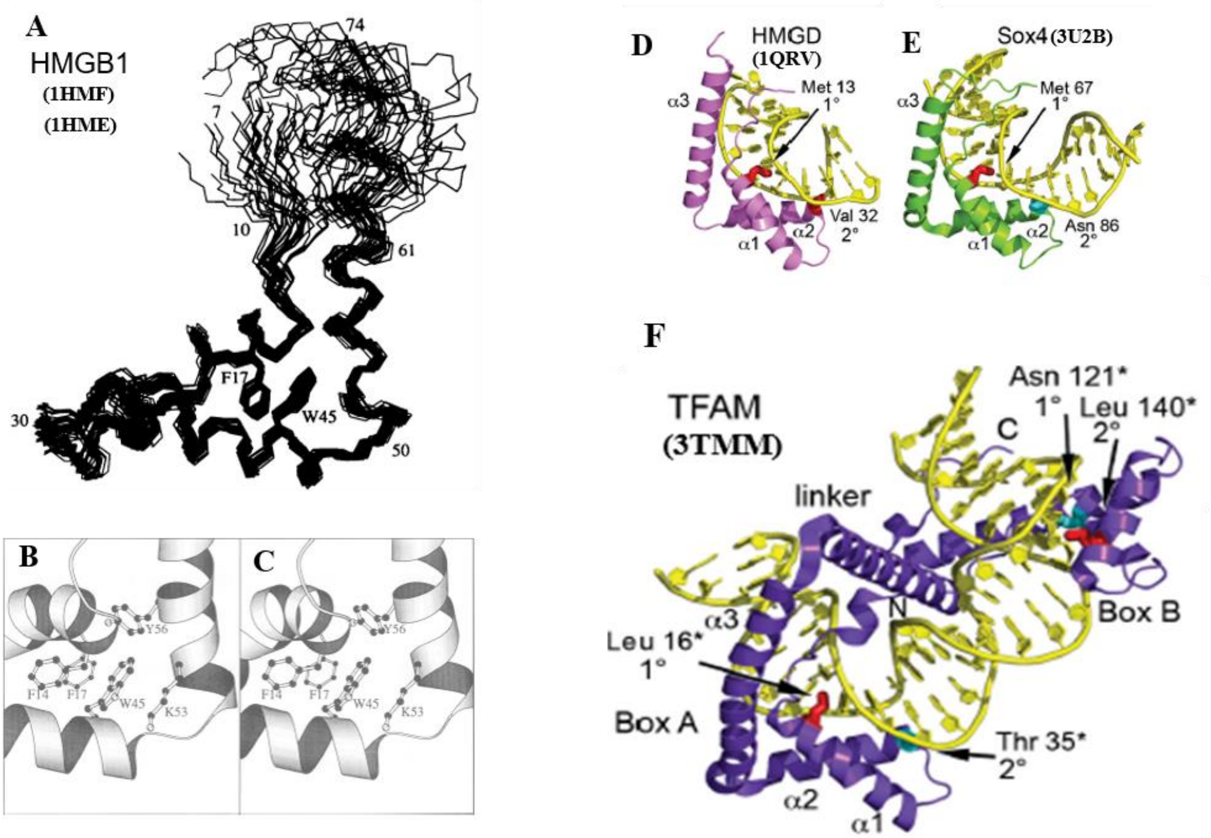


Figure I10. The structure of HMG domains in the absence or presence of DNA. *In* ***(A)****, NMR ensemble of models of the HMG domain B of HMGB1 protein, obtained by NMR analysis in solution.* ***(B*** *and* ***C****) Stereo view of the HMG-box hydrophobic core [77].* ***(D)*** *Structure of the 'Drosophila melanogaster' HMG-box protein HMGD. In red, the Met13 side chain that inserts between DNA bp, breaks the stacking and induces a dramatic DNA distortion of 90ᶜ as shown. The different helices 1, 2 and 3 are indicated [192].* ***(E)*** *'Mus musculus' Sox4 HMG domain is shown in green, and side chains depicted as in (D) [83].-****(F)*** *'Homo sapiens' mitochondrial transcription factor A (TFAM) [40]. (PDB entries are listed for each structure).* Figure extracted and adapted from *[77]*. Weir, Hazel M, et al. 1993, The EMBO Journal and *[69]* Malarkey, Christopher S and Churchill, Mair EA. 2012, Trends in biochemical sciences.

occurs reflected by the decreasing of the twist angle at the inserted base-step, allowing for the overall 90º bend [39], [40], [42], [41], [84].

HMGB proteins may contain single or multiple HMG-boxes. Tandem HMG-box protein (containing two of them) include nuclear proteins such as HMGB-1 (*Homo sapiens*) and mitochondrial proteins including TFAM (*Homo sapiens*), Abf2p (*Saccharomyces cerevisiae*). Tandem HMG-boxes usually contain disordered regions, therefore its structural characterization have only been possible in complex with DNA substrate [39], [40], [42].

TFAM and mitochondrial DNA maintenance in human

The human Mitochondrial Transcription Factor A was the first mtDNA packaging protein to be characterized structurally by X-ray crystallography [39], [40]. TFAM is a flexible protein containing two HMG-boxes connected with a linker that is intrinsically disordered in absence of DNA. At each promoter, TFAM binds to 22bp by using both HMG-boxes, contacting two DNA regions separated by one DNA turn (10 bp). Each HMG-box induces a bend to the DNA of almost 90º, which results in an overall 180º distortion, a DNA U-turn (**Figure I11**). In such a U-turn arrangement, the two HMG-boxes are at different sides of the DNA and the linker connects them by traversing the inner face of the U-turn, the linker is intertwined around the DNA requiring a conformational rearrangement of TFAM. To achieve intertwining of TFAM and DNA, a step-wise binding of the different domains to the DNA occurs. Indeed, single-molecule FRET studies [85] confirmed that TFAM binds to DNA binding commences by HMG-box 1. Upon DNA binding, the linker between the two HMG-boxes folds into a α-helix conformation and establishes salt bridge interactions with the phosphate backbone [86], [39], [85]. Finally, the second box, which has much lower binding affinity for DNA than HMG-box 1 [86], [87], binds and bends the DNA, thus TFAM imposes the U-turn. By virtue of the U-turn, HMG-box 2 and the C-terminal tail are placed close to the transcription initiation site, where they both recruit the RNA polymerase [88] (see **Figure I12**). In addition to its transcription activation role, TFAM compacts the mtDNA by mechanisms of binding and bending that probably, but not necessarily, involve the formation of U-turns. It has been proposed that TFAM compacts the mitochondrial nucleoid by cross-strand binding to DNA, which is conceivable if the two HMG-boxes bind two DNAs independently [31] (**Figure I13**). In addition, TFAM enhances DNA flexibility by local melting, which has also been proposed as a mechanism for mitochondrial DNA compaction [89].

Intra-mitochondrial TFAM levels determine the degree of mtDNA compaction, which also influences different events of mtDNA metabolism. Low levels activate replication and transcription and induce a milder DNA compaction, whereas high TFAM levels inhibit mtDNA transcription and replication due to an excess of compaction that precludes the proper functioning of the transcription and replication machineries [89] [90] (**Figure I14**).

Thus, the absence of TFAM causes the loss of mtDNA and the mitochondrial function, low levels of TFAM contribute to mtDNA maintenance and high of TFAM levels decrease mtDNA copy number [31] [91] [92].

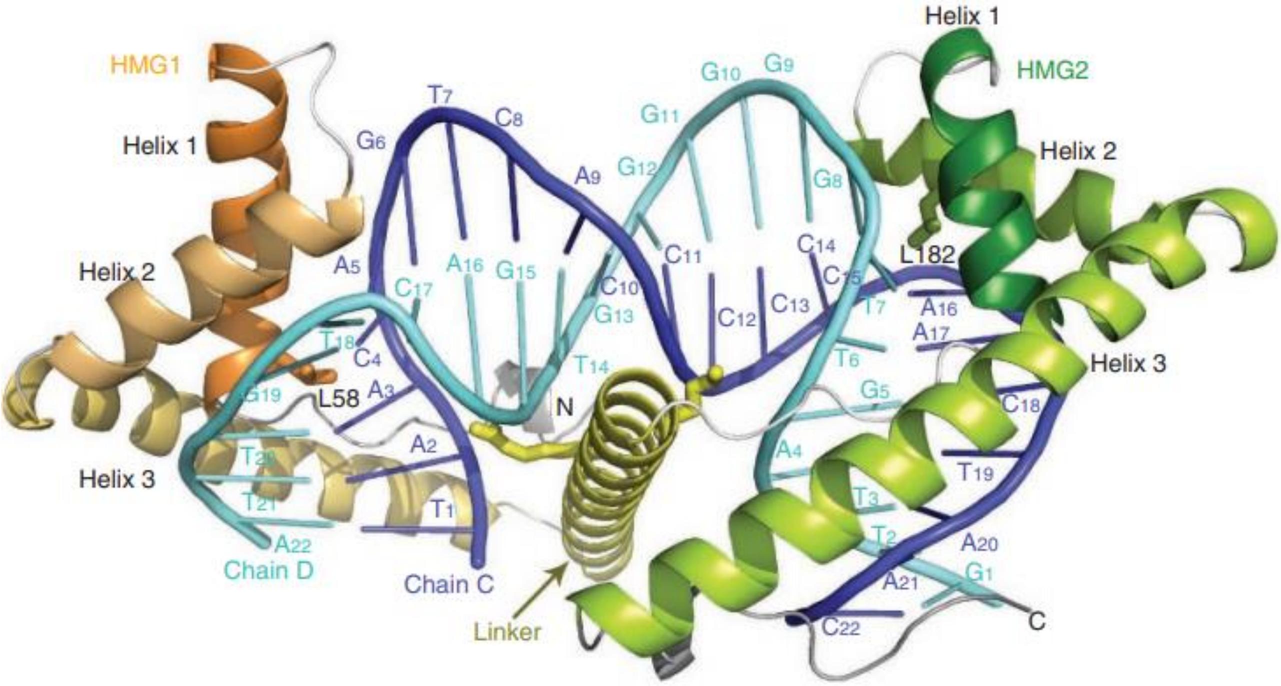


Figure I11, TFAM imposes a U-turn on the mitochondrial Light Strand Promoter. HMG-box 1 (HMG1), HMG-box 2 (HMG2) and the linker (Linker) are labelled and depicted in orange, green and yellow, respectively. Helix 1,2 and 3 are indicated for each HMG-box. Both DNA chains are shown in dark and light blue (Chain C and D, respectively), and the base pairs numbered. The DNA inserting residues Leu58 (L58) and Leu 182 (L182), and two arginines from the linker that contact the DNA are shown as sticks. The N- and C-terminal ends of the protein structure are indicated (N and C, respectively). Extracted from: *[39]*. Rubio-Cosials, Anna, et al. 2011, Nature Structure and Molecular Biology.

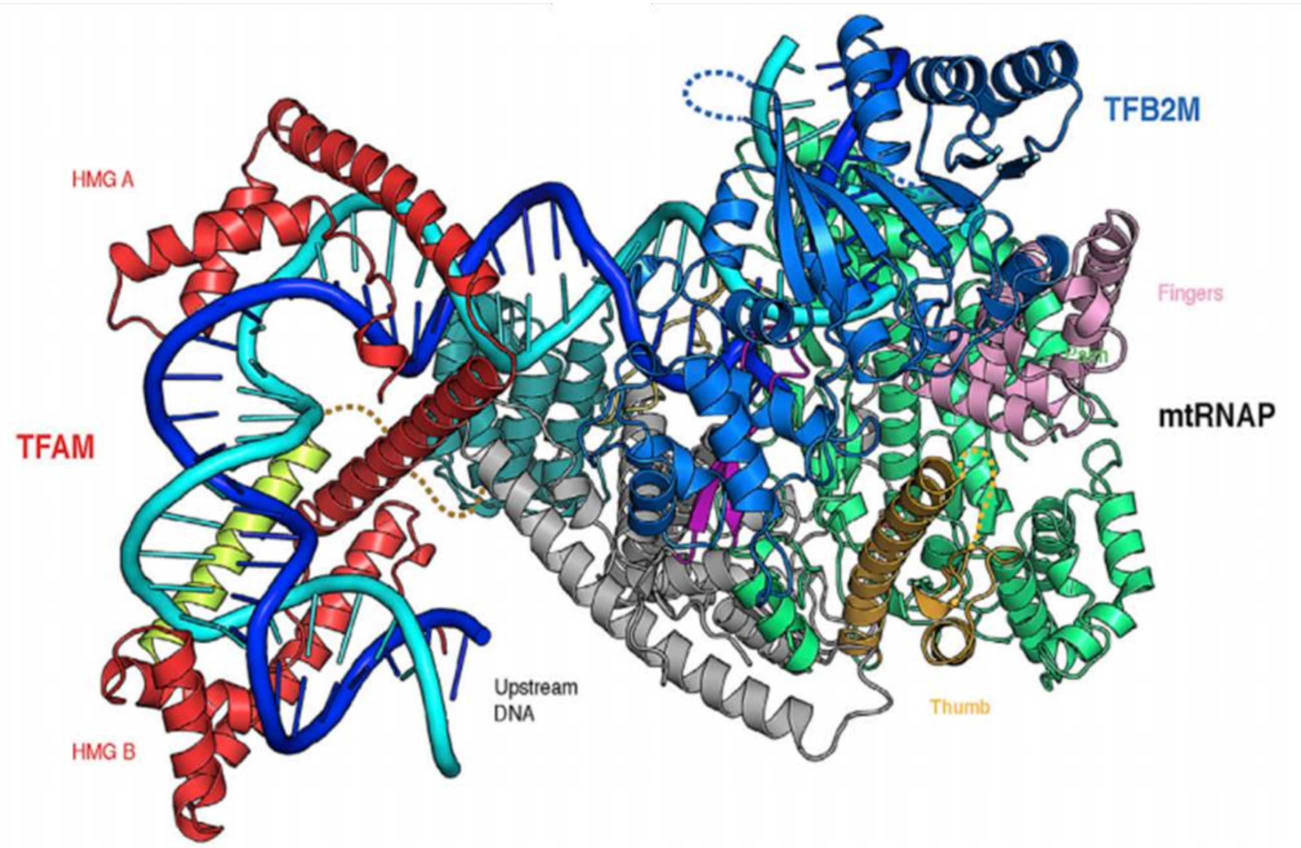


Figure I12, TFAM within the cryo-EM structure of mitochondrial transcription initiation complex. *TFAM (red) binding and bending of DNA have been proposed to locate the core transcription complex in position to initiate transcription. C-terminal tail of TFAM would be the one in charge of interacting with the main proteins of the transcription complex.* Extracted from: *[88]*. Hillen, Hauke S, Temiakov, Dmitry and Cramer, Patrick. 2018, Nature structural and molecular biology.

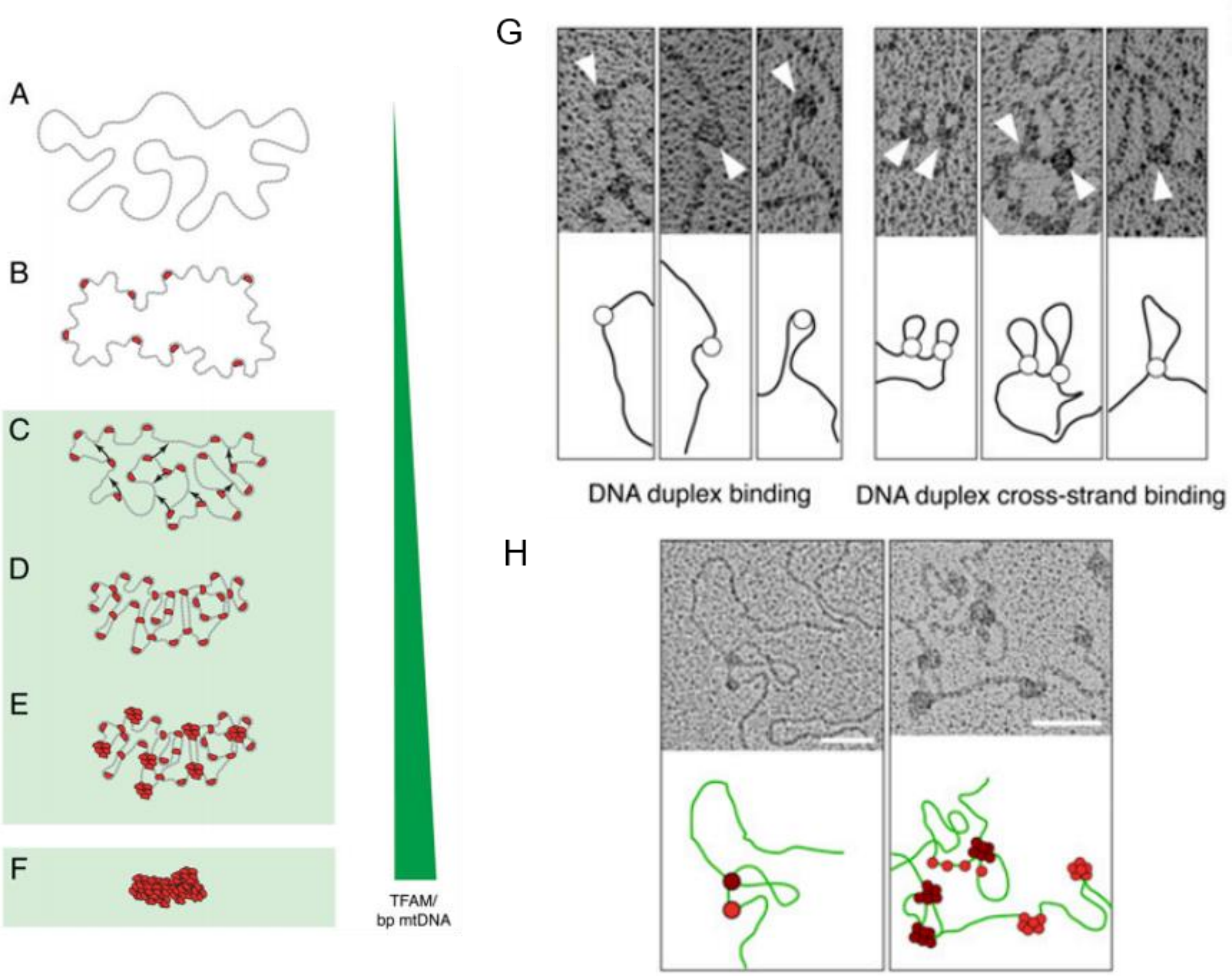


Figure I13. model for the compaction of mitochondrial DNA by TFAM via cross-strand binding. *According to this model, DNA compaction is driven by increasing amounts of TFAM. Mechanisms of binding to a protein-free DNA molecule **(A)** include simple binding **(B)**, cross-strand binding **(C)**, TFAM dimerization **(D)**, aggregation or multimerization **(E)** and collapse of the structure in a nucleoid-like structure **(F)**. **(G)** Interpretation of transmission electron microscopy images showing the binding of TFAM to long DNA duplexes, and the induction of cross-strands. **(H)** Interpretation of transmission electron microscopy images showing TFAM clusters on DNA.* Extracted and adapted from: [31]. Kukat, Christian, et al. 2015, PNAS.

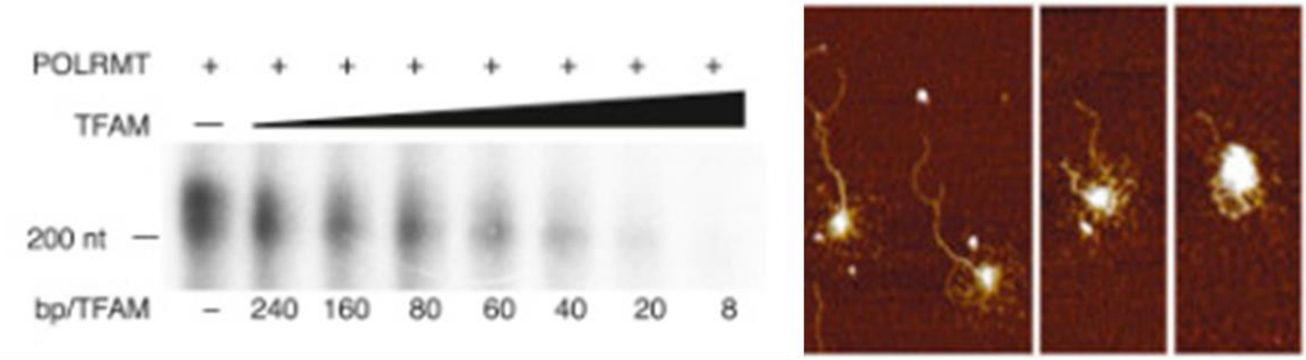


Figure I14. TFAM induced DNA compaction and transcription inhibition. *(A) Effect of increasing TFAM concentration in promoter -independent transcription from a DNA substrate with free 3' tails. (**B**) Densely TFAM-packed structures suggest that TFAM-induced compaction can block the access of polymerase to the DNA template.* Extracted from *[90]*. Farge, Géraldine, et al. 2014, Cell Reports.

Abf2p and mtDNA maintenance in Saccharomyces cerevisiae

Abf2p (ARS (*autonomously replicating sequence*)-binding factor 2 protein) is the mtDNA compacting protein from *Saccharomyces cerevisiae*. As TFAM, Abf2p was extensively characterised in our laboratory [42]. Like TFAM, Abf2p possesses two HMG-boxes (**Figure I15B**). However, in Abf2p the boxes are additionally preceded by an N-terminal helix of 3 turns separated by a very short linker of 2 residues. Furthermore, the protein does not have a C-terminal tail, causing a lack of transcription activation ability [93]. Analogously to TFAM (**Figure I15A**), each HMG-box of Abf2p inserts a residue to the contacted DNA region, which bends by 90º, and thus this protein also imposes an overall U-turn to the DNA (**Figure I15B**). The short linker of Abf2p, impairs intertwining of the protein around the DNA, typical for TFAM binding (**Figure I15A**). Alternatively, Abf2p induces an overall U-turn by binding at one side of the DNA molecule, as a 'staple' (**Figure I15B**). In the absence of DNA Abf2p shows high flexibility in solution, adopting an extended conformation that becomes more compact and less flexible upon DNA binding [42]. The higher rigidity of the 'staple' conformation is conferred by a hydrophobic core between distant protein regions that comprise the short N-terminal helix, helix 3 from HMG-box 1 and the linker region between HMG-boxes, which stabilizes the complex (**Figure I15B**). Abf2p was crystallized with DNA containing symmetric (A_3T_2) and asymmetric (A_4) adenine tracts (A-tracts) (PDB codes 5JH0 and 5JGH, respectively). A-tracts show increased stiffness and a narrowed minor groove [94], and are related to sequence-dependent DNA curvature [95], [96], [97] A-tracts are more abundant in non-coding regions, and its presence has been related to nucleosome positioning and protein-induced DNA structures such as DNA looping [98]. As mentioned above, HMG-boxes bind to the minor groove, and during Abf2p crystallization the DNA A-tracts excluded binding of Abf2p, which contacted other regions of the crystallization oligos, thus positioning the protein at the same DNA site in all protein/DNA complexes, which crystallized. Unlike TFAM, each Abf2p binds to two different DNA molecules. This was the first time that a HMG-box protein was crystallized bound to two different DNA substrates, albeit this could be a feature induced by the sequence of the DNAs used [42].

Abf2p was found to display a phased-binding in the ARS-1 (Autonomously Replicating Sequence) of *S. cerevisiae* nuclear DNA, a region with a high rate of replication initiation events [99]. Abf2p packages mtDNA in nucleoid-like structures, albeit inducing a looser packaging than TFAM [100]. *S. cerevisiae* can grow both in fermentable (anaerobic) and non-fermentable (aerobic) conditions. Under fermentable conditions (in the presence of glucose), mtDNA is not essential for *S. cerevisiae* survival since the encoded OXPHOS subunits are not required. In such conditions, deletion of Abf2p does not affect survival of yeast cells but hampers mtDNA maintenance and leads to mtDNA loss [101]. Therefore, loss of Abf2p (thus, loss of mtDNA) makes *S. cerevisiae* unfit for transitioning back to non-fermentable media (e.g. glycerol) [99]. In yeast containing mtDNA, during the transition to non-fermentable media, the Abf2p:mtDNA ratio is diminished, thus favouring a looser mtDNA packing and an increased rate of the expression of ETC complexes [99]. Abf2p is thus related to mtDNA expression and replication as it regulates appropriate DNA compaction levels. Abf2p HMG-box 1 has a greater DNA binding efficiency than HMG-box 2, and it has been shown to be sufficient for the maintenance of mtDNA and for the *S. cerevisiae* ability to transit from a fermentable to non-fermentable media [42]. Abf2p has been characterized to bind Holliday junctions *in vitro* [102] and to stabilize mitochondrial DNA recombination intermediates *in vivo* [103].

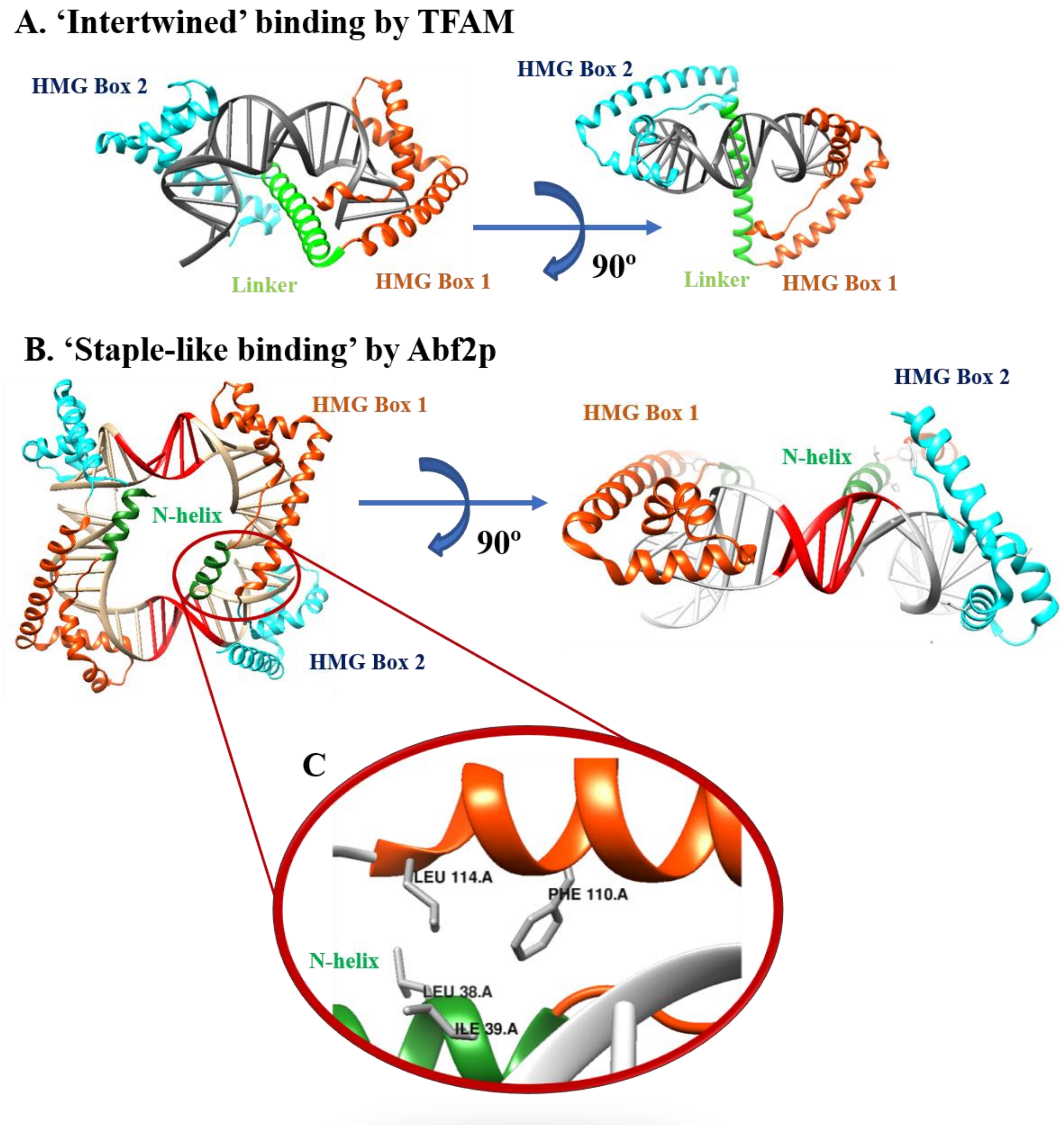


Figure I15. Different binding mechanisms of the mtDNA packaging factors TFAM from human, and Abf2p from S. cerevisiae. (A) *TFAM wraps DNA by means of the helical linker, and both molecules intertwine (PDB.ID:3TQ6) [39].* ***(B)*** *By the 'staple' binding mode of Abf2p, the protein contacts one side of the DNA. This interaction is stabilized by the hydrophobic core formed by the N-terminal helix (N-term), the HMG-box 1 helix 3 and the linker (PDB.ID:5jh0). [42].* ***(C)*** *The DNA bases colored in red correspond to the adenine tract (A-tract) present in the DNA molecule.*

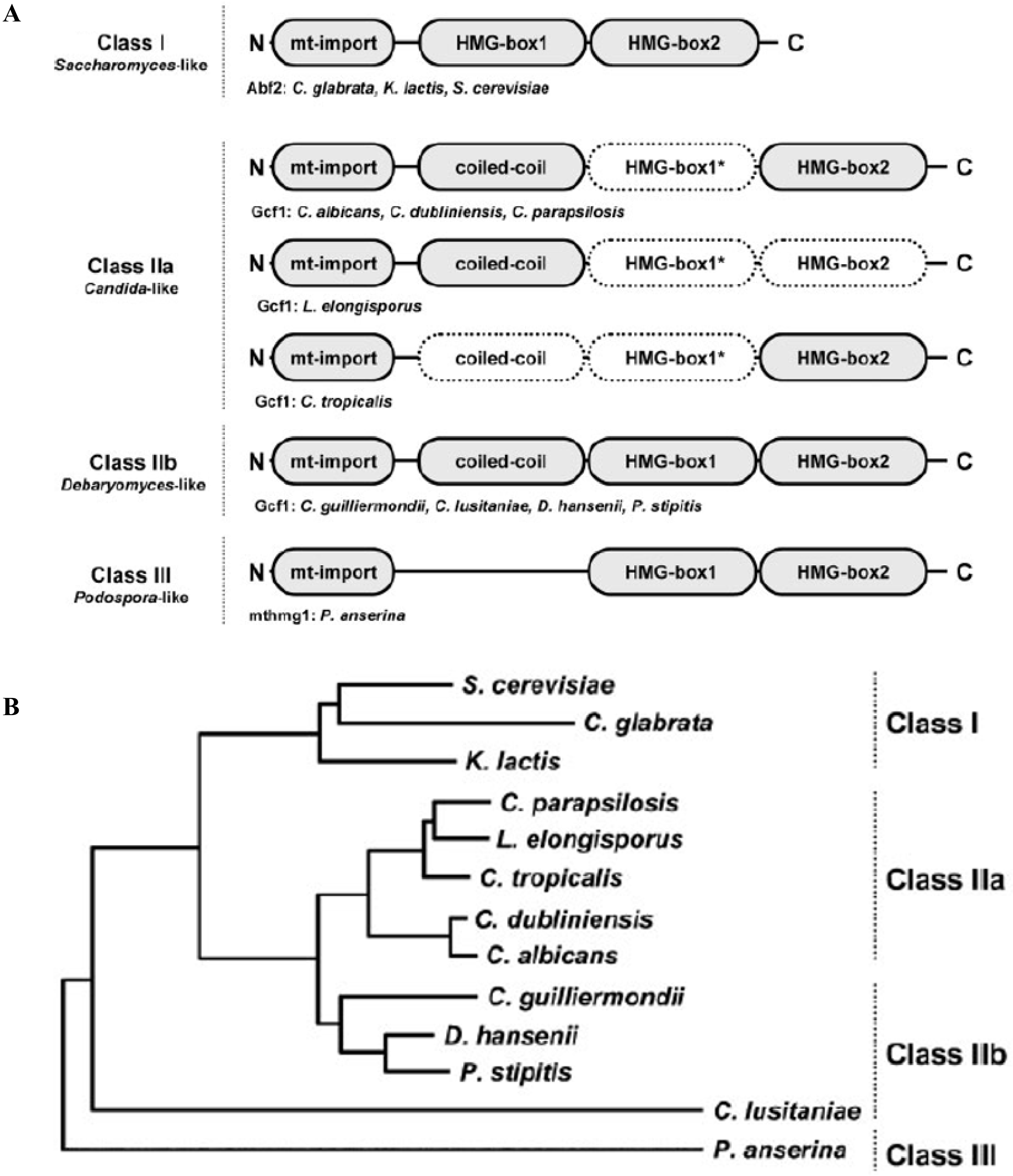


Figure I16, domain prediction and evolutionary relationship between mtDNA maintenance proteins in yeast. ***(A)*** *Prediction of domains in different yeast mitochondrial DNA packaging factors (asterisks on HMG-box1 highlight that the presence of this domain is dubious.* **(B)** *Phylogenetic dendrogram based on mtDNA packaging proteins.* Extracted from: *[104]*. Visacka, Katarina, et al. 2009, Microbiology

Gcf1p, an unconventional mitochondrial HMG-box protein from *Candida albicans*

C. albicans mtDNA is a research subject of high interest, not only due to the high pathogenicity of the microorganism but also due to its mtDNA replication mechanism, which depends on recombination [66], and might provide further insights in mitochondria divergent evolution within eukaryotes [65]. Mitochondrial DNA packaging factor in *Candida albicans* was identified relatively recently, as the product of the ORF19.400/19.8030, assigned as GCF-1 [104]. Gcf1p, the gene product of GCF-1, has been predicted to contain a putative N-terminal coiled-coil domain and only one HMG-box close to its C-terminus. The presence of another HMG-box between these two domains was discussed but although three α-helices were present, it was conclusive as to be related to HMG-box.

Furthermore, *in silico* analyses allowed to draw an evolutionary dendrogram containing the mitochondrial DNA factors of yeast, highlighting the great evolutionary divergence of mitochondrial HMG-box proteins in yeast. Based on the number of HMG-boxes and the presence of a Coiled Coil domain, a classification for yeast mitochondrial HMG was proposed indicating that Gcf1p and other *Candida*-like other mtDNA packaging factors belong to a different phylogenetic branch than that of *Saccharomyces*-like Abf2p (see **Figure I16**).

Gcf1p colocalizes with mtDNA both in endogenous (*Candida albicans*) and heterologous expression (*Saccharomyces cerevisiae*) thus confirming that Gcf1p is associated with mtDNA (see **Figure I17**) and suggesting a similar function to that of Abf2p. Gcf1p overexpression indeed compensates the phenotype induced by Abf2p deletion in *S. cerevisiae* [104].

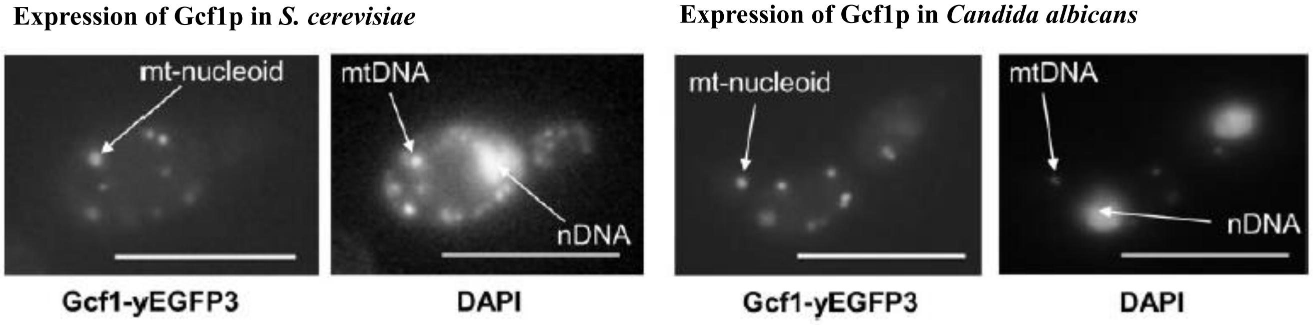

Expression of Gcf1p in S. cerevisiae
Expression of Gcf1p in Candida albicans
mt-nucleoid
mtDNA
nDNA
mt-nucleoid
mtDNA
nDNA
Gcf1-yEGFP3
DAPI
Gcf1-yEGFP3
DAPI

Figure I17, Gcf1p colocalizes with mtDNA when expressed as both endogenous and heterologous protein. Extracted from: *[104]*. Visacka, Katarina, et al. 2009, Microbiology

Importantly, double-allele deletion of the GCF1 gene results in non-viable *C. albicans* cells, indicating that Gcf1p performs an essential function in the organism [104]. Indeed, this is the case. Conditional repression of Gcf1p expression in cells with a single-allele GCF-1 deletion results in a 3200-fold decrease in the number of messenger RNA in mitochondria (mRNA) (**Figure I18A**), as well as a 80% decrease in mtDNA copy number (**Figure I18B**) and a three to five-fold slowed cell cycle [104]. Analysis of the presence of DNA recombination intermediates by bi-dimensional agarose gel electrophoresis of *C. albicans* mtDNA revealed that repression or absence of Gcf1p expression correlate with lower levels of X-DNA structures, i.e. Holliday junction-like recombination intermediates (see **Figure I18C**, **D** and **E**). The close relationship between recombination events and replication initiation in *Candida* mtDNA suggest that Gcf1p is related to mtDNA copy number via the stabilization of recombination intermediates [65], [66], [67]. Similar evidences have been reported for *S. cerevisiae*, Abf2p binds and stabilizes recombination intermediates [102], [103], suggesting a general mechanism for a recombination-based replication of mtDNA in yeast, in which DNA packaging HMG proteins, such as Abf2p and Gcf1p, would play an essential role [65] [62].

Comparative analysis by DNA gel-shift assay using *Candida parapsilosis* Gcf1p (CpGcf1p) *Saccharomyces cerevisiae* Abf2p (ScAbf2p) and *Yarrowia lypolitica* Mhb1p (YlMhb1p) [102] revealed different DNA binding properties (see **Figure I19**). The three protein yielded a clear shift on the gel when binding Holliday-junctions, nevertheless its behavior differed when binding linear double-stranded DNA substrates. Whilst a clear shift was observed when binding DNA substrates longer than 50 bp, no shifted bands were obtained when substrates of 25 bp were used.

The molecular mechanisms underlying Gcf1p binding to DNA are not yet known. The presence or absence of an HMG-box close to the N-terminus has not been confirmed, nor is the function of Gcf1p-specific predicted N-terminal helix. *C. albicans* has biomedical interest and Gcf1p is essential for the *C. albicans* survival, but its targeting requires detailed studies at the atomic level to discern potential similarities and, more importantly, differential features that would allow for specific treatments against the microorganism without hampering the activity of its human counterpart, TFAM.

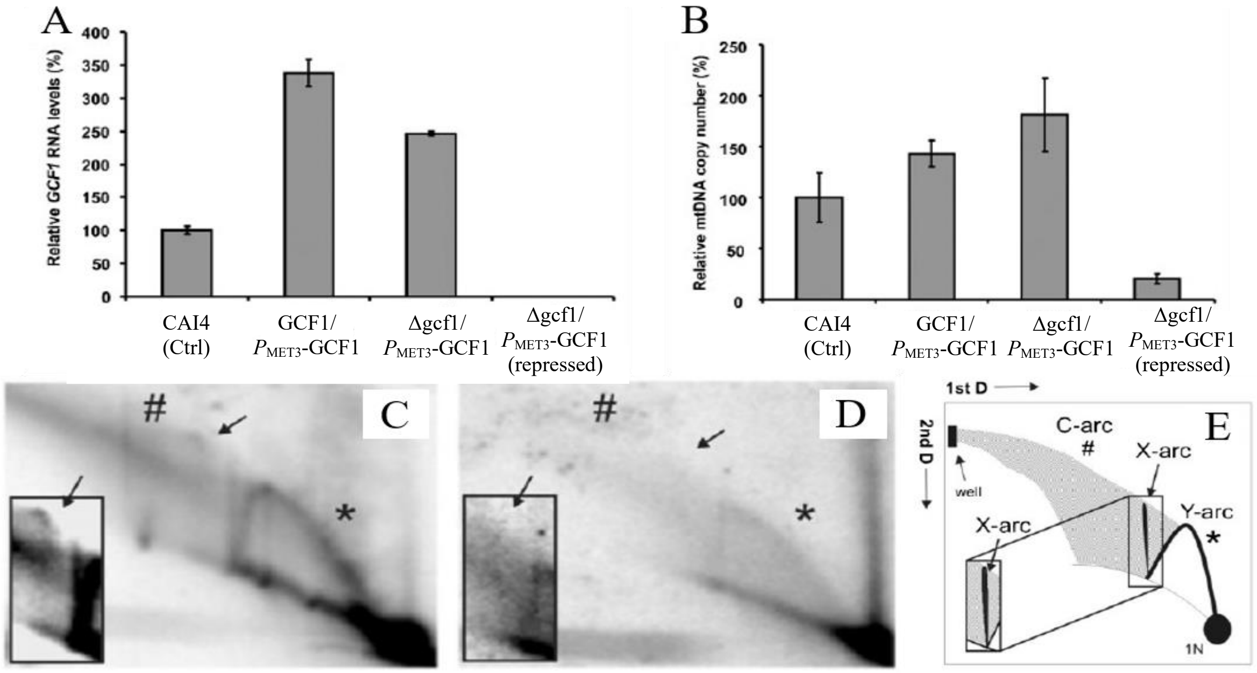


Figure I18, Gcf1p levels correlate with mRNA, mtDNA levels and recombination intermediate levels in Candida albicans. *CAI4 corresponds to the control strain, P*$_{\mathrm{MET3}}$ *-GCF1 corresponds to a conditionally repressible allele and Δgcf1 to a single-allele deletion. In (A) and (B) levels of mRNA and mtDNA copy number respectively are significantly decreased upon GCF1 repression as compared to the un-repressed culture. (C) and (D) correspond to the bi-dimensional agarose gel electrophoresis of the un-repressed and the re-pressed Gcf1p expression, respectively. In (E), an interpretation of the DNA migration indicates that the X-arc (corresponding to Holliday junction-like recombination intermediates) is greatly diminished upon Gcf1p repression*. Extracted from: *[104]*. Visacka, Katarina, et al. 2009, Microbiology

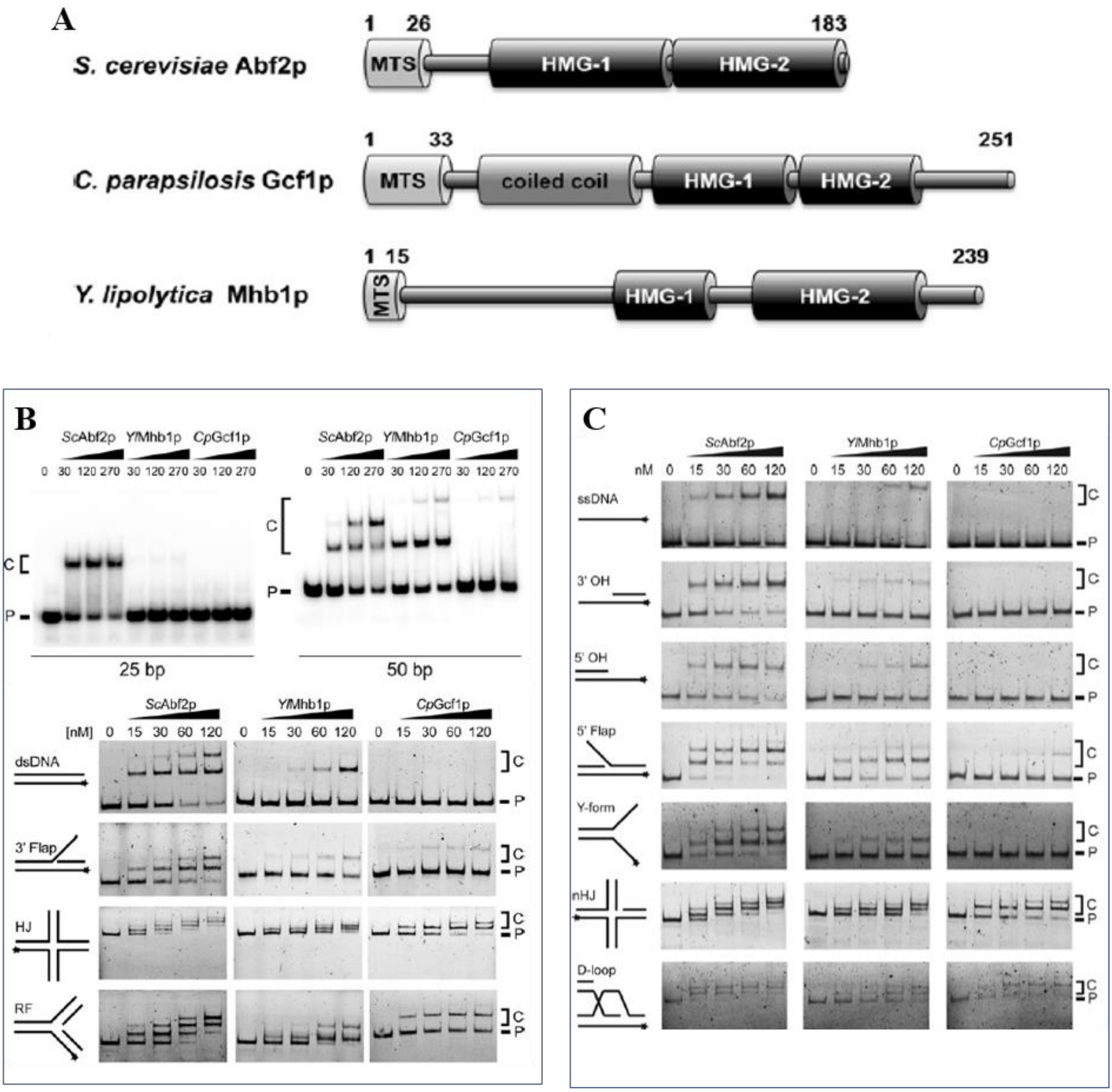


__Figure I19, DNA binding analysis to ScAbf2p, CpGcf1p and YlMhb1p.__ The three proteins included in the assay (A) with its different domains (note that Candida parapsilosis Gcf1p has the same predicted structure than Candida albicans Gcf1p, authors explain in the text that prediction programs failed to identify HMG-box 1 in CpGcf1p although HMG-box1 is depicted based on available orthologues). EMSA results (B) and (C) show the electrophoretic mobility of complex formed at an increasing protein nM concentration keeping a constant DNA concentration of 3 nM. Extracted and adapted from: *[102]*. Bakkaiova, Jana, et al. 2016, Bioscience Reports.

MATERIALS AND METHODS

M1. SEQUENCE ANALYSIS AND STRUCTURE PREDICTION

Analysis of the constructs as well as the design of the oligonucleotides for PCR reactions was performed using SnapGene Viewer software [105].

Structure prediction, search for evolutionary homologs and homology modelling softwares used, were the ones included in the toolkit from Max Planck Tuebingen Institute webpage [106] [107]. Search for homologs was performed using HHPred, sequence alignment with Clustal Ω and homology modelling with Modeller. In parallel to this, predictions of coiled coil forming sequences was performed using HMM algorithms with MARCOIL. Prediction of disordered regions was performed using the independent server IUPred2A [108]. Domain boundaries for the design of constructs was performed using the pfam database [109] and the secondary structure prediction server JPred4 [110].

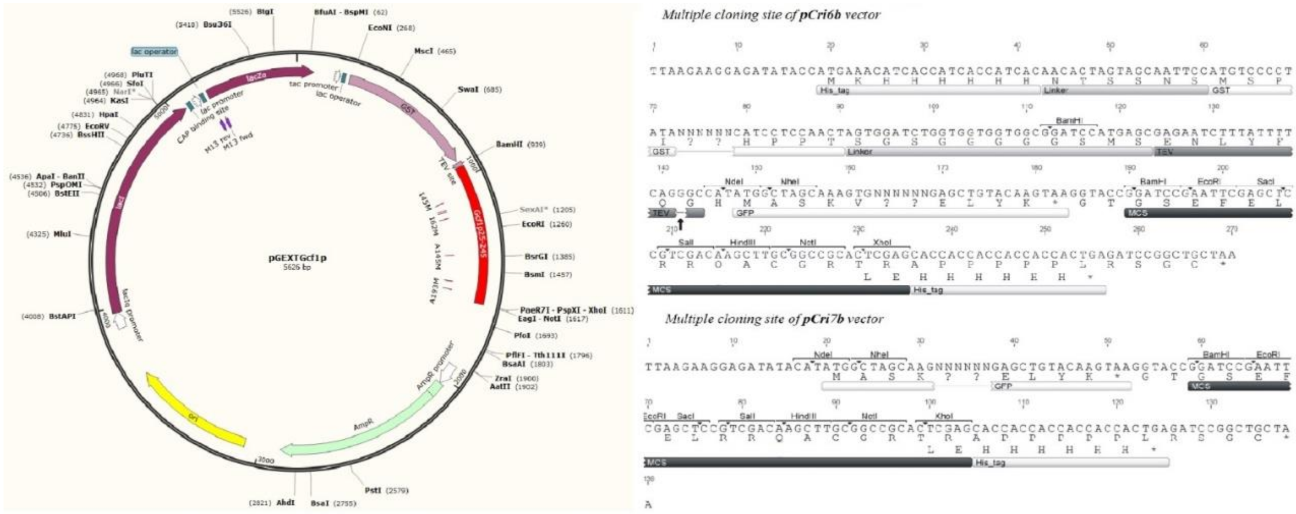


Figure M1: *Schematic map of the pGEXT-4T1 vector that contained the Gcf1p construct, from which all the constructs were amplified and cloned in the expression vectors used in this project (left panel, taken from the software SnapGene Viewer). Schematic representation of the multiple cloning sites of the expression vectors pCri6b and pCri7b (right panel [112, pp. 17 and 18, Supplementary material]*

M2. GENERATION OF CLONES BY MOLECULAR BIOLOGY METHODS: CLONING AND SITE-DIRECTED MUTAGENESIS

The starting material for this project was the original construct used for heterologous expression of Gcf1p in *Escherichia coli*, kindly provided by collaborators Dr. Joachim Gerhold and Prof. Juhan Sedman [104] [102]. The collaborators cloned this construct into a pGEXT-4T1 vector, in which the expression of the gene of interest is under control of the *Escherichia coli* endogenous lac promoter. The construct lacked the Mitochondrial Targeting Sequence (MTS) and included the nucleotide positions encoding for corresponding to aminoacids 25 to 245. Specifically, the insert was cloned between BamHI and XhoI positions. The coding sequence was preceded by a Glutathione S-Transferase (GST) affinity tag intended for affinity purification. Both the affinity tag and coding sequence were connected through a 6-residue long TEV proteolytic site (ENLYFQ) substituting the original thrombin site of pGEX4T1 [111] as it is depicted in **Figure M1**.

M2.1 Preparation of Gcf1p deletions and single site mutants

For all deletion mutants, the around-the-horn PCR was used with non-contiguous, non-overlapping DNA oligonucleotides. The single residue mutations were performed by using *Quickchange system* with self-complementary DNA oligonucleotides. All oligonucleotides used for PCRs are listed in the **Table M1**.

The polymerase used for whole-plasmid amplification was performed by using *Phusion polymerase* (ThermoScientific ®).

Whole-plasmid PCR protocol (20 cycles):

Initial denaturation: 98ºC for 2’

Denaturation: 98ºC for 20’’

Annealing: 55 to 72ºC for 30’’

Extension: 72ºC for 3’

Annealing temperature used by default was 5ºC inferior to the melting temperature calculated using SnapGene Viewer upon oligonucleotide sequence and length [105]. When needed, annealing temperature was screened in steps of 2ºC around the original value in order to optimize the yields of the reaction. Parental DNA digestion followed with *DpnI* restriction enzyme at 37ºC for 2h (ThermoScientific ®) which recognizes methylated adenines within GATC target sequences. The resulting whole plasmid PCR products were then purified by using the *GFX^{TM} PCR and gel band purification kit* (GE Healthcare ®) whose fundament is the reversible precipitation of soluble DNA within a glass fibre matrix. Amplified linear plasmids were subsequently used for chemical transformation of *DH5α* competent cells, using a thermal shock protocol.

Thermal shock transformation protocol

Thaw on ice 20µL of chemically competent DH5α cells
Addition of 1µL of DNA at 10ng/µL
Incubation in ice for 30'
Incubation at 42ºC for 45''
Incubation in ice for 5'
Addition of 980µL of sterile LB media
Incubation at 37ºC for 1h
Plate 150µL of the transformed cells in LB-Ampicillin media.
Grow over-night in a 37ºC incubator.

M2.2 Subcloning of gcf1p insert into pCri7b and pCri6b vectors

Full-length Gcf1p insert (residues 25 to 245 of the gene) was cloned into plasmid vectors pCri7b and pCri6b (Goulas, et al., 2014). For both plasmids, the restriction sites used were *NcoI* (CCATGG) and *XhoI* (CTCGAG) from the pCri system multiple-cloning site [112]. Thus, both sites were added to the 5' and 3' ends of the gcf1p construct, respectively, by PCR and using the oligos listed in **Table M1**. Insert amplification was performed with *KOD HotStart polymerase* (Novagen ®).

Insert amplification PCR protocol (20 cycles):

HotStart activation: 95ºC for 5'

Denaturation: 95ºC for 20''

Annealing: 55 to 72ºC for 30''

Extension: 72ºC for 1' and 30''

PCR products were purified using the *GFX^{TM} PCR and gel band purification kit* (GE-Healthcare ®). Following this, the inserts at a concentration of 50ng/µL were incubated with restriction enzymes *NcoI* and *XhoI* at a constant concentration of 0.1U/µL each for 2 hours at 37ºC, which generated the cohesive ends CCATGG and CTCGAG, respectively. Digestion of the chosen empty vectors was performed at a DNA concentration of 250ng/µL and the same restriction enzymes *NcoI* and *XhoI* at a concentration of 0.1U/µL each. Digestion of the vectors was also performed during 2 hours at 37ºC.

Plasmid DNA digestion mixtures were run in 0.7% agarose gel electrophoresis in 0.5 X TAE buffer (20mM Tris-Acetate pH 8.1, 0.5mM EDTA) at 90V for 45'. Open linear full-length molecules purified from the gel using the *GFX^{TM} PCR and gel band purification kit* (GE-Healthcare ®). Inserts were purified from the digestion mixture by using the same *GFX^{TM} PCR and gel band purification kit* (GE-Healthcare ®).

Ligation of inserts into plasmids consisted in the incubation of the insert with the plasmid DNA at a constant concentration of 5 ng/µL. The insert of interest was added at different vector-to-insert ratios (typically 1:3 and 1:5) for 2h at 16ºC in presence of *T4-DNA ligase* (Thermo). Ligation mixtures were directly used for the transformation of chemically competent DH5α cells following a thermal shock protocol (*see mutagenesis section,* **Materials and Methods 2.1)** and colonies were selected using kanamycin resistance.

Colonies obtained after DH5α transformation were checked for the presence of the insert by colony PCR using Taq polymerase (ThermoScientific ®). Plasmid DNA was extracted from positives colonies using a *QIAprep Spin MiniPrep Kit* (QIAGEN®), a kit for bacterial lysis and plasmidic DNA isolation based on the precipitation of genomic DNA and the reversible binding of plasmidic DNA onto a glass fibre matrix on a spin column. Purified plasmids were sequenced (GATC-Lightrun) and correct constructs were stored at -20ºC for downstream usage. DH5α clones containing the specific plasmids were likewise stored at -80ºC for downstream use.

Colony PCR protocol (30 cycles):

Initial denaturation: 95ºC for 2'

Denaturation: 95ºC for 20''

Annealing: 55 to 72ºC for 30''

Extension: 72ºC for 3'

Table M1: Summary of all the expression constructs used in this project.

Oligonucleotides used for every construct are written, accordingly to the conventions, in the 5' to 3' order.

Depicted in red are the nucleotides that are not present in the original pGEXT-4T1_Gcf1p.

ExoSite PCR for generation deletion constructs	
Construct name	**Oligonucleotides used**
pGEXT-Gcf1p_1-10 *(Residues 10 to 221)*	**Fw:** ACTAAAAAATCCACAACCAAAGC **Rv:** TCCTTGGAAATACAGGTTTTCCAGATCCGATTTTGGAGGATGGTCGCC
pGEXT-Gcf1p_1-18 *(Residues 18 to 221)*	**Fw:** TCACCAAAAACCAAAAAGACTACTAAAAAATCTACCAAACCTCCTAAAG **Rv:** TCCTTGGAAATACAGGTTTTCCAGATCCGATTTTGGAGGATGGTCGCC
pGEXT-Gcf1p_1-28 *(Residues 28 to 221)*	**Fw:** TCTACCAAACCTCCTAAAGTCGATACCAAG **Rv:** TCCTTGGAAATACAGGTTTTCCAGATCCGATTTTGGAGGATGGTCGCC
pGEXT-Gcf1p_1-38 *(Residues 38 to 221)*	**Fw:** GCTATCAGACTTCAGAAGAAG **Rv:** TCCTTGGAAATACAGGTTTTCCAGATCCGATTTTGGAGGATGGTCGCC
pGEXT-Gcf1p_CC+HMG1 *(Residues 1 to 144)*	**Fw:** TGACTCGAGCGGCCGCATCGTGACTGACTGACG **Rv:** CCCGTTGGCACCAAGTTTTGGTTTAGGAGTAAAATAGCTTTTGGC
pGEXT-Gcf1p_HMG1 *(Residues 84 to 144)*	**Fw:** GCAAGAAGCAAGATACATAAATTGGCACCAGGTAATTTTTATTCC **Rv:** TCCTTGGAAATACAGGTTTTCCAGATCCGATTTTGGAGGATGGTCGCC
pGEXT-Gcf1p_HMG2 *(Residues 129 to 221)*	**Fw:** GCCAAAAGCTATTTTACTCCTAAACCAAAACTTGGTGCCAACGGG **Rv:** TCCTTGGAAATACAGGTTTTCCAGATCCGATTTTGGAGGATGGTCGCC
pGEXTGcf1p_HMG1+HMG2 *(Residues 84 -221)*	**Fw:** GCAAGAAGCAAGATACATAAATTGGCACCAGGTAATTTTTATTCC **Rv:** TCCTTGGAAATACAGGTTTTCCAGATCCGATTTTGGAGGATGGTCGCC
Quickchange PCR for single residue mutation	
Construct name	**Oligonucleotides used**
pGEXT-Gcf1p_I49M	**Fw:** GACTTCAGAAGAAGATGAATGAGGCTAGGTCTGC **Rv:** GCAGACCTAGCCTCATTCATCTTCTTCTGAAGTC
pGEXT-Gcf1p_I59M	**Fw:** TTGCAGCAACAAATGAAAGATATTTCCACTCAACACAAG **Rv:** CTTGTGTTGAGTGGAAATATCTTTCATTTGTTGCTGCAA
pGEXT-Gcf1p_I62M	**Fw** TTGCAGCAACAAATGAAAGATATTTCCACTCAACACAAG **Rv:** CTTGTGTTGAGTGGACATATCTTTAATTTGTTGCTGCAA

pGEXT-Gcf1p_I59M+I62M	**Fw:** TTGCAGCAACAAATGAAAGATATGTCCACTCAACACAAG **Rv:** CTTGTGTTGAGTGGACATATCTTTCATTTGTTGCTGCAA
pGEXT-Gcf1p_K79M	**Fw:** CAAGACTTTGAGTAAACAAAGAATGTTCGAAGAAAAAGCAAGAAGCAAG **Rv:** CTTGCTTCTTGCTTTTTCTTCGAACATCTTTGTTTACTCAAAGTCTTG
pGEXT-Gcf1p_A145M	**Fw:** GGTGCCAACGGGTTTATGAAATATGTACAAGAAAATTACATTAGAGG **Rv:** CCTCTAATGTAATTTTCTTGTACATATTTCATAAACCCGTTGGCACC
pGEXT-Gcf1p_A193M	**Fw:** ATACAAGAAGATGCTTGAAAAATGGAAAGAACTTAGATTGAAGG **Rv:** CCTTCAATCTAAGTTCTTTCCATTTTTCAAGCATCTTCTTGTAT
pGEXT-Gcf1p_L209M	**Fw:** AAGGAATACAGTGATTATATGAAATTTAAGGAAAACTACAAAGTGGAGG **Rv:** CCTCCACTTTGTAGTTTTCCTTAAATTTCATATAATCACTGTATTCCTT
Insert PCR for sub-cloning in different expression vector	
Construct name	**Oligonucleotides used**
pCRI6b-Gcf1pwt	**Fw:** AATGCCATGGTCTCCTTGGCAACAAAAGCTGCAACC **Rv:** CCGCTCGAGTCAAAAGTCATCCTCCACTTTGTAGTTTTCC
pCri6b-Gcf1p_1-28	**Fw:** AATGCCATGGTCTCTACCAAACCTCCTAAAGTCGATACCAAGGCTATCAGACTTCAG **Rv:** CCGCTCGAGTCAAAAGTCATCCTCCACTTTGTAGTTTTCC
pCRI7b-Gcf1pwt	**Fw:** AATGCCATGGTCTCCTTGGCAACAAAAGCTGCAACC **Rv:** CCGCTCGAGAAAGTCATCCTCCACTTTGTAGTTTTCC
pCri7b-Gcf1p_1-28	**Fw:** AATGCCATGGTCTCTACCAAACCTCCTAAAGTCGATACCAAGGCTATCAGACTTCAG **Rv:** CCGCTCGAGAAAGTCATCCTCCACTTTGTAGTTTTCC

M3-PROTEIN EXPRESSION AND SOLUBILISATION TESTS

M3.1 Expression and solubilisation tests

The protocol used for our collaborators to produce Gcf1p [104] [102] was adapted to produce the protein in the high amounts and purity required for structural studies.

M3.1.1 Protein expression tests

Protein expression of each different Gcf1p construct was tested in parallel in the different *E. coli* expression strains BL21, Rosetta 2 and Origami. To this end, an aliquot of the corresponding competent cells was transformed with the plasmid containing the construct of interest by using the transformation protocol described previously protocol (*see mutagenesis section,* **Materials and Methods 2.1)**.

From the transformed cells plate, small scale cultures were prepared by picking one colony that was set to grow for 16 hours over-night in a glass tube containing 5mL of LB that additionally contained the antibiotics required for each tested E. coli strain (100 µg/mL ampicillin for both BL21 and Rosetta2 cells; 100 µg/mL ampicillin, 10 µg/mL streptomycin and 10 µg/mL tetracycline for Origami cells; all antibiotics diluted from a 1000X stock). 1 ml of each of these precultures was used to inoculate a volume of 20mL antibiotic-containing LB media in 100 mL Erlenmeyer flasks. Cultures were grown at 37 ºC and 200 rpm until reaching an O.D. of 0.6 UA (λ=600nm) and then induced with 1mM IPTG (dilution from a 1000X stock solution). The expression was tested at different temperatures: at 37ºC, 24ºC, and 16ºC, at different induction times (3h and 6h for 37 and 24ºC and 16h over-night for 16ºC). Note that when the expression T was different to 37 ºC, the culture was cooled at RT before induction. Expression of the different constructs was analysed by SDS-PAGE of 15% 40:1 acrylamide/bisacrylamide and by using *SeeBlue-PreStained Standard* (Thermo Scientific ®) as molecular weight marker. Gel lanes were charged with 10µL of non-induced and induced samples prepared as follows:

Preparation of induced and non-induced sample

Preparation of cell sample at O.D.$_{\mathbf{(\lambda=600nm)}}$ =0.6

Centrifugation of the cells at 4000xg for 5 minutes.

Resuspension of the cell pellet in 20 uL of LB

Addition of Laemmli Sample Buffer (LSB) 2X

Incubation of the sample at 95ºC for 5’.

Note that, samples from further solubility tests as well as chromatography fractions were prepared on a different manner: by mixing 5 parts of the protein solution with 1 part of LSB 6X.

M3.1.2 Solubility Tests

Those strains that showed positive expression results (*see expression tests section,* **Results 2.1.1)** were tested for protein solubility. Solubility tests were performed from identical aliquots coming from a single expression batch. Expression was performed by induction with 1mM IPTG, at 24ºC for 4 hours in a volume of 200mL. Aliquots of 10ml per subsequent solubilisation buffer tested was centrifuged at 4500xg for 30 minutes. Each cell pellet was resuspended in 5mL of a buffer containing 750mM NaCl and 50mM Tris-HCl pH 8.0, based both on the published protocol for Gcf1p purification [104] and for TFAM and Abf2p purification [39], [40] and [42].

Regarding solubility tests, each identical aliquot was supplemented with one of the 12 solubilisation additives from a previously designed battery. Each one was added up to a final concentration of 0.01% triton X-100, 0.01% tween 20, 300mM KCl, 5mM $MgCl_2$, 5mM $CaCl_2$, 1mM EDTA, 100mM urea, 100mM guanidium hydrochloride, 50mM arginine, 50 mM glutamate, 5% glycerol, 5% glucose, and 5% sucrose.

Samples were sonicated in 15mL Falcon® tubes with a 1.5mm tip. Sonication was performed at 20% amplitude for 30'' in 2'' pulses separated by intervals of 4''. Solubility in each of these conditions was assessed by SDS-PAGE: 10µL samples with 1X LDS were loaded in a 15% polyacrylamide, Tris-Glycine buffered gel.

M3.1.3 Strategies for the optimization of protein stability

After expression and solubilization tests, a condition was found in which the protein was soluble and in enough amounts for the characterization of the protein by structural methods, but a minor degradation product was observable in SDS-PAGE (*see expression tests section,* **Results 2.1.1 and Results 2.1.2**). In order to reduce the degradation, changes were added to the protocol consisting in the reduction of the expression temperature to 24ºC and the addition of serin-protease inhibitor PMSF at a concentration of 1mM. PMSF was also used in the purification of the GST-Gcf1p construct and during the affinity tag removal using TEV protease (*see GST affinity chromatography section,* **Materials and Methods 4.1.**).

M3.2 Definitive Gcf1p expression and solubilisation protocols

In this section we describe the definitive protocol for the expression and solubilisation of both native and selenomethionine derivative of Gcf1p that was used for protein characterization.

M3.2.1 Native Gcf1p production protocol

Pre-cultures of chemically competent BL21 cells transformed with the GST-tagged Gcf1p construct were grown in a rotatory cell shaker at 37º C and at 200rpm in 5mL LB, media for 6 hours. Afterwards, the saturated culture was transferred into a 500mL Erlenmeyer flask containing 200mL of LB-Ampicillin media and incubated in a rotatory cell shaker at 37ºC and 200 rpm for 16 hours over-night. Finally, 20mL of the saturated over-night pre-culture was inoculated into 2L Erlenmeyer flask containing 750mL of LB-Ampicillin media and incubated in a rotatory cell shaker at 37ºC and 200 rpm for approximately 2 hours until an optical density (OD) of 0.6 was reached. At this OD, protein expression was induced with IPTG at a final concentration of 1mM, by incubation at 24ºC during 4 hours in a rotatory cell shaker at 200 rpm. Subsequently, cells were pelleted by centrifugation at 4500xg for 30 minutes. Cell pellets were transferred to Falcon tubes, resuspended in PBS supplemented with 10% Glycerol, and subsequently flash-frozen in liquid nitrogen, and stored in -80º C. Successful expression was systematically confirmed by SDS-PAGE (*see expression tests section,* **Materials and Methods 3.1.1)**.

Pellets equivalent to 2L culture, were resuspended in 80 mL lysis buffer containing 750mM NaCl, 100mM HEPES-Na pH 7.25, 1mM EDTA, 1mM beta-mercaptoethanol, 5mM $MgCl_2$, 10 µg/mL DNase and RNase (final concentrations) and protease inhibitor cocktail 1X.

Cell suspension was clarified by sonication in a ultrasonic processor device (Branson®) at 60% amplitude with a 10mm probe. Sonication was performed for 2 minutes in short pulses of 2" followed by pauses of 4". The clarified suspension was finally lysed by using a **CF-1** cell disruptor (Constant Systems Ltd.) at a pressure of 1.36 Bar. After the cell disruptor step, the lysate was supplemented with PMSF and Triton X-100 as they showed to stabilize the protein (*see expression tests section,* **Results 3.1.1 and Results 3.1.2**). Due to its surfactant properties, these two additives are not compatible with the cell disruptor instrument.

Lysate was later centrifuged at 74000xg for 30 minutes in order to separate the soluble fraction containing the protein of interest from the insoluble fraction containing mainly intact cells, membrane fragments and inclusion bodies. Lysate was finally filtered through a 40µm syringe filter before undergoing any purification step.

M3.2.2 Selenomethinone derivatized Gcf1p production protocol

In order to solve the phase problem by experimental phasing methods, selenomethionine-derivatives (Se-Met) of wild-type Gcf1p were produced. The kit for selenomethionine incorporation (Molecular Dimensions®, now Calibre Scientific ®) was used. The bacterial growth medium is composed by three components: a medium base containing buffering solution, all aminoacids except L-Methionine and electrolites; a nutrient mix containing glucose, glycerol, vitamins and trace elements and finally an L-Selenomethionine solution. The bacterial growth medium had the following composition per every 550mL: 500mL of medium base, 40mL of nutrient mix, 7mL of Glycerol 100% 2.5 mL of L-Selenomethionine and 0.5 mL of Ampicillin. The protocol used was adapted from modified version available online in the webpage of Prof. Jan Löwe laboratory [113], [114], which was derived from the originally described in [115]. The Se-Met protein derivatives were prepared by using the same non-auxotrophic BL21strain employed for the native protein.

Precultures of chemically competent BL21 cells transformed with the GST-tagged Gcf1p construct of choice in a rotatory cell shaker at 37ºC and 200rpm in 5mL LB-Ampicillin media for 6 hours. Afterwards, the 5mL saturated culture was transferred in a 500mL Erlenmeyer flask containing 200mL of LB and incubated in a rotatory cell shaker at 37ºC and 200 rpm for 16 hours (O/N). Pre-cultures were centrifuged at 4500xg for 15 minutes and the excess LB was carefully removed by using an automatic pipette. Pellets were subsequently washed with Minimal Medium Base and finally inoculated to 2L Erlenmeyer flask containing 550mL of SeMet media and incubated in a rotatory cell shaker at 37ºC and 200 rpm for approximately 4 hours until an O.D. of 0.5 was reached. At this point, extra 2.5mL Selenomethionine solution at 10mg/mL was added to the growing culture, as well as a mix containing the aminoacids Val, Leu, Ile, Lys, Phe and Thr intended to inhibit the synthesis of endogenous L-methionine through metabolism control mechanisms based on negative feedback (REF). Once an O.D. of 0.6 was reached, cultures were induced with 1mM IPTG (final concentration) and incubated in a rotatory cell shaker at 24ºC and 200 rpm for 4 hours.

Downstream processing including harvesting, storage and protein solubilisation followed the same steps as for the native protein. It is worth noting that SE-Met protein solubilisation as well as subsequent purification steps was carried out in the presence of 5mM beta-mercaptoethanol to ensure that the oxidation state of selenium atoms was not changed along the purification process.

M4- PROTEIN PURIFICATION

M4.1 General notes about protein purification techniques

In order to attempt structural studies, we aimed at having a protein purification protocol that ensured high purity of the sample without compromising the global yield. However, not all the techniques used have the same requirements, so different strategies were undertaken depending on the purpose of the protein sample produced as it is depicted in **Figure M2**. In all cases, after the purification steps, fraction purity was assessed by SDS-PAGE (*see expression and solubilisation tests section,* **Materials and Methods 3.1.**) and UV absorbance. Note that Gcf1p is a DNA binding protein and prone to non-specifically bind nucleic acids. Therefore, a special attention was put on the 260/280 nm ratio, since a value higher than 0.7 for this parameter would indicate nucleic acid contamination.M3.1 Brief background of the used chromatography techniques

M4.1.1 GST Affinity chromatography

Glutathione-S transferase (GST) shows a strong and highly specific interaction with reduced glutathione, with a dissociation constant (K_d) of 6.9 nM [116]. Therefore, is a method of choice as a first step purification from bacterial lysate.

Protocol was adapted from [104]. DNase and RNase were added to the washing buffer in order to enzymatically degrade nucleic acid contaminants. Cell lysates were loaded onto *1mL GSTrap FF* (GEHealthcare) columns at a fixed flow of 0.5 mL/min to maximize binding to the resin. Depending on the volume of lysate processed, several columns were connected one after another at a reason of 1 column per 160mL of lysate. This step was preferentially performed in an AKTA FPLC system to precisely control both the flow and monitor eventual changes in the absorbance and conductivity of the sample. Nevertheless, we often performed this purification by using the peristaltic pump at the bench without any problem.

As the supernatant was directly loaded onto the columns, the loading buffer composition was the same as the lysis buffer (*see expression and solubilisation tests section,* **Materials and Methods 3.1**). Once the lysate was completely loaded, the column was washed with 10 CV of washing buffer (1000mM NaCl, 100mM HEPES-Na pH 7.25, 1mM EDTA, 1mM β-mercaptoethanol, 5mM $MgCl_2$, DNase and RNase at 10μg/mL, Protease inhibitor 1X, PMSF 1mM and Triton X-100 0.01%) until the absorbance at λ=280nm dropped to a stable value.

After the GST-affinity chromatography step, the GST-Gcf1p chimera was digested with the TEV protease. TEV protease is sensitive to NaCl concentrations above 1M, and a wash step was included by using 2 CV of a buffer containing 750mM NaCl, 100mM HEPES-Na pH 7.25, 5mM beta-mercaptoethanol, Protease inhibitor 1X, PMSF 1mM and Triton X-100 0.01%. Subsequently, the TEV protease was injected directly onto the column with a syringe at a final 1 protease: 20 target protein 'mass-to-mass' ratio. The column was left for 16 hours O/N digestion at room temperature. Finally, the digested protein was washed out with 10 CV and collected, using a 750 mM NaCl, 100 mM HEPES-Na pH 7.25, Protease inhibitor 1 X, 1 mM PMSF and Triton X-100 0.01 % buffer. TEV is insensitive to the protease inhibitors used for the stabilization of the protein during cell lysis (*see strategies for the optimization of protein stability,* **Materials and Methods 3.1.3**). Therefore, both PMSF and Protease Inhibitor cocktail (Roche ®) were kept during TEV digestion step.

M4.1.2 Size-Exclusion Chromatography

Stationary phase in a size-exclusion chromatography is composed by porous beads that interact with soluble components from a mixture based on their relative size. It is therefore a useful technique to identify the presence of protein stable multimers and aggregates.

Size-exclusion chromatography was performed in *Superdex* (GE-Healthcare®) columns coupled to an AKTA-FPLC system (GE-Healthcare®). For the purification of the protein for preparative purposes Superdex75 10/300 and Superdex75 26/600 matrix was used, the running buffer contained 750 mM NaCl and 50 mM Tris-HCl pH 8.0 and the flow rate 0.8 mL/min.

For analytical purposes (*see SEC-MALLS,* **Materials and Methods & Results**), a Superdex200 10/300 column was used and the running buffer contained 20 mM NaCl and 50 mM Tris-HCl pH 8.0 and flow 0.5 mL/min.

M4.1.3 Cation Exchange Chromatography

Cation exchange chromatography allows for the separation of the different components of a mixture based upon the preferential binding onto a electronegatively charged stationary phase.

Cation Exchange Chromatography was performed using *MonoS 5/50* columns (GE-Healthcare) coupled to an AKTA-FPLC system (GE-Healthcare). A unique bed column of 1 mL was used in all cases. After the purification step of the protein with the GST-affinity column, the 750 mM NaCl concentration was too high for an efficient binding of the protein to the column. Therefore, the pooled GST fractions were diluted 20 times in 50 mM HEPES-Na buffer at pH 7.5 until a final concentration of 150 mM NaCl (and approximately 0.2 mg/mL of Gcf1p) was achieved. The low concentration of Gcf1p ensured no aggregation despite the buffer was low in salt. Subsequently, the diluted protein sample was loaded onto the MonoS column at a constant flow of 0.5 mL/min. After the whole of the sample was loaded, the column was washed with 10 CV of 50 mM HEPES pH7.5, 300 mM NaCl buffer to discard most of the contaminants. Gcf1p was eluted by a linear salt gradient along 20 CV between a buffer A containing 300mM NaCl and a Buffer B containing 1M NaCl. Both buffers contain 50 mM HEPES-Na pH 7.5.

M4.1.4 Heparin Resin Chromatography

Heparin resin chromatography is a kind of pseudo-affinity column in which this sulphonated glycosaminoglycan in immobilized in the stationary phase. Due to its long and negatively charged structure it may mimic a DNA backbone. This kind of chromatography is often used as a last polishing step for DNA binding proteins.

Heparin resin chromatography was performed with *Heparin HiTrap FF* 1mL columns (GE-Healthcare) coupled to an AKTA-FPLC system (GE-Healthcare). This step was used both for direct purification of the protein following the GST-affinity step or as a last polishing step of the sample after the cation exchange chromatography. Analogously to the MonoS column, samples for the Heparin column were diluted up to 20 times in 50mM HEPES-Na buffer at pH 7.5 until a final concentration of at least 0.2 mg/mL Gcf1p and 150 mM NaCl were achieved. Subsequently, the diluted protein sample was loaded onto the column at a constant flow of 0.5 mL/min. After all sample was loaded, the column was washed with 10 CV of 50 mM HEPES pH 7.5, 200 mM NaCl buffer. Finally, the protein was eluted through a linear salt gradient by using the same buffers as for the cation exchange chromatography, until a final concentration of 1000mM NaCl, 50 mM HEPES pH7.5 along 20 CV.

M4.2 Workflows for the purification of Gcf1p

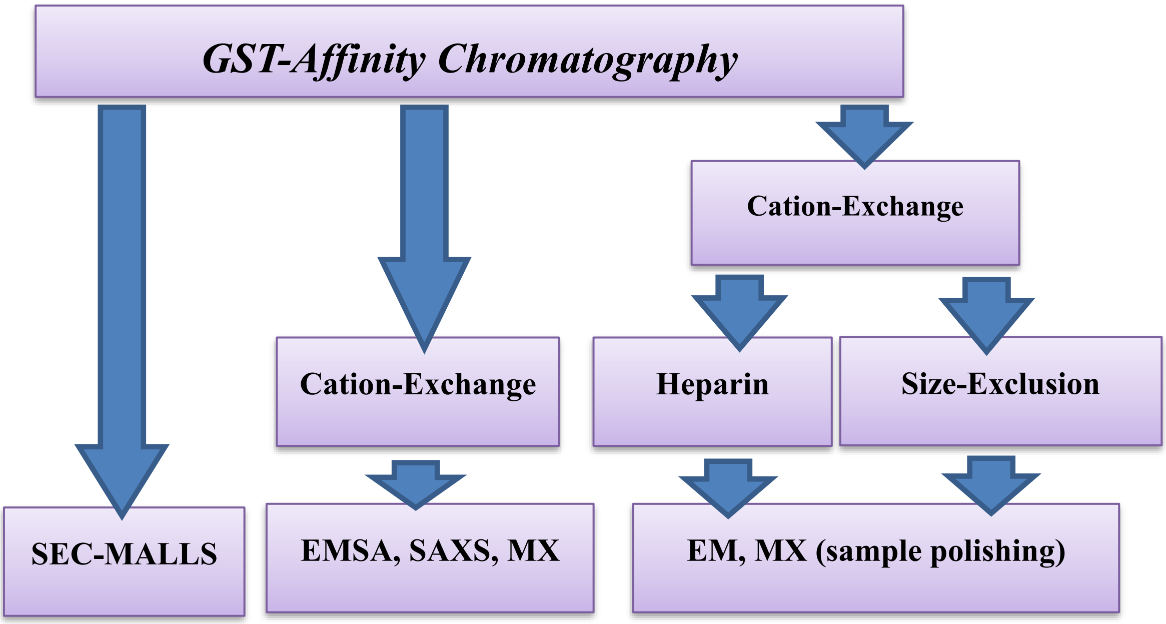


Figure M2: *Basic workflow followed during this entire project for the purification of wild-type Gcf1p and the different mutants, both for native and derivatives.*

Depending on the final purpose of the sample different protein purification workflows were applied (**Figure M2**). In order to optimize the protein production process, the yield of each step and the purity of the sample was assessed (see section **R2**)

M5- ELECTROPHORETIC MOBILITY SHIFT ASSAY (EMSA)

M5.1 Theoretical background

Electrophoretic mobility, defined as the relative movement of a particle subjected to an external electric field, is a function of particle parameters such as size, shape, mass and overall electrostatic charge of such a particle. In our case, changes in the electrophoretic mobility of free Gcf1p upon incubation with DNA indicated formation of a nucleic acid/protein complex. Not only binding can be assessed by such an Electrophoretic Mobility Shift Assay (EMSA), but also information about the homogeneity of the sample can be obtained. Homogeneous samples yield a single band whereas heterogeneous samples can show several bands or a smeared band. Aggregation or unspecific multimerization of the complexes are also detected as upshifted smeared bands that often are not able to enter the gel matrix. Different electrophoretic setups were used depending whether the DNA substrates were short (i.e smaller than 60bp) or long (bigger than 500bp).

M5.2 EMSA with short DNA substrates

Prior to EMSA, DNA duplexes were always prepared by annealing of complementary synthetic primers at equimolar concentration (Sigma-Aldrich®). Upon arrival to the laboratory, the single stranded, complementary DNAs were dissolved in water to a final concentration of 1mM. Afterwards, 45µL of complementary Oligos were mixed with 10µL of annealing buffer 10X (200mM NaCl, 200mM Tris pH8.0 and 50mM $MgCl_2$); samples were afterwards heated up in a thermal block for 5 minutes, and later cooled down at room temperature O/N. Final DNA duplexes were kept at -20ºC at a DNA stock concentration of 450µM in 20mM NaCl, 20mM Tris pH8.0 and 5mM $MgCl_2$. The DNA substrates used are listed in **Table M2**.

Full-length Gcf1p typically eluted from the MonoS column at an approximate salt concentration of 750 mM NaCl in 50m M HEPES pH7.5 (*see cationic exchange chromatography section,* **Results 3.2**). Starting from 20 mM Gcf1p, serial dilutions of Gcf1p were performed using buffer Dº (750 mM NaCl, 50 mM Tris pH 8.0), yielding concentrations of 10 mM, 5 mM, 2.5 mM and 1.25 mM Gcf1p. DNA duplexes at a constant concentration of 1 mM were diluted typically five times in buffer Rº (100 mM NaCl, 50 mM Tris pH 8.0 and 3 % glycerol) up to a final concentration of 200 nM. Finally, 9µL of DNA in buffer Rº were directly added to 1µL of each protein serial dilution and the samples incubated at room temperature for 30 minutes and loaded in native PAGE (formula). DNAs are thus at a final constant concentration of 200 nM with changing protein concentration to protein:DNA ratios of 5:1, 2.5:1, 1.25:1 and 0.63:1. The final buffer concentration can be assumed to be that of the original buffer Rº.

Gels were run at a constant voltage of 90 V for 2 hours in TBE 0.5X. Gels were stained for DNA with SYBR SAFE (Life technologies ®) 1X for 30 minutes and afterwards imaged in a *Typhoon gel scanner* (GE-Healthcare ®). Native gels of 12 cm were casted in a vertical setup (Hoeffner®) with an acrylamide concentration of 8% or 10% (for substrates longer than 35bp or shorter, respectively).

M5.3 EMSA with long DNA substrates

Long DNA substrates used for EMSA were prepared by PCR amplification from control plasmids PBR322 and pUC19. For all these PCR reactions, *DreamTaq polymerase* (ThermoScientific ®) was selected. Standard PCR of three steps per 40 cycles was used. Oligos and templates used for each fragment are listed in **Table M2**. After the PCR run, presence of an amplicon of the desired length and length homogeneity was assessed by agarose gel electrophoresis (1% Agarose, 0.5X TBE, 90 V). Fragments were later purified using a *PCR purification kit* (GE Healthcare®).

The same serial dilution protocol and incubation times used for EMSAs with short DNA substrates was used for long DNAs. After incubation at room temperature of the different complexes, samples were loaded in horizontal electrophoresis gels of agarose percentage ranging from 0.5 to 2% depending on the length of the DNAs used -0.5% for full-length pBR322 and pUC19, 1% for 1000bp DNA substrates, 1.5% for 500 bp DNA substrates and 2% for 200bp DNA substrates). Electrophoresis was typically performed at a constant voltage of 80 V during 6h in TBE 1X running buffer. For optimal sample separation, longer gels and 16h O/N runs at a constant voltage of 45V was also done. Gels were stained in SYBR SAFE (Life technologies ®) 1X for 30 minutes and afterwards imaged in gel scanner Typhoon 9500 (GE Healthcare ®).

Table M2: Summary of all the DNA substrates used in this project.

The main techniques in which DNA substrates were tried are included in the last column of this table. (MX: Macromolecular Crystallography, SAXS: Small-Angle X-ray Scattering and TEM: Transmission Electron Microscopy, SEC-MALLS: Size-Exclusion Chromatography coupled to Multiple Angle Laser Light Scattering). Sequences are annotated in the 5' to 3' direction. For the substrates used in MX, the module AAATT and other poly-adenine tracts that were critical for crystallization in complex with Abf2p in the past (See mitochondrial HMG-box structures section at the Introduction) *is depicted in red.*

Short DNA duplexes

Substrate name (length)	**Oligonucleotides used**	**Techniques used**
Af2_18 (18bp)	**Fw:** TAATAAATTATATAATAT **Rv:** ATATTATATAATTTATTA	**EMSA, MX**
Af2_20 (20bp)	**Fw:** ATAATAAATTATATAATATA **Rv:** TATATTATATAATTTATTAT	**EMSA, SEC-MALLS, SAXS, MX**
Af_22 (22bp)	**Fw:** AATAATAAATTATATAATATAA **Rv:** TTATATTATATAATTTATTATT	**EMSA, MX**
Af2_24 (24bp)	**Fw:** ATATAATAAATTATATAATATAAT **Rv:** ATTATATTATATAATTTATTATAT	**EMSA, MX**

Af2_26 (26bp)	**Fw:** ATAATAATAAATTATATAATATAATA **Rv:** TATTATATTATATAATTTATTATTAT	**EMSA, MX**
Af2_28 (28bp)	**Fw:** AATAATAATAAATTATATAATATAATAT **Rv:** ATATTATATTATATAATTTATTATTATT	**EMSA, MX**
Af2_44 (44bp)	**Fw:** AATAATAAATTATATAATATAAAATAATAAATTATATAATATAA **Rv:** TTATATTATATAATTTATTATTTTATATTATATAATTTATTATT	**MX**
Af2_21 NoT (21bp)	**Fw:** ATAATATTATATATATATATA **Rv:** TATATATATATATAATATTAT	**SAXS, MX**
Af2_20 Shift_1 (20bp)	**Fw:** ATAATAAAATTTATAATATA **Rv:** TATATTATAAATTTTATTAT	**MX**
Af2_20 Shift 2 (20bp)	**Fw:** ATAATATAAATTATAATATA **Rv:** TATATTATAATTTATATTAT	**MX**
Atp9_35bp (35bp)	**Fw:** GGTATTGGTATTGCTATCGTATTATTTAATTAA **Rv:** TTAATTAAATAATACGATAGCAATACCAATACC	**EMSA, MX**
Atp9_40bp (40bp)	**Fw:** GGTATTGGTATTGCTATCGTATTCGCAGCTTTAATTAAT **Rv:** ATTAATTAAAGCTGCGAATACGATAGCAATACCAATACC	**EMSA, MX**
Atp9_50bp (50bp)	**Fw:** GGAGCAGGTATTGGTATTGCTATCGTATTCGCAGCTTTAATTAATGGTGT **Rv:** ACACCATTAATTAAAGCTGCGAATACGATAGCAATACCAATACCTGCTCC	**EMSA, SEC-MALLS, SAXS, MX**
Y_22_Ac (22bp)	**Fw:** TAACAATTGAATGTCTGCACAG **Rv:** CTGTGCAGACATTCAATTGTTA	**EMSA, SAXS, MX**
GC_22 (22bp)	**Fw:** GAAGATATCCGGGTCCCAATAA **Rv:** TTATTGGGACCCGGATATCTTC	**EMSA**
Short DNA duplexes involving three or four independent double-stranded regions		
Substrate name (length)	**Oligonucleotides used**	
3-segment fork (22bp/arm)	**1:** AGCTATGACCATGATTACGAATTGCTTGGAATCCTGACGAACTGTAG **2:** AGCTACCATGCCTGCACGAATTAAGCAATTCGTAATCATGGTCATAGCT	**EMSA**

	3:AATTCGTGCAGGCATGGTAGCT **4:** CTACAGTTCGTCAGGATTCC	
4-way junction J3.12 (12bp/ arm)	**1:**GTCCTAGCAAGGGGCTGCTACCGGAAG **2:**CCGGTAGCAGCCTGAGCGGTGGTTGAA **3:**AACCACCGCTCAACTCAACTGCAGTCT **4:** CTGCAGTTGAGTCCTTGCTAGGACGGA	**EMSA, SAXS, MX**
4-way junction J4W (12bp/ arm)	**1:**GCAAAGATGTCCTAGCAATGTAAT **2:**AGTGCCAGTGATGGACATCTTTGC **3:**GACAGCTCCATGATCACTGGCACT **4:** ATTACATTGCTACATGGAGCTCTC	**EMSA, MX**
4-way junction J4W (25bp/ arm)	**1:** TGGGTCAACGTGGGCAAAGATGTCCTAGCAATGTAATCGTCTATGACGTT **2:** TGCCGAATTCTACCAGTGCCAGTGATGGACATCTTTGCCCACGTTGACCC **3:**GTCGGATCCTCTAGACAGCTCCATGATCACTGGCACTGGTAGAATTCGGC **4:** CAACGTCATAGACGATTACATTGCTACATGGAGCTGTCTAGAGGATCCGA	**EMSA, MX**
4-way junction Elongated-J4W (2x25bp/ arm and 2x12bp/arm)	**1:** TGGGTCAACGTGGGCAAAGATGTCCTAGCAATGTAAT **2:** ATTACATTGCTACATGGAGCTGTCTAGAGGATCCGAC **3:** GTCGGATCCTCTAGACAGCTCCATGATCACTGGCACT **4:** AGTGCCAGTGATGGACATCTTTGCCCACGTTGACCCA	**EMSA**
4-way junction L-shaped-J4W (2x25bp/ arm and 2x12bp/arm)	**1:** TGGGTCAACGTGGGCAAAGATGTCCTAGCAATGTAAT **2:** GCCGAATTCTACCAGTGCCAGTGATGGACATCTTTGCCCACGTTGACCCA **3:** GACAGCTCCATGATCACTGGCACTGGTAGAATTCGGC **4:** ATTACATTGCTACATGGAGCTGTC	**EMSA**

Long DNA substrates

Substrate Id (length)	**Oligos used to amplify of the target**	**Original plasmid**	**Topology**	**Techniques**
pBR_C7 (1131bp)	**Fw:** [Cy5]-CGACGCTCAAGTCAGAGG **Rv:** [biotin]-GATAACACTGCGGCCACC	***pBR322***	***Linear***	***EMSA, TEM***
pBR322 (4361bp)	***Full plasmid extracted from bacteria***	***pBR322***	***Supercoiled (Lk<0)***	***EMSA***
l-pUC191 (2686bp)	***Full plasmid linearized***	***pUC19***	***Linear***	***EMSA, TEM***

r-pUC191 (2686bp)	***Full plasmid relaxed***	***pUC19***	***Circular (Lk=0)***	***EMSA, TEM***
sc-pUC191 (2686bp)	***Full plasmid extracted from bacteria***	***pUC19***	***Supercoiled (Lk<0)***	***EMSA, TEM***

M6-TRANSMISSION ELECTRON MICROSCOPY (TEM)

M6.1 Theoretical background

Nanoscale objects are often difficult to characterize by conventional optic microscopy methods. On the other hand, crystallographic analysis is limited to particles that arrange in a tight three-dimensional array. This is highly improbable for DNA molecules larger than 50 bp. TEM has proven to be successful for the characterization of biological structures of a broad range of sizes, from 100 nm structures up to single proteins down to the atomic level with the recent Cryo-TEM advances [117], and for different systems, from membranes [118] to DNA [119].

The natural substrate of Gcf1p is the mitochondrial DNA of *Candida albicans* [102] [104], which consists of linear molecules of 40kb. Therefore, in order to fully understand the interaction of Gcf1p with long DNAs we analysed its interaction with DNA substrates of at least 1000bp. Transmission Electron Microscopy (TEM) is suitable for visualization of such long linear DNAs and characterize the protein induced distortions [120]. Thus, it is a perfect technique to study interactions that are not observable by Macromolecular Crystallography (MX). Since crystallography reveals short-range interactions, conclusions extracted from MX are highly complementary and non-exclusive of those extracted by TEM.

M6.1.1 Basic elements of a Transmission Electron Microscope

The main body of Transmission Electron Microscopes is a hollow cylinder into which high vacuum is applied (**Figure M3**). On top, an electron gun is located, which consists in a metal wire subjected to high voltage (from 80 to 200kV). From there, electrons are expelled and accelerated, thus forming a roughly defined beam that contains electrons of different kinetic energy. By means of their intrinsic negative charge, electrons can be deflected, focused and/or filtered by their kinetic energy . This is achieved by condenser and lenses placed between the electron gun and the sample, which generate a series of magnetic and electric fields that result in a well-defined electron beam with homogeneous kinetic energy and, therefore, homogeneous wavelength (*Equation 1*).

At the end of the microscope, the electrons interact with the specimen, which is held inside a metallic arm that controls the exact position of the sample (**Figure M3**). Electrons transmitted through the sample typically hit a fluorescence-emitting electron detector from which the image is formed. The projector lenses situated below the sample allow for different modes of image forming.

M6.1.2 Electron dynamics

Following Coulomb's law, when subjected to an external electric field, one electron will move against the lines of such field under uniform acceleration. Under these conditions, the energy of the electron will be constant along the field, e.g. its maximum kinetic energy will be equal to the initial potential energy and, thus, will be a function of the voltage applied. Electron is a particle that shows undulatory-corpuscular duality, and by virtue of de Broglie's law, its wavelength will be a function of electron's momentum as follows:

De Broglie law

$$\lambda = \frac{h}{mv}$$

Where h stands for Planck's constant: 6.62·10^{-34} J·s; m for the mass of the electron, 9.11·10^{-31} kg; v for the velocity of the accelerated electron at a given moment; and λ corresponds to the wavelength associated to the electron at a given velocity.

As explained above, the velocity of the electron will be a function of the applied voltage. Thus, the combination of *de Broglie* equation and *Coulomb's law*, it leads to the following expression where the wavelength λ of an electron in nm can be directly calculated from the applied voltage V, being $\lambda(\text{nm}) \approx \frac{1.23}{\sqrt{V}}$

Upon hitting the sample, electrons will interact both with nuclei and electron orbitals, generating different kind of scattered radiation as it is depicted in **Figure M4**. As it can be observed, electrons mainly pass through the sample without changing energy and/or trajectory. A series of secondary radiation (X-rays, Auger electrons) are generated and only a minor fraction is scattered through the sample and contributing to the TEM image. The interaction with the sample depends on the respective orientations of the electron beam and sample, the beam wavelength, and the energy and atomic number (Z) of the atoms present on the sample [118].

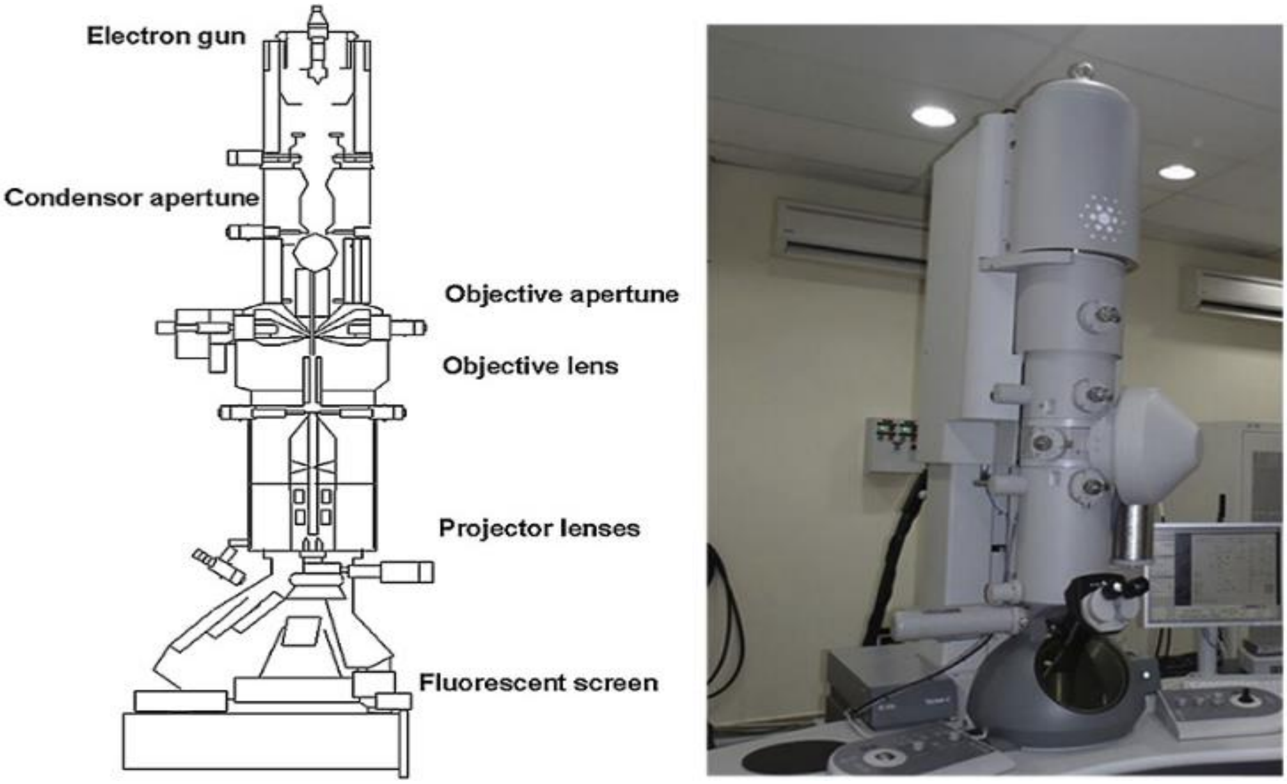


Figure M3: *Schematic depiction of the basic elements forming a Transmission Electron Microscope (left) and actual Transmission Electron Microscope (right image).* Extracted from [118]. C. Tang and Z. Yang, "Transmission Electron Microscopy (TEM)," in Membrane characterization, Elsevier, 2017, pp. Chapter 8: 145-158.

M6.1.3 Sample Contrast and image formation

As it has been shown in the previous section, the interaction of the electrons with the specimen will depend on Z of the specimen atoms. More precisely, atoms with high Z will scatter stronger than those with lower Z. Regarding macromolecules, we are in most of cases limited by atoms with low Z ($_{1}$H, $_{6}$C, $_{7}$N, $_{8}$O, $_{15}$P and $_{16}$S). These atoms are weak electron scatterers, so it is required to stain the samples with heavy atoms (unless the vitrification approach is applied as in CryoEM [121]).

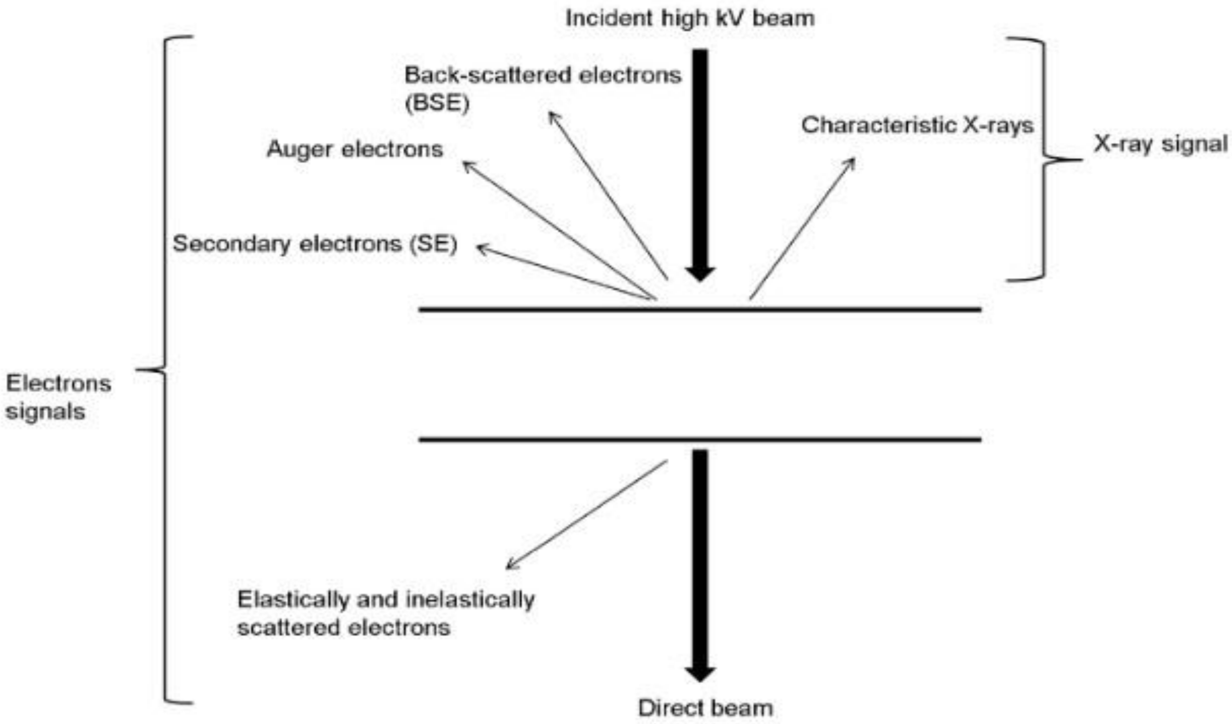


Figure M4: *Summary of the different radiation generated by the impact of the focused electron beam on a TEM specimen*. Extracted from [118]. C. Tang and Z. Yang, "Transmission Electron Microscopy (TEM)," in Membrane characterization, Elsevier, 2017, pp. Chapter 8: 145-158.

Sample staining with heavy-atoms is approached by two different methods, which differ on the type of interaction between the sample and the heavy-atom dye. In negative-staining, the heavy-atom dye is not absorbed by the sample but remains in the interstice between the sample and support, and it thus defines with considerable detail the boundaries of the particle. On the other hand, positive staining dyes directly interact with the sample, and yield specimens that scatter significant amounts of electrons that contrast well with the background. Since negative staining results in nice contour definition for proteins, it is mainly chosen above positive staining. However, positive staining permits a more accurate analysis of DNA parameters as compared to negative staining because it alters less significantly the thickness of the biopolymer. And so, for DNA imaging positive staining is the contrast method of choice. Regarding image generation, two main techniques are applied, namely dark and bright field imaging. Both result in the reconstruction of an image from the sample, but the principle of reconstruction is different between techniques. In bright field, the image is reconstructed upon transmission of the electrons through the sample. Instead, dark field imaging makes use of the diffracted electrons only, as shown in **Figure M5**. Therefore, for this latter, a series of electromagnetic lens refocus the diffracted beams on the detector, so the image obtained is directly in the real space (compared to X-ray crystallography, in which the data is at the reciprocal space due to the lack of a lens). Dark field images have less contribution from the unscattered electrons, so they characteristically show a better contrast than bright field imaging. As it can be observed in **Figure M5**, this effect is dramatic for DNA, which is a thin polymer that can only be observed in detail using positive staining contrast combined with dark field imaging so this was the option chosen during the experiments exposed further [118].

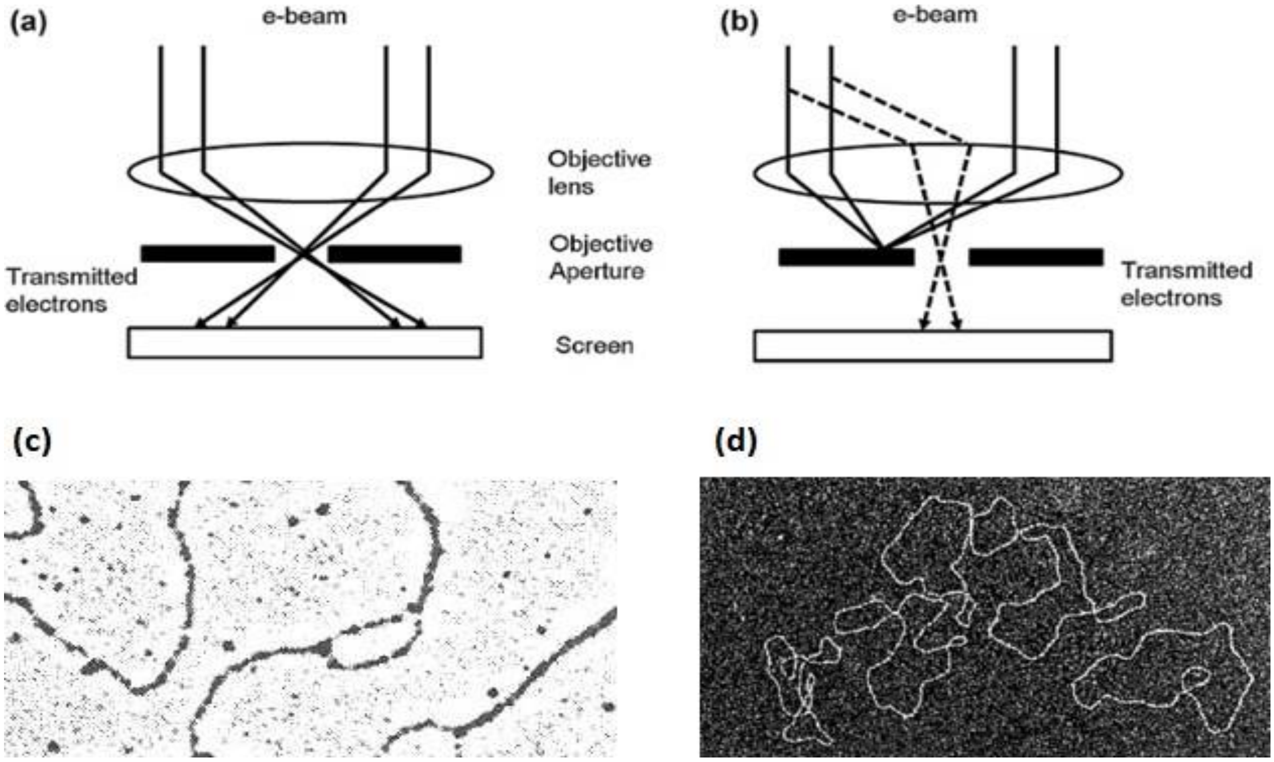


***Figure M5:** Different contrast methods and image formation in Transmission Electron Microscopy. Differences between bright (**a**) and dark-field (**b**) imaging modes are shown. Images of DNAs taken with TEM using positive contrast in both Bright-field (**c**) and Dark-field (**d**).* Extracted from [118]. C. Tang and Z. Yang, “Transmission Electron Microscopy (TEM),” in Membrane characterization, Elsevier, 2017, pp. Chapter 8: 145-158.

M6.2 Sample preparation for TEM

M6.2.1 DNA substrate preparation

pBR_C7 (1161 bp): 1000bp DNAs for TEM were amplified from PBR322 templates between positions 2576 and 3707 using the oligonucleotides pBR_C7-Fw and pBR_C7-Rv, listed in **Table M2.** Substrates were amplified using *Phusion* polymerase (ThermoScientific ®)

TEM substrates PCR protocol (30 cycles):

Initial denaturation: 98ºC for 2'
Denaturation: 98ºC for 20''
Annealing: 57ºC for 30''
Extension: 72ºC for 30''

PCR products were diluted ten times in water in order to lower the salt concentration. Samples were then injected onto a *MiniQ column* (GE Healthcare ®) at a constant flux of 0.1mL/min. *MiniQ columns* contain an anionic exchange bed column that traps the negatively charged DNA. Being so, the principles of interactions and purification workflow is almost identical to that described for cationic exchange chromatography (*See Protein purification,* **Materials and methods 4.3**). After all volume is loaded, a constant salt gradient was applied using a buffer A -50 mM NaCl, 20 mM Tris-HCl pH 8.0- and a buffer B -1 M NaCl, 20 mM Tris-HCl pH8.0- during 50CV at a constant flux of 0.1 mL/min. Fractions containing DNA were collected based on its absorption at 260nm wavelength UV light. Due to the strong electrostatic charge of DNA phosphate backbone, elution occurs at NaCl concentrations higher than 650mM NaCl. DNA was precipitated by addition of 100% ethanol and a 3 M Sodium Acetate pH 4.5 solution up to a final concentration of 50% ethanol and 300 mM sodium acetate. DNA pellet was washed three times by resuspension in Ethanol 70%, thus assuring DNA purity. During all this process, presence of the DNA pellet was easily tracked down thanks to the Cy5 marker present at 5' which gives a characteristic blue colour, making it easier to remove the supernatant by careful pipetting. At this point DNA substrates were stored at -20ºC for downstream processing and analysis.

sc-pUC191 (2686bp): Full-length pUC19 was obtained by DNA extraction from *Escherichia coli* culture using a *QIAprep Spin MiniPrep kit* (Qiagen®) *(See DNA manipulation for subcloning,* **Materials and methods 2.2**). In this kind of purification DNA was mainly recovered maintaining its natural negative supercoil tension.

r-pUC191 (2686bp): Relaxed circular DNA substrates were relaxed using topoisomerase I assay thus yielding a Gaussian distribution of topoisomers around Lk=0. Substrates were purified after the assay using *GFX^{TM} PCR and gel band purification kit* (GE Healthcare ®). This substrates were prepared by our collaborator Sonia Baconnais from prof. Éric le Cam laboratory.

l-pUC191 (2686bp): Linear DNA substrates were prepared by one-site digestion of full-length pUC19 molecule using blunt-end generating single-cut restriction enzyme. Substrates were purified after the assay using *GFX^{TM} PCR and gel band purification kit* (GE Healthcare ®). This substrates were prepared by our collaborator Sonia Baconnais from prof. Éric le Cam laboratory.

M6.2.2 Nucleoprotein complexes preparation for TEM analysis

Series of protein-DNA complexes were prepared maintaining a constant DNA concentration of 1 ng/μL and increasing protein concentrations, from 20nM to 500nM. Protein-DNA mixtures were diluted in reaction buffer up to a final concentration of 100 mM NaCl buffered with 20 mM Tris-HCl pH 8.1 and incubated for 30 minutes at RT, 20ºC and 37 ºC. Steps in complex preparation are identical to those described previously (*See EMSA with long DNA substrates,* **Materials and methods 5.3**). Following this procedure, the salt concentration was decreased after DNA addition thus promoting complex formation but reducing the typical precipitation of the protein due to low ionic strength.

Protein-DNA aggregates were removed following two strategies. On one side, protein-DNA complexes were loaded immediately after the incubation onto a gel filtration column (*See Protein purification section,* **Materials and methods 4**). Peaks corresponding to the aggregates and the complex of interest were followed using UV-light absorbance at 260nm wavelength. Samples corresponding to the peak of interest were directly loaded onto the grids from the FPLC sample fractionator in order to reduce sample manipulation. Alternatively, aggregates were removed by addition of 1 μL of sc-pUC191 at 10 ng/μL concentration to the already formed complexes. The excess of DNA captured the free protein and thus prevented its aggregation in the low salt conditions. Removal of aggregates was not needed in the complexes containing circular supercoiled or circular relaxed DNA (*See TEM images,* **Results 5.2**).

M6.2.3 Grid preparation

Once the nucleoprotein complexes were prepared, the mixture was applied onto carbon-coated copper grids with no delay and carbon coats had to be prepared before the complexes. First, graphene was deposited onto a glass microscope slide by thermal evaporation of carbon threads in vacuum. The thin carbon layer was then gently transferred on top of the grids by progressive evacuation of a water bath containing the grids at the bottom and the carbon film slides in the water-air interface. Such carbon films are hydrophobic so, once deposited on the grids, they were functionalized by glow-discharge in an atmosphere of 1-pentilamine for 5 minutes. During the first hour after functionalization, 5 µL of the sample were deposited on top of the grids and subsequently 5 µL of uranyl acetate were applied to spread the sample homogeneously in addition to act as a contrasting agent. After one-minute incubation, grids were then dried by manually blotting out the excess of liquid with filter paper and stored for further imaging at room temperature.

Microscopy images were acquires using 80 kV and 120 kV Leo-Zeiss microscopes.

M7- SEC-MALLS

M7.1-Theoretical background

Size-Exclusion Chromatography coupled to a Multiple Angle Laser Light Scattering (SEC-MALLS) detector was used to determine the molecular weight of both Gcf1p and the complexes formed with different DNA substrates in solution. In SEC-MALLS the sample is first passed through a gel filtration column in which the different components of the sample are separated following its hydrodynamic radius. A MALLS detector is coupled to the end of the chromatography and the UV-light scattering of the samples is measured as they leave the column. A refractometer is also coupled at the end of the column which allows to measure the exact peak concentration. The values for the calculated light scattering and peak concentration allows to obtain an exact value of molecular weight for each peak.

In the case of Gcf1p it was also an important step in determining the conditions in which our protein-DNA complexes were to be prepared prior to attempt structural characterization of our system. The sample preparation protocol detailed in the following subsection was also used for structural characterization in solution and for protein crystallization.

M7.2-Sample preparation

DNA-binding proteins require high salt conditions for stability which, in turn, prevents DNA binding. Stable complexes of Gcf1p with DNA substrates were prepared by mixing 1mL of Gcf1p -in 750 mM NaCl, 50 mM HEPES-Na pH 7.5- at 1mg/mL, i.e. 38.6 µM, and 42.8 µL of the chosen DNA substrate -in 20 mM NaCl, 20 mM Tris-HCl pH 8.0, 5mM $MgCl_2$- at 450 µM, i.e. a final 2:1 protein:DNA ratio, DNA addition had negligible effect in salt concentration. This mixture was diluted by addition of 4 mL of a buffer containing 350 mM NaCl, 50 mM Tris-HCl pH 8.0, until a final 430 mM NaCl concentration and a final 0.2 mg/mL protein concentration was achieved. The diluted 5mL mixture was loaded in a 3000 Da Molecular Weight cut-off dialysis bag (Sartorius ®) and subsequently exposed to three consecutive buffers of decreasing salt concentration, 1L of buffer A (350 mM NaCl, 50mM Tris-HCl pH8.0) for 2h; then to 1L of buffer B (180 mM NaCl, 50 mM Tris-HCl pH 8.0) for 2h; and to 2L of buffer C (20mM NaCl, 50 mM Tris-HCl pH 8.0) for 16h over-night incubation.

Complexes were concentrated with 3000 molecular weight cut-off *Protein Concentrators* (MerckMillipore ®) until a protein concentration of 5mg/mL was achieved. Protein concentration was measured using colorimetric measurement using the bicinchoninic acid method (Smith P. K., et al., 1985), *Pierce^{TM} BCA^{TM} protein assay kit* (ThermoScientific®) and alternatively the Bradford reagent method (Bradford, 1976), *Bio-Rad protein assay* (Bio-Rad®). For both methods, calibration curves were performed using Gcf1p samples within the dynamic range of each technique, between 20 µg/mL and 2000 µg/mL for BCA^{TM} and between 200 µg/mL and 1400µg/mL of known concentration assessed by UV absorbance at 280 nm wavelength.

M7.3-Scattering measurements and data treatment

Both the refractive index measurements and the scattering measurements required for protein concentration and molecular mass calculations were performed using a scattering DAWN-HELEOS-II-detector (Wyatt Technology ®). Data treatment was performed using ASTRA software. Calculations were performed by Laura Company Sapiña from the PCB crystallography platform; she also equilibrated the columns and supervised the different SEC-MALLS experiments:

M8- SAXS

M8.1 Theoretical background

Small Angle X-ray Scattering (SAXS) is a method for the characterization of both structured and disordered macromolecules. It is a structural biology technique performed in solution that allows for the calculation of the main structural parameters of a particle at low resolution. This information is used for shape and size description as well as for the reconstruction of low-resolution models. SAXS can also quantitatively determine the sample polydispersity in solution, and thus it describes conformational heterogeneity and dynamics (Kikhney & Svergun, 2015).

Elastic scattering arises from the resonance condition stablished between incident electromagnetic radiation and electron orbitals from the particle. Ideal scattering from a single electron is described in *equation 1.*

Eq1: Thomson scattering of X-rays by a single electron

$$I_e(2\theta) = \frac{r_0^2}{R^2} I_0 = r_0^2 \cdot \left(\frac{1 + cos^2(2\theta)}{2}\right) \frac{1}{R^2} I_0$$

Where I_e stands for the intensity of the scattered X-rays and I_0 for the intensity of the incident X-rays, r0 is the classical electron radius 2.817x10-15 m, R is the distance between the source of the scattering and 2θ is the angle between the incident photons and the scattered ones.

The direction of the incident (I_0) and scattered radiation (I_e) is determined by the wave vectors $\overrightarrow{k_0}$ and $\overrightarrow{k_1}$, respectively. In Thomson scattering, the magnitude of both vectors is identical and being a travelling wave its value is equal to $\frac{2\pi}{\lambda}$. The difference between vectors $\overrightarrow{k_0}$ and $\overrightarrow{k_1}$ gives rise to a vector $\vec{q}$ parallel to the detector plane and whose magnitude $|\vec{q}|$ is defined as momentum transfer (Svergun & Koch, 2003). From now on and following the most commonly used nomenclature of the field, vectors $\overrightarrow{k_0}$, $\overrightarrow{k_1}$ and $\vec{q}$ will be listed as $\boldsymbol{k_0}$, $\boldsymbol{k_1}$ and $\boldsymbol{q}$ and its magnitude as k_0, k_1 and q respectively. It is interesting to note at this point that the momentum transfer, here defined as q can be found in related bibliography indistinctly as q, s, h or k.

Eq2: momentum transfer in X-ray scattering

$$k = k_0 = k_1 = \frac{2\pi}{\lambda}$$

$$\boldsymbol{k_0}(k, \Omega) = \left(\frac{2\pi}{\lambda}, 0\right); \boldsymbol{k_1}(k, \Omega) = \left(\frac{2\pi}{\lambda}, 2\theta\right); \ \boldsymbol{q}(q, \Omega) = \boldsymbol{k_0} - \boldsymbol{k_1}$$

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

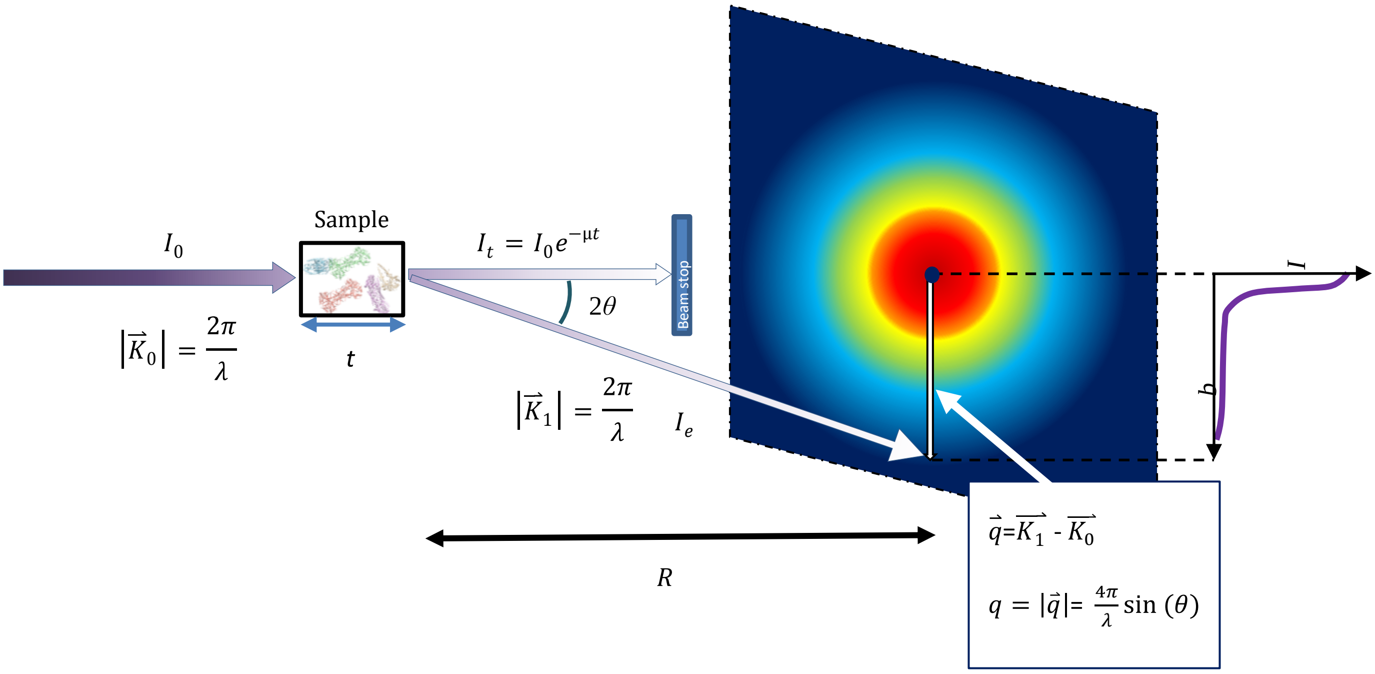

I_0
Sample
$I_t = I_0 e^{-\mu t}$
Beam stop
2θ
$|\vec{K}_0| = \frac{2\pi}{\lambda}$
t
$|\vec{K}_1| = \frac{2\pi}{\lambda}$
I_e
I
q
R
$\vec{q}=\vec{K_1} - \vec{K_0}$
$q = |\vec{q}|= \frac{4\pi}{\lambda}\sin(\theta)$

Figure M6: *Schematics of a SAXS experiment on a regular beamline. Incident radiation I_0 is a monochromatic, unpolarized beam with wavevector* ***k_0*** *directed in perpendicular to the vertical of the sample. After impacting the sample of macromolecules in solution most of the incident radiation is transmitted without changing the direction. Transmitted beam I_t is a function of the incident beam I_0, the absorbance coefficient of the sample μ and its thickness t and it is deflected by a metallic beam stop in order not to alter the scattering pattern of the sample or damage the detector. The beam I_e is scattered without energy transfer thus the modulus of the scattered wave k_1 is equal to that of the incident k_0 but changes its direction by a 2θ angle as described in equations 1 and 2. Momentum transfer vector* ***q*** *and its module q are calculated from the difference between incident and scattered wave vectors. The typical diffuse, spherically averaged scattering signal of solution experiments can be plotted in one dimension as I as a function of the momentum transfer q. Inspired in figures present in (Svergun & Koch, 2003), (Cornell Synchrotron, BioSAXS) and (Koch, 2011), shown elements are not in scale.*

For the study of the scattering from molecules the concept of the scattering density distribution ρ(r) (see *Equation 3*) equal to the total scattering length of the atoms per unit volume (integrated over the volume in *Equation 3*. The X-ray scattering signal of macromolecules in solution is generally weak and rarely significantly stronger than that of the bulk solvent. To circumvent this, for each measurement, the scattering of a buffer control must be subtracted from the curve of the macromolecular sample, assuming the solvent has a constant scattering density $\Delta\rho_s$. The difference scattering amplitude from a single particle over the equivalent solvent volume V is defined by the Fourier transform ($F[\Delta\rho(r)]$ in *Equation 3*) of the excess scattering length density $\Delta\rho(r) = \rho(r) - \rho_s$. Amplitude of the scattered wave A(s) is not directly measured in a scattering experiment but the intensity of the scattering I(q)=A(q)A*(q), where A*(q) is the complex conjugate of the amplitude A(s) (Svergun & Koch, 2003).

Eq.3: Scattering amplitude and scattering density

$$A(q) = F[\Delta\rho(r)] = \int \Delta\rho(r)e^{iqr}\, dr$$

Following the properties of the Fourier transform, reciprocity between dimensions in the real space (i.e. dimensions within the particle) and dimensions in the reciprocal space (i.e. length of the scattering vector in the detector, q) is present in a SAXS pattern. Hence to that, lower q values contain information of long intraparticle distances, i.e. low-resolution features of the particle, and vice versa.

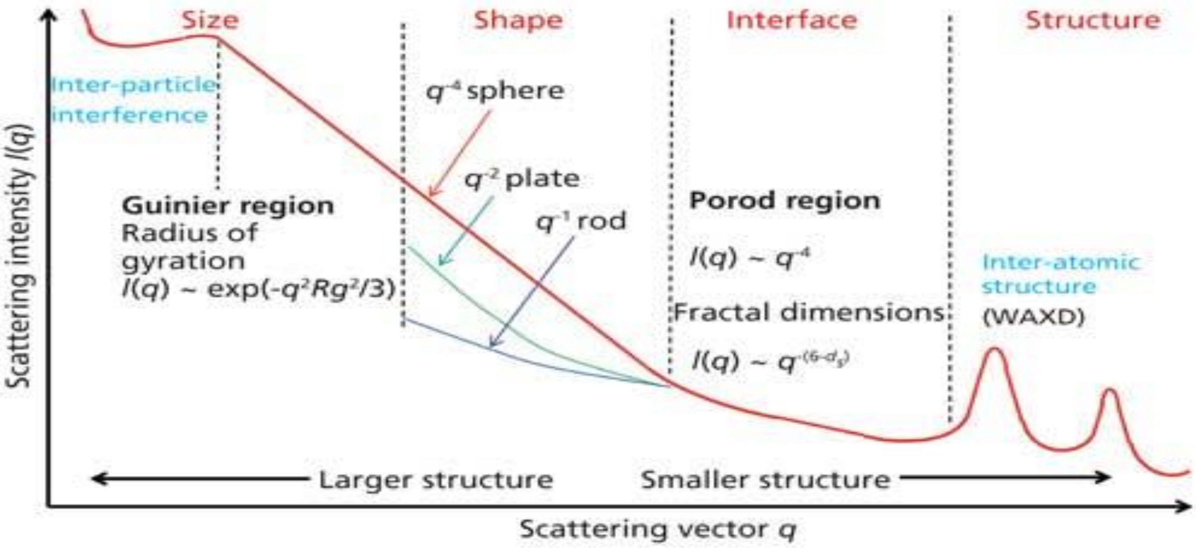


Figure M7: *Example of a SAXS profile representing the scattering Intensity i.e. I(q) as a function of scattering vector q.* Extracted and adapted from [160] (Murthy et al. 2017)

M8.1.1 The Guinier Region. Information about particle size

Regardless of particle shape, at very small scattering angles, *I(q)* can be expressed as a function of the radius of gyration (*Rg*). The radius of gyration or *gyradius* is a core concept of polymer physics that describes the dimensions of a polymer chain without regard of its shape or size. Mathematically, the *gyradius* of a particle in solution is understood as the root mean square deviation (rmsd) of the particle's parts distance to the centre mass of such a particle. At scattering angles (*2θ*) in which *q<1.3/Rg*, at the region known as Guinier regime, Guinier law becomes valid, by which:

Eq 4. Guinier Law

$$I(q) = I_0 \cdot e^{-\frac{1}{3} \cdot R_g^2 \cdot q^2}$$

Thus, by means of *equation 3* a representation of ln *I(q)* vs q^2 (i.e. the Guinier Plot) would yield a first order linear function for those values of *q* satisfying the condition for the Guinier regime. In this function, the slope will be the *Rg* and the intercept with the y axis would be the parameter I_0. The intersect I_0 is directly related to the molecular mass (MW) of the protein. Therefore, the MW can be assessed in solution using the I_0 values for protein standards of known molecular weight, typically lysozyme and bovine serum albumin (BSA). Deviations from linearity within the Guinier regime make the Rg calculation inconsistent with the actual composition of the sample. Such deviations normally arise from two phenomena. First aggregated particles, which have a molecular mass much greater than that of the particle of interest, and thus affect R_g calculation. Lack of linearity may also be due to the presence of positive or negative interparticle interactions, causing respectively over and underestimation of the R_g calculation. For these reasons, the sample analysed by SAXS must be completely free from aggregated particles. In addition, it is also highly advisable to measure the samples at different concentrations in order to do an effective extrapolation to zero concentration and obtain meaningful values for both I_0 and R_g.

M8.1.2 The Kratky Plot. Information about particle compactnesss and flexibility

One of the significant strengths of SAXS analysis is the possibility to detect the presence of intrinsically disordered regions and characterize them. Qualitative information of the particle compactness and globularity is obtained by visual inspection of the so-called Kratky plot, where $q^2I(q)$ is plotted against *q*. In the high angle region, corresponding to smaller distances in the real space, scattering is no longer a function of size but reflects flexibility, which has a strong impact on the *I(q)* values (**Figure M8**). More specifically, at higher angles, an intrinsically disordered protein will scatter stronger than the globularly folded counterpart. This information becomes evident with the Kratky plot, in which a globular protein will show a Gaussian distribution while that of an intrinsically disordered protein will show a continuous increase along the x axis (Brennich, Pernot, & Round, 2017) (**Figure M8).** Partially flexible proteins will show higher $q^2I(q)$ values at high q than those of a globular protein. Therefore, the Kratky plot inspection gives hints of total, partial or no flexibility of the analysed sample.

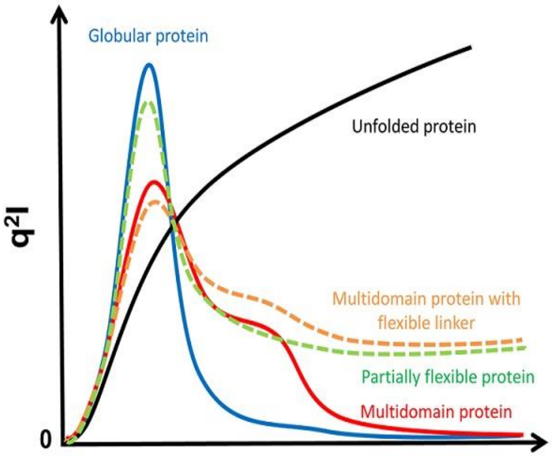


Figure M8: *Differences between different compaction states of particles can be quickly assessed by visual inspection of the Kratky plot.* [158]

https://www-ssrl.slac.stanford.edu/~saxs/analysis/assessment.htm.

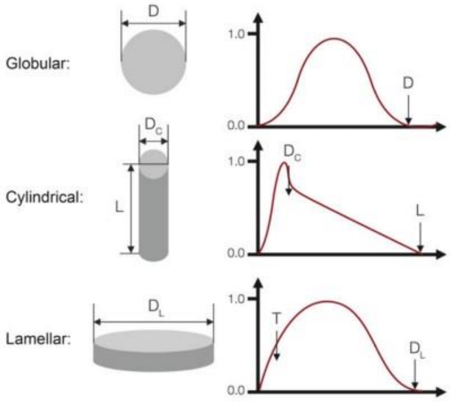


Figure M9*: The distance distribution function reflects the geometrical features of randomly oriented objects in a confined space.* Extracted and adapted from [161]. Alford et al., 2017

M8.1.3 The distance distribution function. Information about particle shape

Another very informative plot that can be extracted directly from the SAXS curve is the distance distribution function or p(r) plot, in which the probability of the different intra-particle distances p (r) is plotted against the distance *r* in a smoothed histogram. This is a real space plot directly related to the scattering curve by an inverse Fourier transform (*equation 5*).

EQ.5: DISTANCE DISTRIBUTION PLOT

$$p(r) = \frac{r^2}{2\pi^2}\int_0^\infty I(q) \cdot q^2 \cdot \frac{\sin q\, r}{q} dr$$

The p(r) plot shape contains information about the geometry of the particle, which can be intuitively approached by its visual inspection, e.g elongated shapes will scatter differently compared to spheres (**Figure M9)**. Finally, the p(r) plot is the inverse Fourier transform of the scattering function of a sample. As the scattering pattern corresponds to the direct Fourier transform of the electron density *Equation 5*, we can apply the inverse mathematical operator to reconstruct a real-space system from reciprocal space observations.

M8.1.4 Ab initio modelling from SAXS scattering pattern

As it has been mentioned before, SAXS is a technique in which the scattering of the solvent and that of the studied particle can be assessed independently by subtraction of a perfectly matching particle-free solvent (or buffer in the case of macromolecules). The difference scattering amplitude (A(*s*)) from a single particle respect to that of the equivalent solvent volume is related with the excess electron density (ρ) by a Fourier transform (F) . (Svergun & Koch, 2003)

With this relation, an *ab initio* model is calculated by filling with 'dummy atoms' the region with positive electron density , which corresponds to the particle, and leaving the rest of the volume empty, which is assumed to be the solvent. This is achieved by random positioning of the dummy-atoms using a probabilistic Monte Carlo approach and, afterwards, the resulting model is refined against the original scattering data and misplaced atoms are removed (Svergun, Restoring low resolution structure of biological macromolecules from solution scattering using simulated annealing., 1999). Model fitting to the experimental data is in each case assessed using a *chi-squared distribution* (*Equation 6*). Other programs, like GASBOR, follow an approach were the dummy atoms are substituted by dummy residues constrained by imposing the 3.6A to 3.8A distance characteristic of the inter-Cα distance (Svergun, Petoukhov, & Koch, Determination of domain structure of protein from X-ray solution scattering, 2001). In order to determine the uniqueness of SAXS *ab initio* models it is necessary to compare, and average of the different models obtained in separate runs with software designed for this purpose(Volkov & Svergun, 2003).

Eq 6: Chi squared expression:

$$\chi^2 = \frac{1}{K-1}\sum_{j=1}^{K}\left[\frac{\mu I(s_j) - Iexp(s_j)}{\sigma(s_j)}\right]^2$$

M8.1.5 Computation of scattering from a known structure

As it has been previously introduced (*Equation 3*) the scattering of a macromolecule is a function of the excess electron density of a single particle over the buffer averaged over all particle orientations, thus accounting for rotational tumbling in solution. Inversely, from a known high-resolution crystal structure one can calculate a theoretical scattering and fit it to an experimental scattering curve [122]. CRYSOL, a program that allows us to calculate such theoretical scattering curves also calculates the agreement between theoretical curve and experimental data using a chi-squared analysis (*Equation 6*). CRYSOL also performs an adjustment of the solvation shell thickness of the particle, thus accounting for the effect that the ordered solvent (i.e., in close contact with the particle) has in the final scattering curve over bulk solvent [122]. This approach allows for the validation of crystal structures in solution. Other approaches can be complementary to CRYSOL, for mixes of oligomeric states of a protein, OLIGOMER can be used in order to calculate the theoretical SAXS curve for each complex and the degree in which each component contributes to the final curve [123]. For the modelling of conformational changes from a high-resolution crystal structure, SREFLEX can be used, which imposes normal mode analysis to introduce flexibility in different regions of the particle and then assesses the agreement with the data using a chi-squared analysis (*Equation 6*) [124].

M8.1.6 Structural analysis of unstructured regions

For the study of intrinsically disordered proteins or proteins that have non-structured regions, an approach is to use conformational ensembles created from available 3D structures to model the corresponding fully or partly disordered proteins in solution. This approach is known as the Ensemble Optimization Method (Bernadó, Mylonas, Petoukhov, Blackledge, & Svergun, 2007) (Tria, Mertens, Kachala, & Svergun, 2015) and is performed in two stages. As a first step, the boundaries between domains are defined so that inter-domain aminoacids are assigned as to have a random coil conformation and defined as dummy residues. If we consider the dihedral angles between such dummy residues of the random coil, an astronomic number of possible conformations can be generated. At this stage, thousands of models, up to 10000, are built. In a second stage, these models are randomly grouped in ensembles of 10 to 50 conformations and fitted to the experimental curve by using a *chi-squared* distribution. This process is performed iteratively following a genetic algorithm by which the ensembles of conformations that best fit the data are selected, combined and fitted again, in an *in-silico* fitness approach. At the end of the process, good values of *chi-squared* renders those sub-ensembles that better describe the data. The most relevant output of the EOM is the distribution plot of *Rg* in the experimental sample. Nonetheless, representation of the selected sub-ensemble models gives an intuitive picture of the conformational space explored by the molecule in solution.

M8.2 Sample preparation for SAXS

As it has been described before (See section **M7.2**) complexes with different DNAs were prepared by successive dialysis to reduce the salt to 20mM NaCl, 50mM Tris-HCl pH 8.0 . In addition, samples containing the protein alone were prepared in 750mM NaCl and 50mM Tris-HCl pH8.0. Samples were concentrated with 3000 *Molecular Weight Cut-Off Amicon filters* (Merck ®) and the flow-through was kept in all cases to have a perfectly matching buffer for subtraction. Sample concentration was assessed through UV absorbance at 280nm in NanoDrop for the protein alone and by using the *Pierce BCA method* (ThermoScientific ®) for the protein-DNA complexes (Smith, et al., 1985). Finally, samples containing DNA alone were prepared by serial dilutions of the annealed double and four-stranded DNA molecules in milli-Q water. In order to assess the contribution of interparticle interactions to the overall scattering, the samples were prepared at different concentrations (0.25, 0.5, 1.0 and 2.0 mg/mL for DNA complexes; 1.25, 2.5, 5.0 and 10.0 mg/mL for protein alone and 0.5, 1, 3 and 6 mg/mL for DNA alone)

M8.3 Data collection and data reduction

SAXS measurements were performed at the BM29 beamline at the European Synchrotron Radiation Facility in Grenoble (ESRF) and at the P12 beamline in the Deutsches Electronen-Synchrotron in Hamburg (DESY) in collaboration with scientists from the BioSAXS group headed by Prof. Dmitri Svergun from European Molecular Biology Labs (EMBL) in Hamburg. In both cases, data was collected at 20ºC in the flow mode using 35 to 60 µL volume per injection.

The program Primus from the *ATSAS* package (Franke, et al., 2017) (Konarev, Volkov, Sokolova, Koch, & Svergun, 2003) was used for analysis of the curves. The radius of gyration (*Rg*) for each sample and the I_0 was obtained using the Guinier approximation. Programs Gnom and Autognom, also from *ATSAS* package were used to determine the distance distribution or P(*r*) plot and the maximum diameter of the particles (*Dmax*) (Svergun D. , 1992).

For shape and conformation analysis, programs Dammif and Damaver were used to build the dummy-atom models of the protein-DNA complexes in solution (Svergun, Restoring low resolution structure of biological macromolecules from solution scattering using simulated annealing., 1999) (Volkov & Svergun, 2003). Finally, the EOM approach (Bernadó, Mylonas, Petoukhov, Blackledge, & Svergun, 2007) (Tria, Mertens, Kachala, & Svergun, 2015), was also used to make conformational ensembles of the free protein with the dedicated support of Dr Pau Bernadó (CBS, Montpellier, France).

M9-CRYSTALLIZATION

M9.1- Theoretical background:

Crystallization of a soluble particle is a phase-transition process in which such a particle undergoes an ordered precipitation process. Particles forming a crystal arrange in a regular three-dimensional lattice that may or may not present internal symmetry.

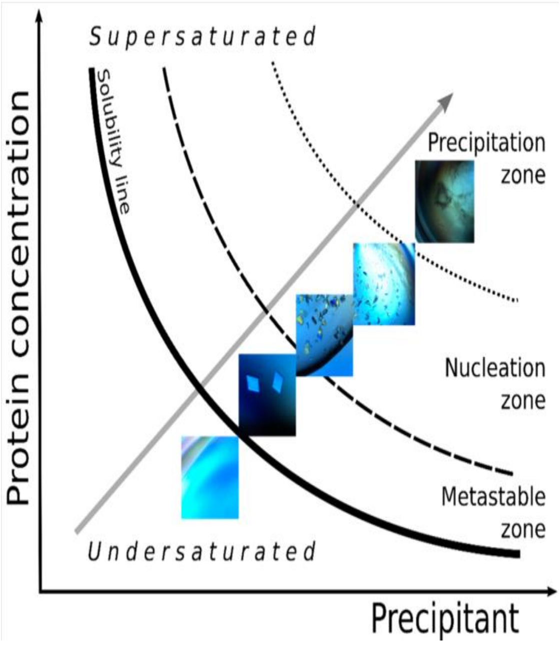
In the crystallization process, the most energetically expensive step is the nucleation , in which the ordered crystalline nuclei form. To reach this point, different energetic barriers are crossed, especially the loss of the particle hydration shell. This phenomenon is only favourable in supersaturated solutions. Once the crystal nuclei are formed, the energetic cost for other particles to leave the solution and attach to the crystal surface is greatly reduced. Thus, a lesser concentration is required for an already formed nucleus to grow as compared to the *de novo* formation of nuclei. In a typical crystallization experiment, different points of the solubility curve of the protein is explored by changing both the protein concentration and the concentration of precipitant, a phase diagram is generated that explains the crystallization behaviour for a particular compound **Figure M10** (Asherie, 2004).

In a phase diagram, several characteristic regions are found. The labile zone is in which effective nucleation may take place. Once the nuclei start to form, as they locally trap molecules in a solid phase, the surrounding concentration of soluble molecules will decrease. This local depletion of the available molecules would ideally move the solution to the metastable zone in the phase diagram, where the formation of nuclei is less probable than the enlargement of already formed nuclei. Thus, the formed nuclei will grow into crystals.

By knowing this behaviour and starting from chemical conditions in which nuclei appear, the experimenter changes the protein and precipitant concentrations and, by this, will observe how the number and size of crystals will change consequently. In principle, the longer the crystallization solution stays in the labile zone, more crystals will appear but the smaller they will be. Thus, by trial and error within the metastable zone, crystals of enough quality to perform diffraction data collection might be obtained.

The crystallization technique *in batch* exploits just one point in the phase diagram. However, it is highly improbable to reach the exact point in which crystals of the desired size and quality will appear. As an alternative, several approaches have been developed that explore the different regions of the phase diagram, among which *vapor diffusion, dyalisis and capillary counter-diffusion.*

Among all of them, we have chosen the vapour diffusion technique as the one to identify initial conditions of crystallization and optimization of crystal size and quality. In this setup, precipitant and sample are arranged as shown in the **figure M10.** In general terms, the vapor diffusion approach consists in a drop containing equal volumes of sample solution and crystallization condition and is equilibrated against a reservoir solution, both isolated from the external environment. The crystallization solution in the drop is diluted twice as compared to the reservoir due to the protein solution. Thus, because of osmolarity differences, water is evaporated from the sample drop and absorbed by the reservoir, so that both the protein and precipitant concentration increases in the droplet until the osmotic pressure between reservoir and sample drop reaches equilibrium. During this process, the sample travels through the phase diagram and may reach the nucleation point, leading to crystals. As it is shown in **Figure M10**, there are two types of vapor diffusion setups, the sitting and the hanging drop. It is not possible to foresee the crystallization condition for a new protein. Therefore, extensive crystallization trials are performed to find promising conditions in which crystalline precipitation, phase separation, nuclei or crystals appear.



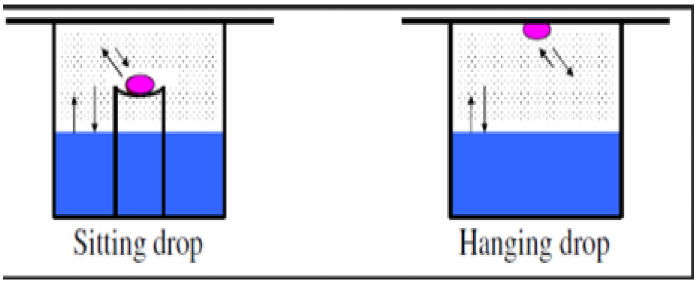


Figure M10: *on top a schematic image depicting the main experimental setup for vapor diffusion crystallization: the hanging and sitting drop). On the left, the different regions of the phase diagram and examples of drops containing protein at each of them.*

Extracted from: [172] A. McPherson and J. A. Gavira, 2013. Acta Crystallographiva Section F

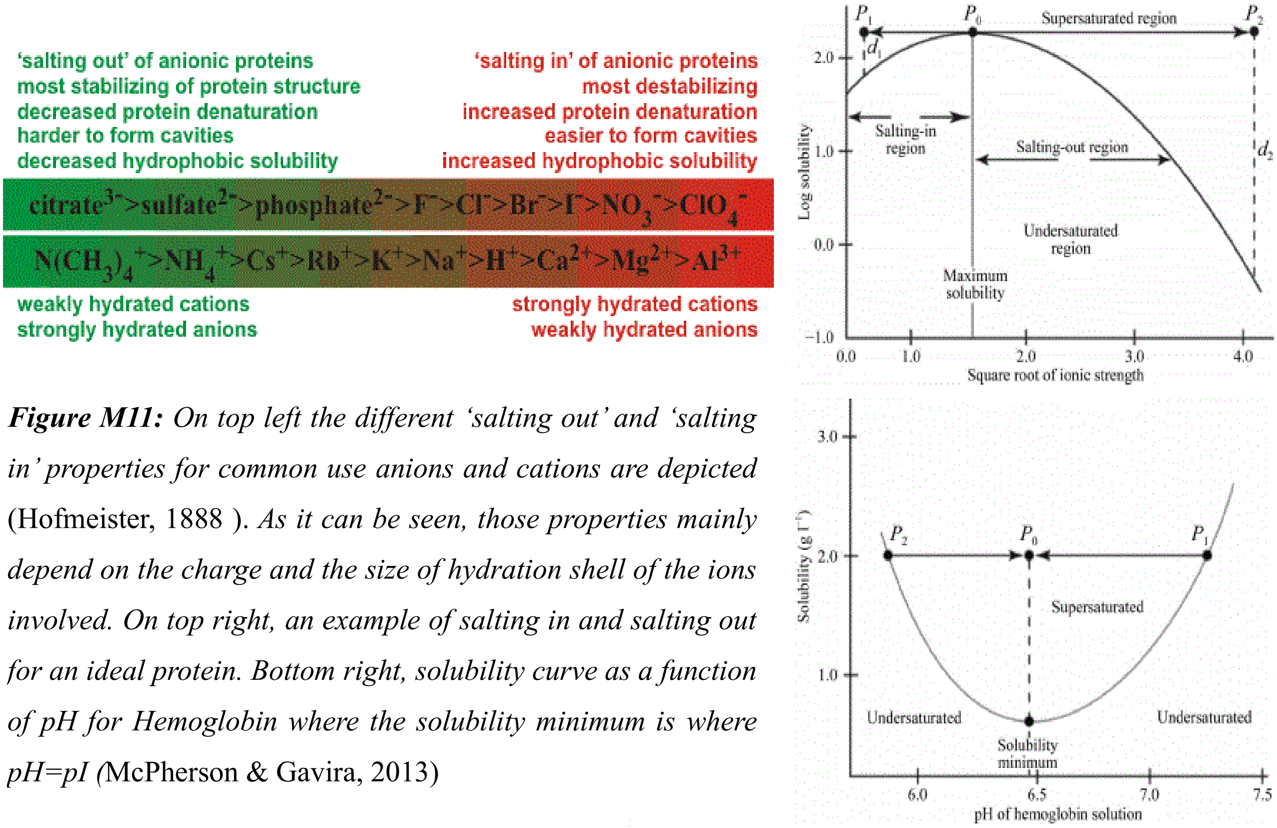


Figure M11: *On top left the different 'salting out' and 'salting in' properties for common use anions and cations are depicted* (Hofmeister, 1888). *As it can be seen, those properties mainly depend on the charge and the size of hydration shell of the ions involved. On top right, an example of salting in and salting out for an ideal protein. Bottom right, solubility curve as a function of pH for Hemoglobin where the solubility minimum is where pH=pI (*McPherson & Gavira, 2013)

The vapor diffusion technique is nowadays automatized, which extraordinarily facilitates the screening for initial crystallization conditions in 96-well plates. At the Automatized Crystallization Platform from the Barcelona Science Park, several automatic dispensers are available.

Crystallization screening conditions contain a precipitant compound, a buffer to stabilize the pH, and one or more additives such as organic molecules or polycations, among others. All these components will affect the final solubility of the protein. "Classical" precipitants include salts, notably ammonium sulphate. These ionic compounds, either organic or inorganic, affect protein solubility through the phenomena of salting in (increasing ionic strength stabilizes external charged residues) and salting out (increasing ion concentration sequesters water molecules from the protein surface, promoting aggregation). Accordingly, the solubility of proteins as a function of ionic strength reaches a typical maximum at the point where the salting out phenomena starts to overcome the salting in effect. This maximum of solubility, as well as sample stability, depends on the sample and the ions involved as shown in **Figure M11** (Hofmeister, 1888). By changing the ionic strength towards the salting out region promotes aggregation and, occasionally, crystal formation. Despite salting out is the most common approach when crystallizing a protein using a salt precipitant, it is remarkable that salting in (by changing the ionic strength below the solubility maximum) is also often used to promote crystallization using a dialysis setup (McPherson & Gavira, 2013).

Another classical compound present as a precipitant in the crystallization screens are the hydrophilic organic polymers polyethyleneglycol of different molecular weight (PEG400, PEG1000, PEG12000 etc.). PEGs promote protein aggregation through a combination of two physicochemical phenomena. On one side, PEGs are highly hygroscopic, so they remove a significant amount of water from the protein hydration shell. On the other, due to their elongated and/or branched structure they have a high hydrodynamic radius which reduces the effective available volume of bulk solvent thus generating a crowding effect that also promotes particle aggregation. Differently to salting in and salting out phenomena, the plot of protein solubility as a function of non-polar organic precipitants is always decaying. Also, contrarily to salt-based precipitants, PEGs don't significantly increase the ionic strength of the sample. For this last reason, PEGs are normally chosen for the crystallization of macromolecular complexes, especially those that are formed upon salt bridges e.g. most of protein-DNA complexes.

Protein solubility is also directly related with the pH of the solution (see **FigureM11**). In this case, protein solubility reaches a minimum when the media pH reaches the isoelectric point (pI) of the particle and its overall charge is neutralized (McPherson & Gavira, 2013).

Lastly, the third main component of a crystallization condition is the additive. Those may be small molecules of different nature that may increase the quality of the crystal by the introduction of extra crystal contacts (e.g. silver bullets) or by finely modulating the solubility of the sample (salts, polar compounds, chaotropes or cosmotropes). Some additives may also be salts containing heavy atoms to co-crystallize with the sample in order to perform experimental phasing as discussed in the diffraction physics section chromatography (*See phase determination by SAD,* **Materials and methods 10.3.2**).

M9.2- Crystallization of Gcf1p in complex with DNA

Taking into account the results of Gcf1p characterization in solution as well as the available information of Gcf1p homologs Abf2p and TFAM (Rubio-Cosials, et al., 2011) (Rubio-Cosials, et al., 2011) (Ngo, Kaiser, & Chan, 2011) (Chakraborty, et al., 2017) it was clear that crystallization would only be possible in the absence of DNA.

Gcf1p complexes whre prepared at protein:DNA ratios of 1:1.2 and 2:1.2 following the three-step dialysis protocol described in previous sections (see section **M7.2**)Tested DNA substrates are summarized in **Table M2**. Once the complexes were formed, the screening for the optimal crystallization condition was performed in the

Automated Crystallography Platform (PAC) from *Parc Científic de Barcelona* (PCB-UB) with help and assistance of Dr. Joan Pous and Xandra Kreplin. ®). By use of automated drop dispensing robots *Cartesian* and *Phenix* (TECAN Life Sciences ®), we screened in parallel the conditions for different complexes. Different sparse matrices were tested, which are designed based on available screenings and prepared by by the PAC personnel, mainly: *PAC1, PAC2, PAC5, PAC10, ProPlex* and *Top96* (Hampton Research ® and Molecular Dimensions Diffracting protein-containing crystals were obtained only for the following DNA sequences: *Af2_20, Af2_20Shift_1* and *Af2_20Shift_2* and only in condition F2 from PAC-2 crystal screen: 20% PEG1000, 0.1M Na/K-Phosphate pH6.2 and 0.2M NaCl. Interestingly this was also the condition for which first initial hits were obtained for TFAM in complex with the human mitochondrial Light Strand Promoter.

In order to get bigger crystals, crystallization drops were set manually using both sitting and hanging drop approaches, and 1 µl volumes for both protein solution and chemical condition. In parallel to this, screening additives from the *96-additive kit* (Hampton Research ®) where tried and some of them improved crystal size and cryogenic protection. Since the final structure solution required multi-crystal averaging, the conditions in which those crystals where obtained are the following:

Crystals were obtained in the following conditions:

For native protein (Sitting drop in all cases):

20%PEG1000, 0.1M Na/K-Phosphate pH6.2 and 0.2M NaCl

20%PEG1000, 0.1M Na/K-Phosphate pH6.2 and 0.2M NaCl, 50mM $FeCl_2$

20%PEG1000, 0.1M Na/K-Phosphate pH6.2 and 0.2M NaCl, 3% Trehalose

For Seleno-derivative protein (Both Hanging and sitting drop)

28%PEG1000, 0.1M Na/K-Phosphate pH6.2 and 0.2M NaCl, 3% Trehalose

28%PEG1000, 0.1M Na/K-Phosphate pH6.2 and 0.2M NaCl, 3% Glycerol

18%PEG1000, 0.1M HEPES-Na pH7.5 and 0.2M NaCl, 6% Glycerol

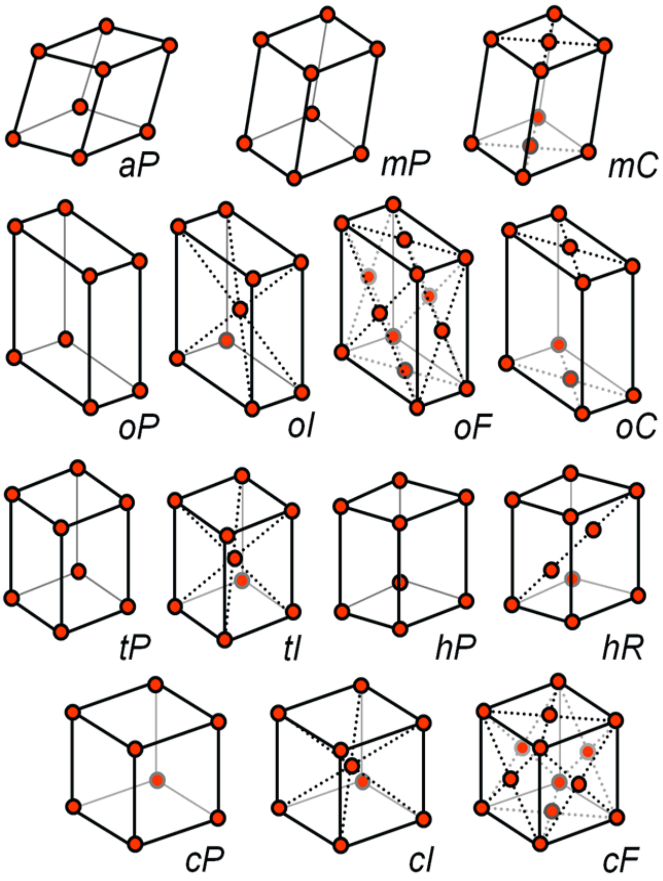
M10. DIFFRACTION PHYSICS, DATA COLLECTION AND STRUCTURE SOLUTION

M10.1- Theoretical background

M10.1.1- Crystal symmetry

A crystal is a solid state of matter in which all its components are arranged with the same orientation in a continuous three-dimensional array. In this arrangement, one can define the minimal operator to geometrically describe the system. Such a mathematical operator is termed lattice and is repeated *ad infinitum* along the crystal system. The translations that describe the periodicity of this lattice are defined by three basic vectors that are not coplanar and that form a parallelepiped referred as the unit cell of the crystal. This unit cell is the building block of the crystal, repeated in three dimensions. Its size and shape are defined by both the length of the three parallelepiped edges (a, b and c) and the angles (α, β, γ) between them.

The edges of the unit cell are indeed rotation axes of order *n*, which due to the lattice constrictions is limited to n=6, 4, 3, 2 or 1 (no symmetry). Combinations of these operators at any of the three a, b and c edges give rise to 32 different point groups (or crystal classes). In addition, cell edges are related with each other by angles of 90°, 120° or by a non-fixed angle. Thus, considering the rotational symmetry of axes and the cell angles, seven crystal systems appear: Triclinic, Monoclinic, Orthorhombic, Tetragonal, Trigonal, Hexagonal or Cubic. The seven crystal systems are further modified by the position of the centre of symmetry, which can be at the cell vertices; at one, or at all cell faces; or at the cell centre. This gives rise to the different 14 Bravais lattices, namely primitive (P), body-centered (I), face-centered (F), rhombohedral (R) or centered (C), as it is summarized in **Figure M13.** Finally, there is a further variation, which is a rotation combined with a translation, that gives rise to an improper rotation termed screw axis. The combination of the 14 Bravais lattices with the 32 different point groups, together with other symmetry elements such as screw axes, mirror planes and glide planes give rise to a total of 230 total space groups. However, since macromolecules show chirality (L-aminoacids), symmetry operators such as mirror planes an inversion centres, are excluded. Therefore, the space groups compatible with macromolecules are the 65 *non-enantiogenic* space groups. Space groups are the combination of all the symmetry operators needed to reconstruct the unit cell and the whole three-dimensional arrangement of a crystal from a single, smallest element without internal crystallographic symmetry: the asymmetric unit (ASU). The ASU may still show some internal non-perfect symmetry, which is known as non-crystallographic symmetry (NCS). Those NCS are not restricted to the lattice restraints and can be of any possible order (apart the ones above, also e.g. 5-, 7-, 11-, 21-fold symmetries or more), which is relatively common in macromolecular crystallography. It is often an indicator of the biological oligomeric unit and is usually helpful for structure solving and model building.



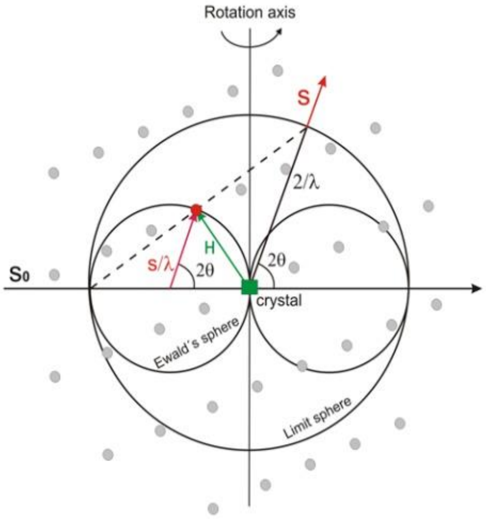


***Figure M13:** On the left, the seven crystal classes combined with the four possible lattice centerings give rise to the **14 Bravais lattices**. On the top images, Bragg's Law combined with crystal lattice gives rise to the construction known as the **Ewald's Sphere**. Grey dots symbolize the reflections collected, which depend on the crystal orientation.* Extracted from *[176]*. https://www.xtal.iqfr.csic.es/Cristalografia/index-en.html

M10.1.2- X-ray and X-ray diffraction

Diffraction is a physical phenomenon extensively described for both electromagnetic and mechanical waves. Diffraction is a combination of different events, firstly when a wave encounters an object in its trajectory, this object can oscillate in resonance with the incoming wave and thus become a new wave origin. If two wave origins are close enough to each other, phenomena of positive and negative interference occur, which generates a characteristic discrete diffraction pattern, as it was described by Young in its celebre double-slit experiment.

Diffraction of X-rays is originated by the elastic scattering of photons when interacting with electron orbitals from atoms (see, Thomson scattering in *SAXS theoretical background section*, **Materials and Methods M8.1**). When we consider a three-dimensional crystal system with ideally infinite planes, and the crystal is hit by an X-ray beam, the cumulative effect of both positive and negative interference between outgoing scattered waves from the different planes restricts the positive interferences, which appears as reflections, to those that accomplish the Bragg' s condition or Bragg' s Law. (Bragg & Bragg, 1913)

Eq.1: Bragg's Law

$$2dsin\theta = n\lambda$$

*Where d is the distance between crystal planes, n is an integer number, λ is the wavelength and θ is the scattering angle, as described with anteriority in this book (see, Thomson scattering in SAXS theoretical background section, **Materials and Methods M8.1**).*

An elegant and intuitive representation of the Bragg's condition is the Ewald's sphere construction (Figure M13). By rotation of the diffraction origin (the crystal) with respect to the incoming radiation source, a set of different reflections accomplish the diffraction condition at the different rotation angles. According to the Bragg's Law, the diffraction angle *θ* is inversely proportional to the distance between planes *d*. Hence, the smaller is the distance between the diffracting planes of the crystal the greater will be the diffraction angle *θ*. That is, the diffraction pattern (at the *reciprocal space*) generated by a three-dimensional crystal (at the *real space*) will show inverted reciprocity with the internal arrangement of such a crystal. The diffraction reflections, which belong to the *reciprocal space*, will follow the same symmetry pattern than that of the crystal unit cell in the *real space*. Being so, the diffraction pattern shows a *reciprocal unit cell* with axes that are inversely proportional to the crystal unit cell at the *real space* with 1/a, 1/b and 1/c cell parameters (also termed a*, b* and c*, respectively). In a perfect crystal system with a completely regular and infinite lattice, is related to the reciprocal space diffraction pattern or lattice (which arises from the Bragg's condition), by the Fourier transform, as depicted in **Figure M14**.

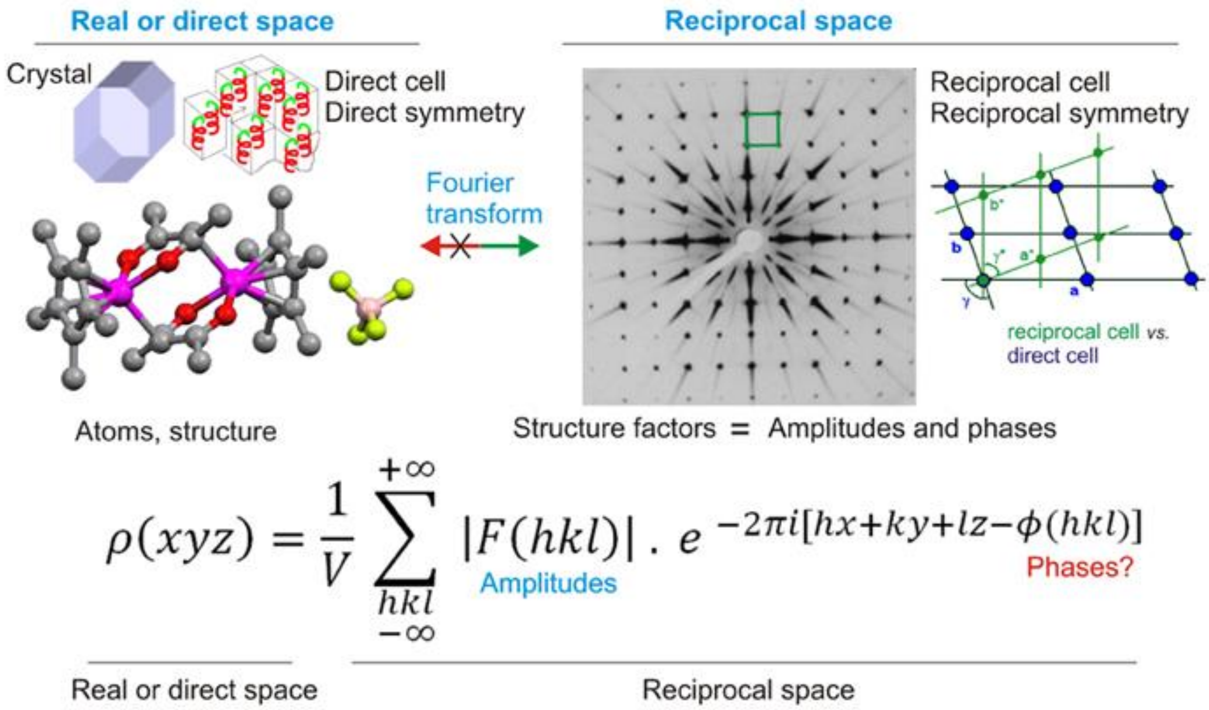


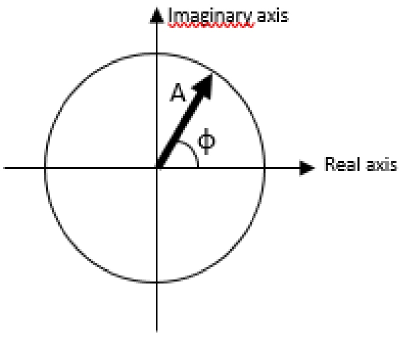
***Figure M14:** As shown at the webpage of Instituto de Química Física Rocasolano (IQFR-CSIC, Madrid) (IQFR-CSIC, 2020), the reciprocity between real space (the crystal) and reciprocal space (data obtained from diffraction images) are represented by the Fourier Transform equation.*

Ideally, with the Fourier Transform expression, real space information such as the electron density at a specific point ρ(xyz*)* can be calculated from reciprocal space parameters such as phase and amplitudes. However, X-ray diffraction only provides the amplitudes (the square root of the measured intensity of each reflection) but not the phases. A more detailed view of this two elements, amplitude and phase, of the wave function Ψ(r) and a more extended explanation of the Fourier Transform shown in **Figure M14** is provided in the next point.

M10.1.3- X-rays are electromagnetic waves with an amplitude and a phase.

X-ray radiation has electromagnetic nature. In accordance to this, two components can be described, an oscillating electric field, and an oscillating magnetic field perpendicular to the former. Both oscillating fields have the same wavelength, but they are out of phase by 90°. To include them in the same mathematic expression

Ψ a complex number is needed, and they are represented by the Argand diagram representing the real and the imaginary axis (**Figure M15**).



$$Eq.2{:}\ \Psi = A\cos(\varphi) + \boldsymbol{i}A\sin(\varphi)$$

$$Eq.3{:}\ \Psi(x) = Ae^{i\varphi}$$

$$(where\ \varphi = 2\pi\left(\frac{x}{\lambda}\right) and\ \boldsymbol{i}\ is\ defined\ such\ as\ i^2 = -1$$

Figure M15: *Argand diagram representing phase and amplitude for an electromagnetic wave, which is expressed in equations 2 and 3. The mathematic expression describing the wave Ψ can be expressed as the sum of sines and cosines of the phase (φ) contributed by coefficients amplitude (A) and the imaginary unit* ***i****.*

Using the Euler formula, *equation 2* in **Figure M15** can be simplified to *equation 3*, where A represents the oscillation amplitude, ϕ stands for the phase and λ for the wavelength of the radiation. Both factors amplitude and ϕ are necessary to reconstruct the wave function Ψ. Nevertheless, A is a real number and ϕ a complex number.

M10.1.4- Diffraction: path difference and phase difference. Introducing the phase problem in crystallography

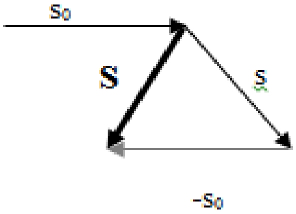
Only elastic scattering is assumed in this discussion. Therefore, the different incident and outgoing waves will maintain the original value of amplitude (A). In order to obtain the phase difference between them, we need to consider the difference in the path followed by the radiation. The difference between the incident beam (s_0) and the diffracted beam (s) will define this path difference, summarized in *equation 4* and *equation 5*. Where *r* is the magnitude of the position vector ***r*** defining the position of the scattering source point.

$$Eq.4{:}\ Path\ difference = r \cdot s_0 - r \cdot s$$

$$Eq.5{:}\ Phase\ difference = \left(\frac{2\pi}{\lambda}\right)(r \cdot s - r \cdot s_0)$$

The presented notation can be simplified defining a scattering vector **S.** It is the path difference between vectors s_0 and s as it is represented in the next scheme. Vector **S** is defined in *equation 6* and allows a more simplified description of the phase difference in *equation 7*.

When describing scattering of a real-life object, e.g. a molecule, a much more complex situation takes place, since the scattered X-rays from all atoms contribute to the final signal. The summation of this scattered beams is represented by the integral described in *equation 8*, where Ψ (S) is the resultant X-ray in the direction specified by the scattering vector S in the last construction, and the factor dV_r refers to the volume of the scattering voxel.



$$Eq.6: scattering\ vector\ \ S = \frac{1}{\lambda}(s - s_0)$$

$$Eq.7: Phase\ difference\ (S) = 2\pi(r \cdot S)$$

$$Eq.8: \Psi(S) = \int e^{i(2\pi(r\cdot S))} dV_r$$

In the former expressions 7 and 8, the phase difference (defined as 2π(r·S)) refers exclusively to the phase change induced by the path difference defined in the former vector constructions. In the context of radiation-matter interaction, the incident photon excites the electron cloud that resonates with the incident wavelength and produces a new, scattered X-ray photon. In this situation, amplitude does not change but a phase change component (Δφ(r)) is added to the scattered photon. The extra value of phase change Δφ(r) is proportional to electron density ρ (r); Δ ϕ (r)= σ ρ (r), being σ a scattering factor characteristic of the atoms involved in each individual scattering event.

This additional phase change Δ ϕ (r) is added to the last expression in the following series:

$$Eq.9: \Psi(S) = \int e^{i(2\pi(r\cdot S)+\Delta\varphi(r))} dV_r$$

$$Eq.10: \Psi(S) = \int e^{i\sigma\rho(r)} e^{i(2\pi(r\cdot S))} dV_r$$

For small phase angles, Taylor series allow us to simplify expφ≈φ.

$$Eq.11: \Psi(S) = \int i\sigma\rho(r)e^{i(2\pi(r\cdot S))}dV_r$$

$$Eq.12: F(S) = \frac{\Psi(S)}{i\sigma}$$

$$Eq.13: F(S) = \int \rho(r)e^{i(2\pi(r\cdot S))}dV_r$$

Throughout equations 11 to 13 we have defined the structure factor (F(S)) as the result of the division of the diffracted beam ($\Psi(S)$) divided by the complex number iσ. F(S) is therefore the Fourier Transform of the electron density. Isolation of the electron density parameter results in equation 14:

$$Eq.14: \rho(r) = \frac{1}{V_r}\int F(S)e^{-i(2\pi(r\cdot S))}dS$$

Therefore, the electron density of an object in *real space* is the inverse Fourier Transform of the structure factors at the *reciprocal space*, which are contained in the diffraction pattern. In addition, in a crystal, the same motif is periodically repeated in space infinite times, following the symmetry pattern. At the reciprocal space, such a periodicity corresponds to the Bragg planes (orthogonal to each other), each plane defined by the Miller indices h, k, l axes, which progress to the infinite. For a crystal, the expression of the inverse Fourier Transform is thus modified to *equation 15*, which was introduced in **Figure M14**.

Eq. 15:

$$\rho_{(xyz)} = \frac{1}{V}\sum_{\underline{hkl}\atop -\infty}^{+\infty}|F(hkl)|e^{-2\pi i\,[hx+ky+lz-\varphi hkl]}$$

Where the term hkl refers to the coordinates at the reciprocal space, and the real term F(hkl) is the structure factor amplitude at that point.

The values of the amplitudes of diffracted photon waves are equal to the squared root of the intensities recorded on the detector. Therefore, the F(hkl) term can be directly extracted from the diffraction pattern by applying the square root to the recorded intensities, a step termed *truncation*. A different situation applies to phases φ, since they are a complex number that is not contained in the diffraction pattern. Unluckily for crystallographers, there are no lenses that can re-focus the diffracted X-rays like it was described for the electrons previously (*see TEM theoretical background section,* **Materials and methods M6.2**). This gives rise to the so-called *phase problem* in crystallography, by which the values for the phases of the diffracted photons are lost in the diffraction experiments and must be calculated by using different strategies.

M10.1.5- Reciprocal space centrosymmetry, Friedel’s law and anomalous signal.

The first studies on the properties of the X-ray diffraction that patterns were centrosymmetric (Friedel, 1913). Therefore, every F_{HKL} has an identical partner, with the same intensity, at inverted coordinates $F_{-H,-K,-L}$. Indeed, all symmetry partners of F_{HKL}, which are F_{-HKL}, F_{H-KL}, F_{HK-L}, F_{-H-KL}, F_{-HK-L}, F_{H-K-L}, and F_{-H-K-L} share the same intensity. This is because of the following expressions, in which the function F(hkl) is elevated to the square (I), so the imaginary part in the Fourier Transform f(xyz) disappears:

Friedel’s law shows that every F has a F^- by centrosymmetry and an antisymmetric φ

Eq. 16: $F(hkl)=\int_{-\infty}^{+\infty} f(xyz)e^{i[hx+ky+lz]}dx$

Eq. 17: $|F(hkl)|^2=|F(-h-k-l)|^2$

Eq. 18: $\varphi\ (hkl)= -\varphi\ (-h-k-l)$

Thus, by virtue of the Friedel’s law, the function F(hkl) will be centrosymmetric implying that for a reflection hkl exists an identical reflection –h-k-l (its Friedel pair), which has the same amplitude but inverse, 180° shifted, phase value. However, the Friedel law is broken due to the phenomenon known as anomalous dispersion, which takes place when a scatterer is irradiated with an X-ray beam with the same energy as the absorption edge of one of its electron shells. The energy of the photon is absorbed by the electrons at this shell, so that the electrons promote to an upper, unoccupied shell. Since part of the incoming energy is thereby absorbed, the scattering factor f_0 of the atom is altered to a new f, according to the following equation (see below for parameter defintition):

$$f= f_0+f' +\mathbf{i}f''$$

The new scattering factor f is a complex number formed by the dispersive component f' (real) and the anomalous component if'' (complex). The dispersive component is 180° out of phase with respect to f_0 while the anomalous component is 90° out of phase. This effect is used for the calculation of the phases by experimental methods (*see solving the phase problem using SAD section*, **Materials and methods M10.3.2**).

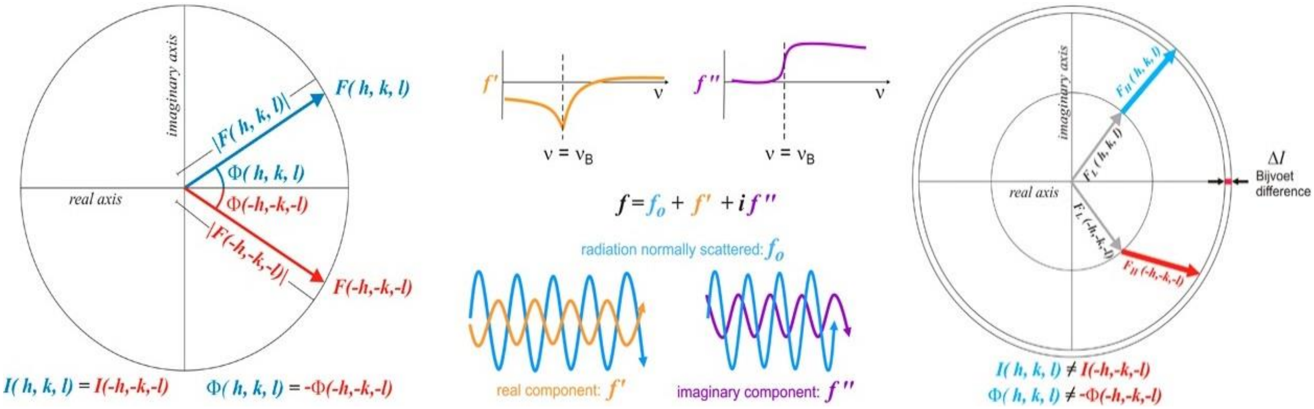


Figure M16. Friedel's law and anomalous diffraction*, from left to right, the Argand diagram shows two symmetrically related pair of reflections that follow the Friedel's law since they have identical amplitudes. At the central image, when the X-ray energy is at the absorption edge of an atom, there is an enhancement of the anomalous component f" that introduces a 90^{o} shift thus breaking Friedel's law as it is shown in the Argand diagram of the right. In addition, at the inflection point of the f"curve, the dispersive component f' reaches its minimum*. Extracted and adapted from https://www.xtal.iqfr.csic.es/Cristalografia/parte_07_4-en.html

M10.2- General concepts regarding data treatment

M10.2.1- Getting the best of our data: data processing and quality indicators in crystallography

Data set collection involves the rotation of the crystal by a certain number of degrees, depending on the crystal orientation and space group. The total oscillation refers to the total number of degrees collected, whereas the oscillation range are the degrees collected per image. Depending on the intensities of the reflections, which is related to the crystal quality, the diffraction data is collected at specific exposure time and intensity of the incident beam, both parameters experimentally determined during the diffraction experiment. Upon dataset collection, all the reflections recorded on the detector are processed so that a 3D image of the reciprocal space is reconstructed, and a complete data set of reflections is obtained. To perform this task, one of the programs available is XDS (Kabsch, 2010), or DIALS. In XDS, the 3D integration of a reflection intensity (I) collected in several images is performed with a fine estimation of the intensity at each 2D image. Because the oscillation range is small in each image, a reflection is collected in several images but the signal-to-noise ratio for the corresponding I is enhanced.

During processing of the data, the indexing of the reflections is the step by which the h, k and l indexes are assigned to each intensity, which corresponds to a reflexion. Afterwards, a Bravais lattice is assigned to the data, to make an initial guess of possible point (Laue) groups. Scaling the intensity of each reflexion throughout the total oscillation follows. At this point, the symmetry-related intensities are still unmerged. Data reduction programs such as Aimless (Evans & Murshudov, 2013) merge the intensities of symmetrically related reflections and further assess data quality by using different statistical operators. Truncation of merged intensities to the structure factors F(h) is also performed at this point (French & Wilson, 1978). Prior to this, a check of systematic absences is performed (Evans, Scaling and assessment of data quality, 2006) in order to identify the presence of screw axes, thus stablishing the most probable space group. In any case, space group is a hypothesis until the definitive structure determination.

Descriptors of data quality

Prior to solving the phase problem, it is important to evaluate the quality of the data. Several statistic operators exist in order to check data quality. Here I present the most relevant ones:

Completeness and multiplicity:

Completeness is defined as the relation between observed over expected reflections. A good completeness would be above 98% at low resolution while a completeness around 60-40% could be acceptable at high resolution (more than 90% overall). Low resolution completeness is very important for the crystallographic phase determination.

Multiplicity is defined as the ratio of the total number of measured observations over the unique reflections. In other words, how many times a reflection is measured during data collection. In addition, the intrinsic crystal symmetry makes possible that the symmetry-equivalent reflections that are measured and expected to be have the same intensities, are merged. Multiplicity is not *per se* a data quality indicator, but a better accuracy in the determination of the intensity value of a reflection is expected at higher multiplicity, as well as decreasing the noise associated to the measurements.

I/σ(I) and dAno/sigAno:

The signal-to-noise ratio is defined as the ratio between the intensity (I) compared to the noise (σ(I)), which is associated to the background of this intensity. The I/σ(I) parameter indicates how strong is the reflection intensity in the dataset. With former detectors, the reflections had to have $I/\sigma(I) \geq 2$, and the weakest portion of the dataset was discarded due to unreliability. Today, with more sensitive detectors, this parameter is not critical in determining the quality of the data set, albeit is taken into account in the overall balance between parameters. Other, newer parameters are considered to select the reliable measurements, such as R-meas and CC* (both explained below), which depend more on the internal consistency of the reflections than on the relative strength of such reflections (Diederichs & Karplus, 2013). New structure solution programs like *Phaser* (McCoy, Grosse-Kunstleve, Adams, Winn, Storoni, & Read, 2007) succeed in dealing with weak data by use of data corrections that are not based on the elimination of weak reflections but are pondered based on the agreement with the whole dataset. Similarly, the indicator dAno/SigAno is used in order to determine which range of data has useful anomalous information for experimental phasing.

R_{merge} and R_{meas}:

Historically, the indicator R_{merge} has been used extensively as the reference indicator of aggregated statistical properties. This is a normalized residual that indicates consistency between measurements. It is the summation of the absolute difference between individual reflection intensity I to the I mean (between equivalent reflections), normalized by the summation of the intensities (Arndt, Crowther, & Mallett, 1968)

$$R_{merge} = \frac{\sum_i \sum_{j=1}^{n_i} \left| I_j(hkl) - \overline{I(hkl)} \right|}{\sum_i \sum_{j=1}^{n_i} I_j(hkl)}$$

> *Regardless of the extensive use of R_{merge}, it was proved that it performs poorly with high multiplicity of the data. A more robust indicator, $R_{meas,}$ which is independent of data multiplicity, was introduced (Diederichs & Karplus, 1997). A factor $\sqrt{n_i/(n_i-1)}$ (where n represents the number of reflections within the set) is applied to modify each factor of the R_{merge} equation. Thus, the R_{meas} for reflection I_i (h) (h refers to the Miller index hkl) is as follows:*

$$R_{meas} = \frac{\sum_h \sqrt{\frac{n_h}{n_h - 1}} \sum_i^{n_h} \left| \hat{I}_h - I_{h,i} \right|}{\sum_h \sum_i^{n_h} I_{h,i}}$$

$$\text{where } \hat{I}_h = \frac{1}{n_h} \sum_i^{n_h} I_{h,i}$$

R_{meas} is a robust indicator for unmerged data quality. Similar indicators for data after merging, which compares the symmetry-related reflections, are $R_{mrgd\text{-}I}$ for Intensities, and $R_{mrgd\text{-}F}$ for amplitude structure factor (F). They are both robust and show how data quality improves with higher multiplicity (Diederichs & Karplus, 1997).

$CC_{1/2}$, CC^* and CC_{ano}:

$CC_{1/2}$ is a Pearson correlation coefficient that indicates the agreement between the intensities of two randomly selected half-datasets. CC* gives information about the agreement between the experimental data and the underlying true signal. Both indicators provide information about the statistical significance of the reflections of our dataset regardless of reflections strength/weakness (Diederichs & Karplus, 2013).

*Correlation coefficients $CC_{1/2}$ and CC**

$$CC_{1/2} = cov\frac{(X,Y)}{\sigma_x \sigma_y}$$

$$CC^* = \sqrt{\frac{2CC_{1/2}}{1 + CC_{1/2}}}$$

Similarly, a correlation coefficient for the anomalous signal (CC_{ano}) provides the agreement of two randomly selected half-datasets treating each Friedel pair as independent reflections. Significance of this value depend on the number of reflections considered but values greater than 0.3 for $CC_{1/2}$ are taken as acceptable.

M10.2.2 Descriptors of data quality and crystal pathologies: anisotropy

When processing data, the values of data quality descriptors may deviate from the acceptable. This might be due to intrinsic problems of the crystals that introduce systematic errors, and which can be circumvent by applying corrections to X-ray diffraction datasets. In this thesis project, we had to deal with moderate anisotropy of the data. Anisotropy is defined as a deviation from the ideal sphericity of the reciprocal space. In anisotropic datasets, data strength is direction dependent and one or two of the reciprocal axes show significantly weaker reflections. This can be due to the preferential growth of the crystal towards one direction (thus, there is more crystal mass diffracting in that direction) or due to an intrinsic lack of crystal contacts (i.e. disorder) in one or two of the axes of the unit cell.

Different approaches are developed to deal with data anisotropy. One is based in the definition of an ellipsoid that excludes weak data, such as in *Staraniso* (Tickle, et al., 2018) (Global Phasing Ltd.). Or, weaker reflections are enhanced based on the internal agreement within the rest of the dataset, such as in *Phaser* (McCoy, Grosse-Kunstleve, Adams, Winn, Storoni, & Read, 2007).

M10.3- Solving the crystallographic phase problem

M10.3.1- Phase determination by molecular replacement (MR)

Molecular Replacement (MR) approach consists in the use of calculated phases from structures of homolog or similar macromolecules to solve the phase problem. This is a rather old technique [122] that is nowadays the most commonly used technique for structure solution due to the increase of available structures in the Protein Data Bank (PDB) as well as the improvements of the available software.

In general, molecular replacement searches are done in two independent steps, in one the searching model is rotated at the origin of the cell to search for the orientation/s that best fits the orientation of the molecule in the crystal, which is assessed by the agreement between the calculated and the experimental intensities. The second step consists in translation searches to position the rotated molecule in the crystallographic cell. In the case of the program *Phaser-MR*, after the respective searches a Z-score for both rotation function (RFZ) and translation function (TFZ) are computed as well as a log-likelihood gain (LLG) indicating how well the model-derived reflections fit with the experimental ones (McCoy, Grosse-Kunstleve, Adams, Winn, Storoni, & Read, 2007).

M10.3.2- Phase determination by single wavelength anomalous diffraction (SAD)

Single and Multiple Wavelength Anomalous Diffraction (SAD and MAD, respectively) are methods to solve the phase problem and both exploit the anomalous dispersion phenomena (*see Friedel's Law and anomalous diffraction section,* **Materials and methods M10.1.5**). The phase values are calculated experimentally from the diffraction data from a single crystal. In order to maximize the anomalous effect, the energy of the X-ray beam is tuned to the absorption edge of one of the atoms present in the crystal. It is worth noting that any atom can generate anomalous dispersion at (or near to) its atom-specific absorption edge wavelength. This requires an X-ray source that can reach such an appropriate wavelength. The X-ray energies usually available in most of the synchrotron sources range from 7 to 15 keV. One strategy is to introduce atoms of a Z significantly higher than the ones naturally present in the macromolecule, which is achieved either by soaking the crystal into a heavy metal containing solution (Hg, Au, Pb) or by the incorporation of higher atomic number elements to the primary structure of the protein, typically by substitution of L-methionines for Seleno-L-Methionines for proteins, or Br-thymine for DNA.

SAD data is collected only at one wavelength, typically the value of choice is that at which f" reaches its maximum based on the X-ray fluorescence profile of the sample as exemplified in **Figure M16**. In MAD strategy, additional data sets are collected at different wavelengths exploiting both anomalous f'' and dispersive f' differences (**Figure M16**) to perform univoque phase determination. SAD only uses measurements at one wavelength and thus there is an intrinsic 180° phase ambiguity that must be solved. Current methods for density modification are generally able to solve this ambiguity by improving first phase calculation. Common density modification procedures are solvent flattening, NCS-averaging and histogram-matching. Visual inspection of the connectivity of the density upon density modification will give the final assessment of appropriate crystallographic phase resolution.

M10.4- Diffraction data collection and structure solving of Gcf1p/DNA crystals

Native crystals were cryo-protected by stepwise addition of higher concentrations of PEG Low-Molecular Weight Smear, up to a maximum of 26%, or PEG400 up to 26%. depending on the case. During X-ray diffraction, crystals were kept under cryogenic conditions by using a N_2 cryostream to preserve crystal stability. Datasets were collected at ID-30 Beamline from the European Synchrotron Radiation Facility (ESRF), Grenoble (France); ID-04 and ID-24 from Diamond Light Source, Didcot (United Kingdom); and I02 from DESY Hamburg (Germany). These data collections were crucial to discern the quality of the crystals. All the datasets used to attempt structure solution and model building were collected in XALOC Beamline at the synchrotron ALBA, Cerdanyola (Spain).

Up to 20 useful datasets were collected from native crystals with an oscillation range of 0.1° at a beam energy of 12 KeV. Datasets were collected at a nominal resolution of 2.2 Å. For seleno-derivative crystals, up to 15 datasets were collected at a nominal resolution of 2.2 Å with an oscillation range of 0.1° or 0.5°, depending on the case. Datasets for seleno-derivatives were collected aiming at high multiplicity and thus crystal integrity was preserved by collecting at a transmission attenuated to 10%, with a flux of 10^{12} photons/s. Prior to data collection, the selenium absorption edge was determined experimentally. This is crucial since the absorption edge of Se bound to a protein may shift with respect to the theoretical one due to the local environment. Absorption edge was typically determined at 12.661 KeV, not far from the characteristic K-edge of free selenium atoms (12.616 KeV).

M10.4.1- Data processing and space group determination

Data was integrated with the XDS suite. Different datasets were combined and scaled using XSCALE (Kabsch, 2010), treating the Friedel pairs separately in dataset from seleno-derivatives. The space group was determined with Pointless (Evans, Scaling and assessment of data quality, 2006) and datasets were merged and truncated using Aimless (Evans & Murshudov, How good are my data and what is the resolution?, 2013). Staraniso server

(Tickle, et al., 2018) was used to determine the maximum resolution limit of the best diffracting reciprocal axes and generate an elliptically truncated dataset.

M10.4.2- Structure determination using MR-SAD and model building

MR searches were performed with Phaser (McCoy, Grosse-Kunstleve, Adams, Winn, Storoni, & Read, 2007) by using the data from Se-Met derivatives without imposing previous corrections nor imposing resolution limits. Search of heavy atom sites was performed with Phaser EP using the MR partial solutions. Initial phases were improved by density modification (DM) as implemented in Phenix_Resolve, which rendered Fourier maps that allowed us to trace fragments of α-helices. Phenix-Autobuild was used to trace protein residues in such initial Fourier synthesis. The molecules were fully built by iterative cycles combining manual model building in the real space with Coot, alternated with reciprocal space refinement with Phenix_Refine (Liebschner, et al., 2019)

RESULTS

R1 SEQUENCE ANALYSIS AND SECONDARY STRUCTURE PREDICTION

R1.1 Sequence analysis of Gcf1p

The product of *Candida albicans* gene GCF1, i.e. Gcf1p (Uniprot reference Q59QB8-1) is a 245-residue long protein with a molecular weight of 28486 Da. The sequence includes a mitochondrial targeting sequence (MTS) that was predicted by both servers TargetP-2.0 [123] [124] and Mitoprot II [125] [126], yet both programs did not fully agree on the exact MTS length. According to TargetP-2.0, the MTS spans between residues 1-25, whereas Mitoprot II predicts residues 1-33, with a probability of 99.17% and 99.46%, respectively. The construct was cloned by Joachim Gerhold from the laboratory of our collaborator Juhan Sedman (University of Tartu, Estonia), it spanned from residues 25-245 and it is the one that we, from now on, will also term 'full-length' protein as it is the whole functional part of the protein once inside mitochondria. Thus, MTS will be considered for residue annotation, i.e. residues 1-245, thus following the Uniprot entry Q59QB8-1.

Analysis of the 25-245 protein sequence with the server ProtParam [127] indicated a molecular weight of 26047 Da, an extinction molar coefficient of 38390 $M^{-1}cm^{-1}$ at 280 nm (ε), and a calculated isoelectrical point of 9.83. Sequence composition analysis reveals a total of 54 basic residues considering arginine and lysine residues, the two latter representing the 21% of the total sequence. Sequence is poor in sulphur-containing residues -no cysteines and 2 methionine residues only.

R1.2 Design of mutants: secondary structure prediction

Secondary structure prediction with JPred4 [128] [110] yielded 8 alpha helices with a score greater than 6 (the possible maximum score is 9) for residues 64 to 73, 79 to 101, 114 to 119, 125 to 132, 136 to 156, 167 to 175, 183 to 194 and 197 to 240. JPred4 suggested lack of secondary structure for the N-terminal residues 25 to 63 and the presence of a coiled coil forming region for residues 62 to 102, thus encompassing the two first alpha helices predicted by jnetpred. Likewise, coiled coil analysis of residues 25 to 105 using Coils server [127] [129] predicted a coiled-coil between residues 62-103 with a probability greater than 98%, using MTDIK scoring matrix with a window of 28 residues. Predictions of disordered regions and globular domains were performed using the IUPred2.0 [108] server, which suggested a globular domain between residues 114 and 245. Residues 25 to 114 were predicted as disordered, with scores between 0.7 and 0.8 over 1 for residues 36-58, and around 0.5 for residues 71-113.

R1.3 Design of mutants: Definition of Gcf1p domains

In order to explore potential globular domains, we performed a search for similar protein domains within Pfam database [130] [131] but no significant matches were found. Distant matches were nevertheless found for some regions of the proteins, residues 182-222 from Gcf1p matched with box 1 from HMGB1 protein with an E-value of 0.00079 whilst residues 128-156 scored an E-value of 16 when compared with the same protein. In addition, residues 61-99 matched with coiled-coil regions of *Saccharomyces cerevisiae* autophagy-related proteins Atg16 and Apg6 [132]) with E-values of 0.057 and 0.16 respectively.

Alignment of the Gcf1p sequence with HMG-box proteins from mitochondria or the nucleus with available 3D structures, including TFAM, SRY and Abf2p, was performed using Expresso [133] [134], which revealed a clear signature for an HMG-box domain at residues 180-240 that consisted in three α-helices separated by short linkers of 2 and 4 residues. Intriguingly, while the rest of the Gcf1p sequence did not align with any of the TFAM, SRY and Abf2p HMG-boxes, JPred also predicted three helices for residues 114-156. At this point we named the region as ‘helical region’. In conclusion, our predictions suggested four domains for Gcf1p. A first domain consisted in an unstructured N-terminal tail covering residues 25-62, followed by a coiled-coil domain comprising residues 63-103. A globular region between residues 114 and 245 was predicted to fold into two different domains involving residues 114-156 (helical region) and 180-240 (HMG-box) respectively (see **Figure R1**).

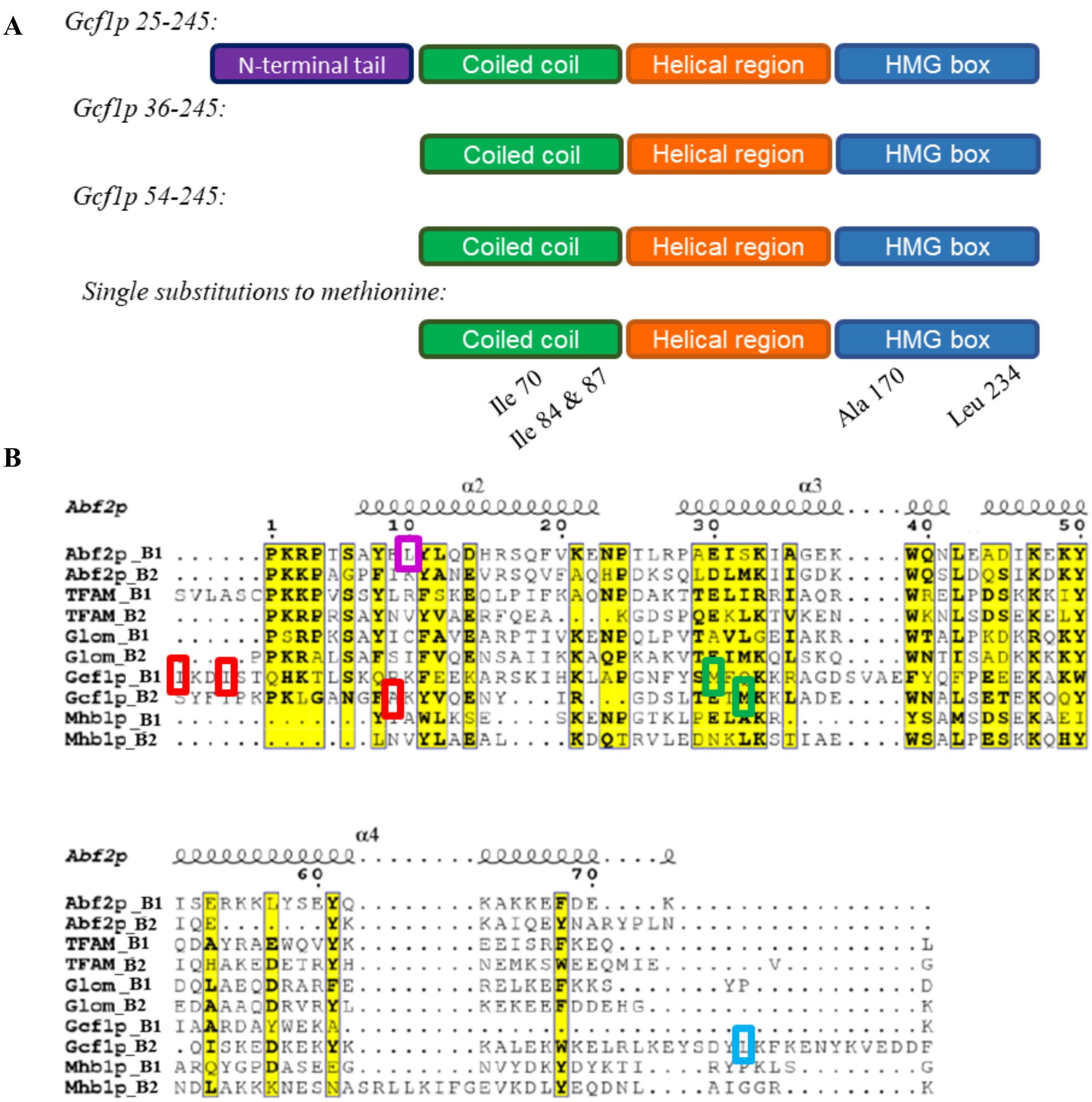

A
Gcf1p 25-245:
N-terminal tail
Coiled coil
Helical region
HMG box
Gcf1p 36-245:
Coiled coil
Helical region
HMG box
Gcf1p 54-245:
Coiled coil
Helical region
HMG box
Single substitutions to methionine:
Coiled coil
Helical region
HMG box
Ile 70
Ile 84 & 87
Ala 170
Leu 234
B
Abf2p
α2
α3
1
10
20
30
40
50
Abf2p_B1 ......PKRPTSAYFLYLQDHRSQFVKENPTLRPAEISKIAGEK....WQNLEADIKEKY
Abf2p_B2 ......PKKPAGPFIKYANEVRSQVFAQHPDKSQLDLMKIIGDK....WQSLDQSIKDKY
TFAM_B1 SVLASCPKKPVSSYLRFSKEQLPIFKAQNPDAKTTELIRRIAQR....WRELPDSKKKIY
TFAM_B2 ......PKRPRSAYNVYVAERFQEA....KGDSPQEKLKTVKEN....WKNLSDSEKELY
Glom_B1 ......PSRPKSAYICFAVEARPTIVKENPQLPVTAVLGEIAKR....WTALPKDKRQKY
Glom_B2 ....PPKRALSAFSIFVQENSAIIKKAQPKAKVTEIMKQLSKQ....WNTISADKKKKY
Gcf1p_B1 IKDISTQHKTLSKQRKFEEKARSKIHKLAPGNFYSMFQKKRAGDSVAEFYQFPEEEKAKW
Gcf1p_B2 SYFIPKPKLGANGFAKYVQENY...IR...GDSLTEIMKKLADE....WNALSETEKQQY
Mhb1p_B1 ...........YYAWLKSE....SKENPGTKLPELAKR.........YSAMSDSEKAEI
Mhb1p_B2 ..........LNVYLAEAL....KDQTRVLEDNKLKSTIAE....WSALPESKKQHY
Abf2p
α4
60
70
Abf2p_B1 ISERKKLYSEYQ.........KAKKEFDE....K.................
Abf2p_B2 IQE.........YK.........KAIQEYNARYPLN...............
TFAM_B1 QDAYRAEWQVYK.........EEISRFKEQ.................L
TFAM_B2 IQHAKEDETRYH.........NEMKSWEEQMIE.......V..........G
Glom_B1 DQLAEQDRARFE.........RELKEFKKS........YP..........D
Glom_B2 EDAAAQDRVRYL.........KEKEEFDDEHG..................K
Gcf1p_B1 IAARDAYWEKA...............................K
Gcf1p_B2 .QISKEDKEKYK.........KALEKWKELRLKEYSDYIKFKENYKVEDDF
Mhb1p_B1 ARQYGPDASEEG.........NVYDKYDYKTI....RYPKLS.........G
Mhb1p_B2 NDLAKKKNESNASRLLKIFGEVKDLYEQDNL....AIGGR..........K

Figure R1. Results from sequence analysis: *(A) A summary of the constructs most used along this thesis work. The different domains appear as predicted using pfam. The 'helical domain' was not predicted to be an HMG box according to previous results* [102]. ***(B)*** *Results of the sequence alignment with different HMG boxes that was used for the design of single-residue mutations to methionine. Purple square corresponds to the leucine that was successfully mutated to methionine in Ábf2p* [42]. *Red squares correspond to I84, I87 and A170, mutated by us to methionine, these mutants did not produce crystals at all. Green square corresponds to L209, mutants L209M yielded crystals, albeit with poor diffraction quality.*

R2 PRODUCTION OF GCF1P AND ITS MUTANTS

R2.1 Mutagenesis and subcloning

Deletion mutants of Gcf1p were generated by removal of the following segments of the N-terminal tail, namely 25-35, 25-43, 25-53 and 25-63, thus generating the constructs **Gcf1p 36-245**, **Gcf1p 44-245**, **Gcf1p 54-245** and **64-245** (Figure **R1A**). Single point mutations to methionine were generated in positions Ile70, Ile84, Ile82, Ala170 and Leu234, thus generating mutants **I70M**, **I84M**, **I87M**, **A170M** and **L234M**. Double and triple methionine mutants were also obtained by combination of the previous ones, producing mutants with up to 5 methionine residues in its sequence. The criteria for the substitutions was based on the alignment with TFAM, Abf2p, Glom and Mh1bp (HMG-box proteins from *Homo sapiens*, *Saccharomyces cerevisiae, Physarum polycephalum* and *Yarrowia lypolitica* respectively) (**Figure R1B**). Gcf1p 25-245 was subcloned successfully in pCri6a and pCri7a, which introduce an N-terminal His-GST-TEV site or a C-terminal Histidine tag, respectively.

R2.2 Protein expression and purification

GST-Gcf1p 25-245 construct was expressed in BL21 and Rosetta2 following the protocol described in section **M3.2**. Best expression levels where achieved when cultures were induced for 4 hours (at an O.D.$_{600nm}$ of 0.6, 1mM IPTG, 24º C, 200 rpm), as a longer expression time resulted in protein loss while higher temperatures generated a Gcf1p C-terminal degradation product, as assessed by peptide mass fingerptinting (Centro de Investigaciones Biológicas CIB-CSIC, proteomics and genomics services) (**R2B**, Lane GST-Fraction 2).

GST-Gcf1p single point mutations to methionine were expressed in the same conditions and the same applied to deletion mutants, except for the GST-Gcf1p 63-245, which we were not able to express following our expression protocol for wild-type Gcf1p (see above). Gcf1p 25-245 construct neither showed expression when cloned into pCri7a plasmid, which adds a C-terminal His-tag. The pCri6a-Gcf1p 25-245 construct was not tried, provided that the pGEXT-Gcf1p 25-245 construct showed a production yield enough to undergo crystallization trials.

Our standard protocol for the purification of Gcf1p constructs consisted in a two-step chromatography (see **Figure R2**) that included a GST affinity column followed by in-column digestion with TEV protease, and elution of Gcf1p using the appropriate wash buffer. The major eluting protein after GST purification ran as a 30 kDa protein in 15% SDS-PAGE (see **Figure R2B**) and was identified to be Gcf1p by peptide mass fingerprinting (Centro de Investigaciones Biológicas CIB-CSIC, proteomics and genomics services). Only some minor contaminants were present, indicating that the digested GST affinity-tag and the undigested chimeric protein were efficiently retained in the stationary phase. The Gcf1p-containing samples were pooled, diluted to reduce the salt content and protein concentration and loaded to the following cationic exchange chromatography, which efficiently removed the minor contaminants from the sample. Gcf1p eluted as a major peak at 650 mM NaCl (**Figure R2 C**,**D**).

As it has been explained in Materials and Methods, we applied different purification workflows depending on the final purpose of the sample (see section **M4.2**). The specific yields for each procedure are summarized in **Table R1**. A third purification step consisting either in a heparin column or a gel filtration did not cause any improvement of crystal diffraction quality. Therefore, protein used in crystal formation was produced following the two-step protocol indicated above.

Table R1: ***Yields of each purification step starting from 2L of induced culture (A(mg) indicating total amount in mg and [C] (mg/mL) the concentration of the most concentrated sample in mg/mL). Numbers in brackets [ ] represent the values for the selenomethionine derivative for the constructs in which it was produced. Whenever either gel filtration or heparin were used, it was always after cationic exchange in two alternative three-step protocols. (N.D. stands for not done)***

	GST Affinity		Cationic Exchange		Gel filtration		Heparin	
	A (mg)	[C](mg/mL)	A (mg)	[C](mg/mL)	A (mg)	[C](mg/mL)	A (mg)	[C](mg/mL)
25-245	12 [10]	2 [1.5]	5 [3]	3.5 [2.7]	4 [2.5]	1.5 [0.8]	3.5 [2]	3.3 [2.8]
35-245	8	1.2	1.2	0.9	N.D.	N.D.	N.D.	N.D.
53-245	6 [4]	1.0 [0.8]	1.2 [0.8]	0.8 [1.4]	1 [0.6]	0.3 [0.1]	N.D.	N.D.
L224M	12 [10]	2 [1.5]	5 [3]	3.5 [2.7]	N.D.	N.D.	3.5 [2]	3.3 [2.8]
A170M	13	2	7	5	N.D.	N.D.	N.D.	N.D.
I59MI62MA170M	13	2	7	5	N.D.	N.D.	N.D.	N.D.
I45MI59MI62M	13	2	7	5	N.D.	N.D.	N.D.	N.D.

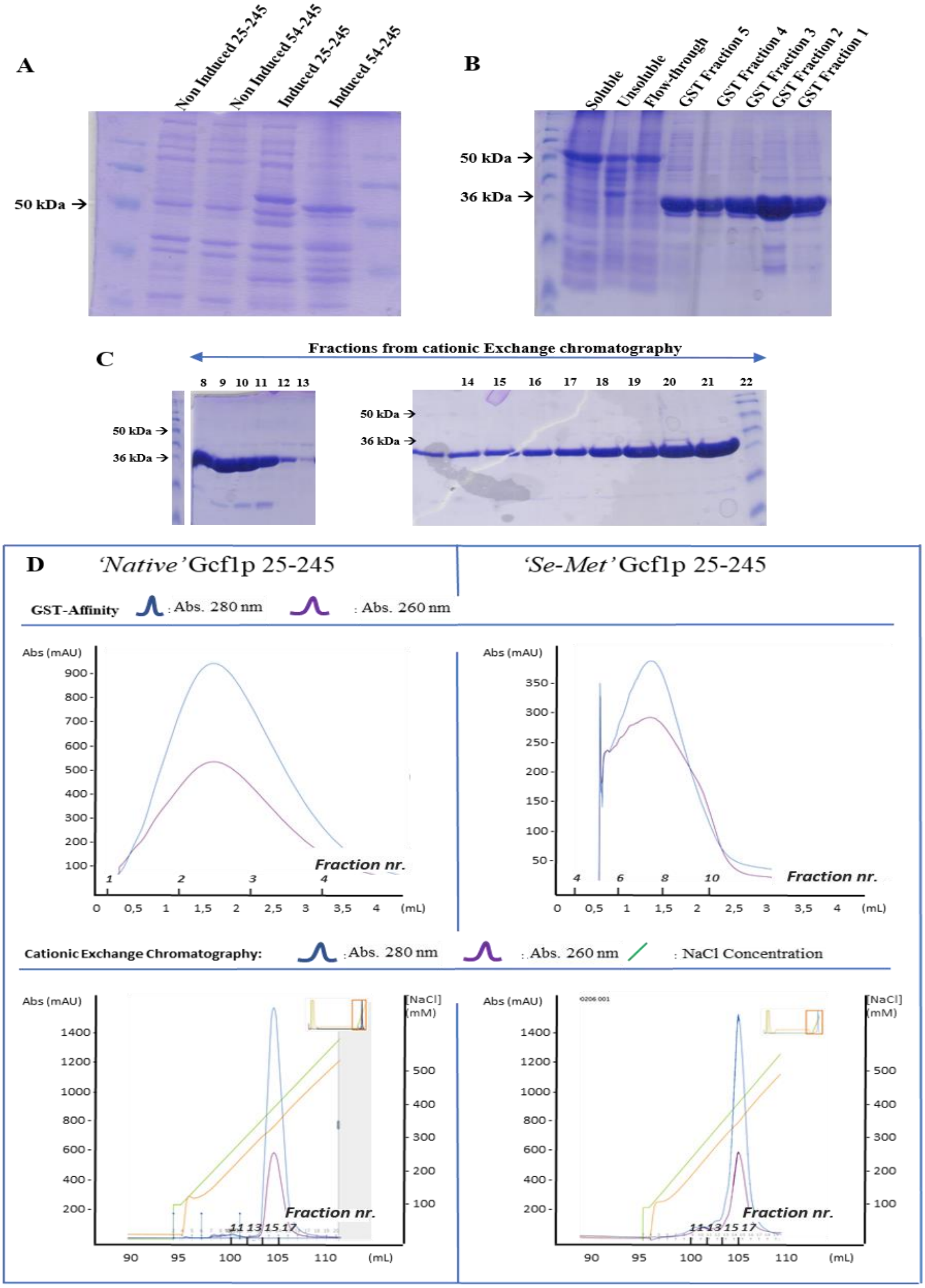


__Figure R2, Gcf1p production:__ SDS-PAGE of Gcf1p expression (__A__), GST affinity (__B__), and exchange chromatography purification (__C__). Chromatograms corresponding to both the native and 'SeMet' purifications (__D__).

R3 X-RAY STRUCTURE OF GCF1P IN COMPLEX WITH DNA

R3.1 Analysis of candidate DNA substrates for co-crystallization

R3.1.1 Gcf1p shows a dual behaviour in EMSA depending on DNA substrate

As for both TFAM and Abf2p, which were both previously crystallized in complex with DNA in the laboratory [42], [39], Gcf1p shows a high isoelectric point (pI=9.83) characteristic of DNA-binding proteins, whose stability requires a high salt content in the buffer solution. On the other hand, long regions in Gcf1p were predicted flexible (see section **R0** and **Figure R1**). Both TFAM and Abf2p showed a more compacted an rigid particle upon DNA binding, and none crystallized in the absence of DNA [42], [39]. Therefore, analogously to TFAM and Abf2p, co-crystallization of Gcf1p with an appropriate DNA was chosen as the strategy to obtain an atomic model of the protein [41], [42], [39], [40].

Electrophoretic Mobility Shift Assays were performed using DNA substrates of variated structure and sequence (**Table M2**). In all cases Gcf1p was kept at a salt concentration of 750 mM NaCl and only after DNA addition NaCl concentration was decreased at 100 mM NaCl. Serial dilutions of the protein sample were performed in order to scan different protein:DNA ratios. As shown in **Figure R3**, depending on the length of the DNA, different types of shifts were observed. Short linear DNAs of 25 or 35 bp did not show any well-defined band shift but a smear that collapsed to aggregates at highest protein concentrations. Instead, DNAs of 50 bp showed a band shift that eventually smeared, whereas longer DNAs (60 bp) showed the formation of a clear second shift before smearing. Branched DNA substrates recreating Holliday junctions showed formation of a double band that did not smear at the protein concentrations tested.

The smears of short DNAs in the EMSAs suggested that the procedure to prepare the samples could have affected the interaction of Gcf1p, despite the prevention of never decreasing the salt content in the absence of the DNA during the serial dilutions. Thus, we prepared the samples by a three-step dialysis that included an O/N step (see section **M6**). Substrate of choice was the dsDNA Af2_22 (5'-AATAAT**AAATT**ATATAATATAA-3') whose rigid poly-adenine tract facilitated Abf2p crystallization [42]. In order to screen the effect of different DNA lengths, Af2_22 sequence length was modified in steps of 2 bp thus generating the new DNA substrates Af2_18, Af2_20, Af2_24, Af2_26 and Af2_28 (**Table R2)**. As shown in Figure **R3**, Gcf1p shifted such linear DNA substrates in a single defined band.

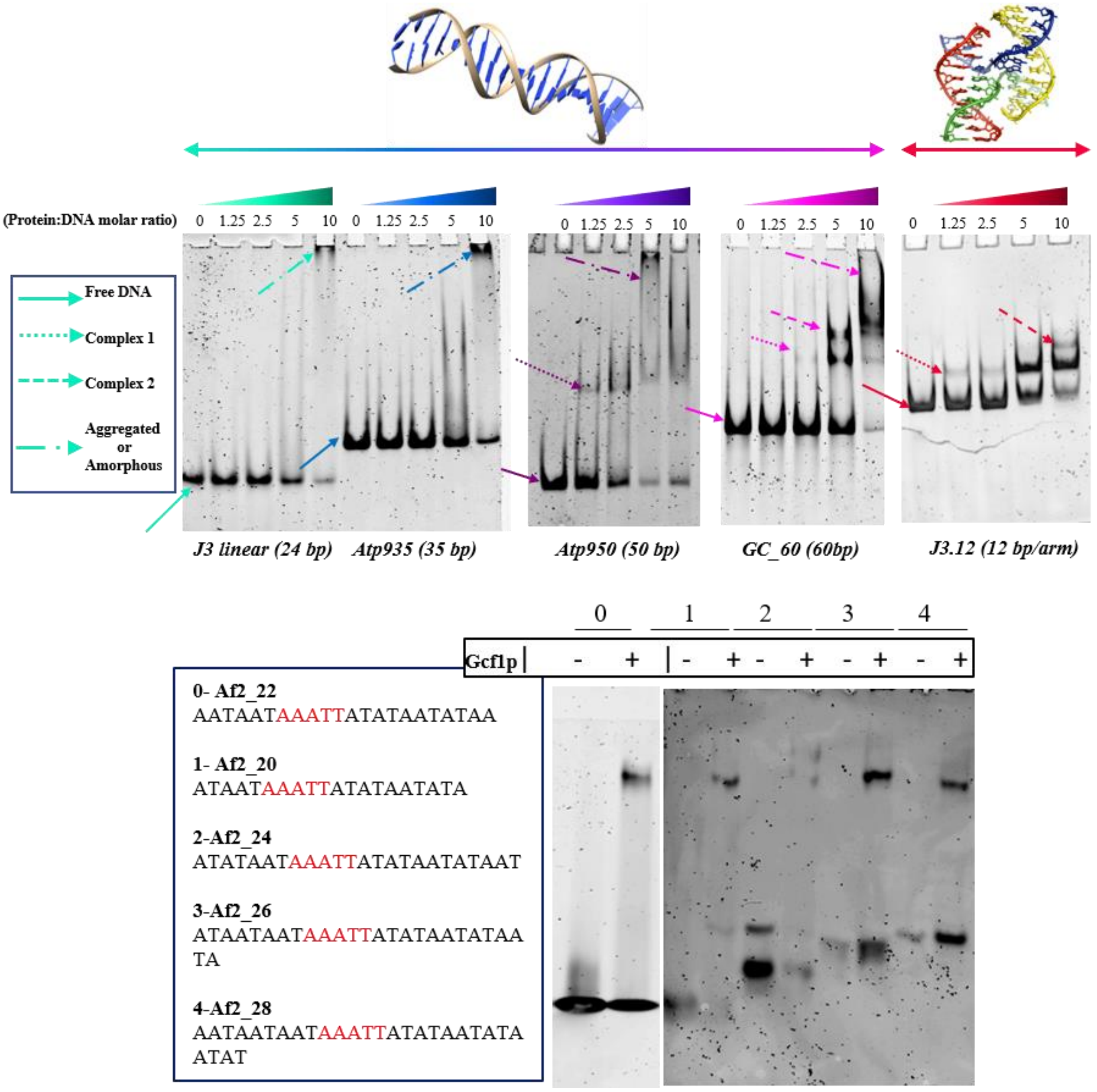


Figure R3: Binding of Gcf1p to different substrates. (A) *Binding of Gcf1p to different DNA types, analysed by EMSA, using linear and branched substrates recreating a Holliday as indicated on top of each sample. Sequences of each substrate are annotated in* ***Table M2****.* ***(B)*** *Samples containing complexes of Gcf1p/Af2 DNA variants stabilized by three step dyalisis, samples with and without protein are labelled as + and – respectively.*

R3.2- First crystallization attempts with different DNA substrates

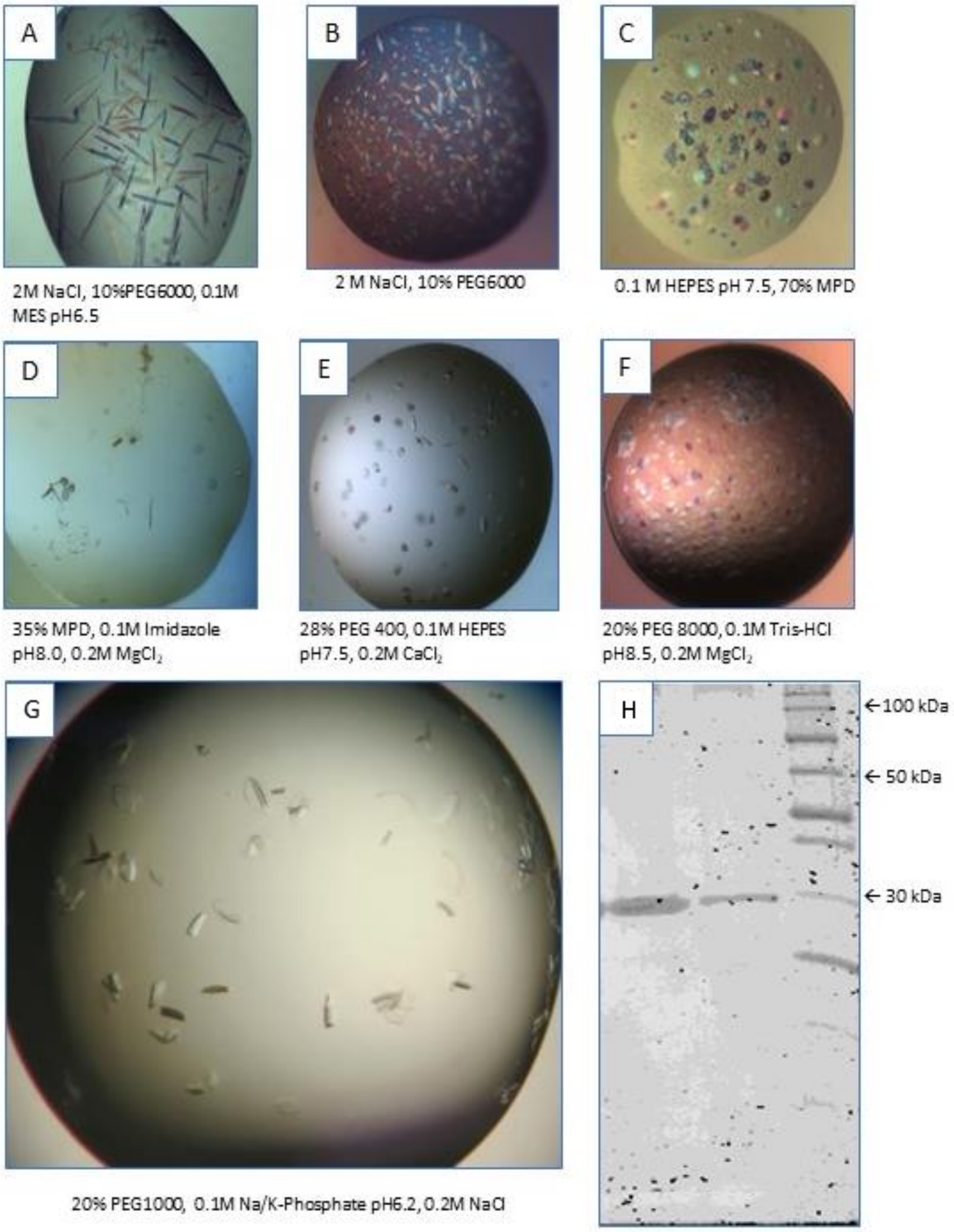
R3.2.1- DNA junctions and 40 bp DNA substrates

First DNA fragments selected as candidates for crystallisation of the Gcf1p/DNA complex were those showing a pattern of one or more clear shifted bands on native PAGE i.e. DNA junctions or DNA substrates longer than 40 bp. Following the EMSA results, the protein:DNA ratio for crystallization screenings was set at 1:1.2. The complexes were prepared following a three-step dialysis protocol (see above). Initial and extensive crystallization screenings plates were done by employing the nano-drop dispensing robots at the Automated Crystallography Platform (PAC) from the Barcelona Science Park.

First crystalline hits appeared for samples containing Gcf1p in complex with J3.12 junction (Figure **R4A**) and Atp950 dsDNA substrates (Figure **R4B** and **C**). in high salt concentration (2M NaCl, 10% PEG6000). The crystallization drops showed thin highly bi-refringent needles. Those crystals were optimized and tested in ESRF ID23-1 beamline. The crystals diffracted at 15 Å. In addition, strong phase separation was obtained for the linear DNA substrates Atp950 (50 bp), Atp940 (40 bp) and Atp935 (35 bp) in different chemical conditions (20% PEG 3350, 50 mM Tris-HCl pH 8.5, 10 mM $MgCl_2$), (0.2 M $(NH_4)_2SO_4$, 0.1 M Na Acetate pH 4.6, 30% PEG-MME 2000) and (28% PEG 400, 0.1 M HEPES pH 7.5, 0.2 M $CaCl_2$, respectively). These crystals were extremely fragile and none of them could be fished out from the mother liquor to flash-freeze them in order to test their diffraction (see **Table R2**).

R3.2.2- Af2_22 derived DNA substrates

In order to test shorter DNAs, the Af2_22 DNA (from the crystal structure of Abf2p [42]) and the corresponding different variants were also tested for crystallization. Complexes formed with all DNA substrates (Af2_18, Af2_20, Af2_22, Af2_24, Af2_26 and Af2_28) yielded small hexagonal protein-DNA crystals in magnesium containing conditions (**Figure R4D** to **R4F**), which showed clear edges and notable bi-refringence although none of them showed diffraction spots (see **Table R2**). Strikingly, Af2_20 resulted in an important breakthrough regarding the crystal diffraction quality. Af2_20 was the only one that yielded crystals in the F2 condition from PAC2 (20% PEG1000, 0.1M Na/K-Phosphate pH6.2, 0.2M NaCl) (see **Figure R4G**), the same condition that was used for the crystallization of TFAM in complex with LSP [39]. Indeed, the sample containing Gcf1p and Af2_20 yielded crystals that showed diffraction spots up to 6 Å resolution which could be indexed. The presence of protein inside these crystals was confirmed by SDS-PAGE (**Figure R4H**). Crystals were fished from the crystallization drop, thoroughly washed in three soaking steps of 10 seconds in mother liquor, smashed and loaded onto SDS-PAGE.



2M NaCl, 10%PEG6000, 0.1M MES pH6.5

2 M NaCl, 10% PEG6000

0.1 M HEPES pH 7.5, 70% MPD

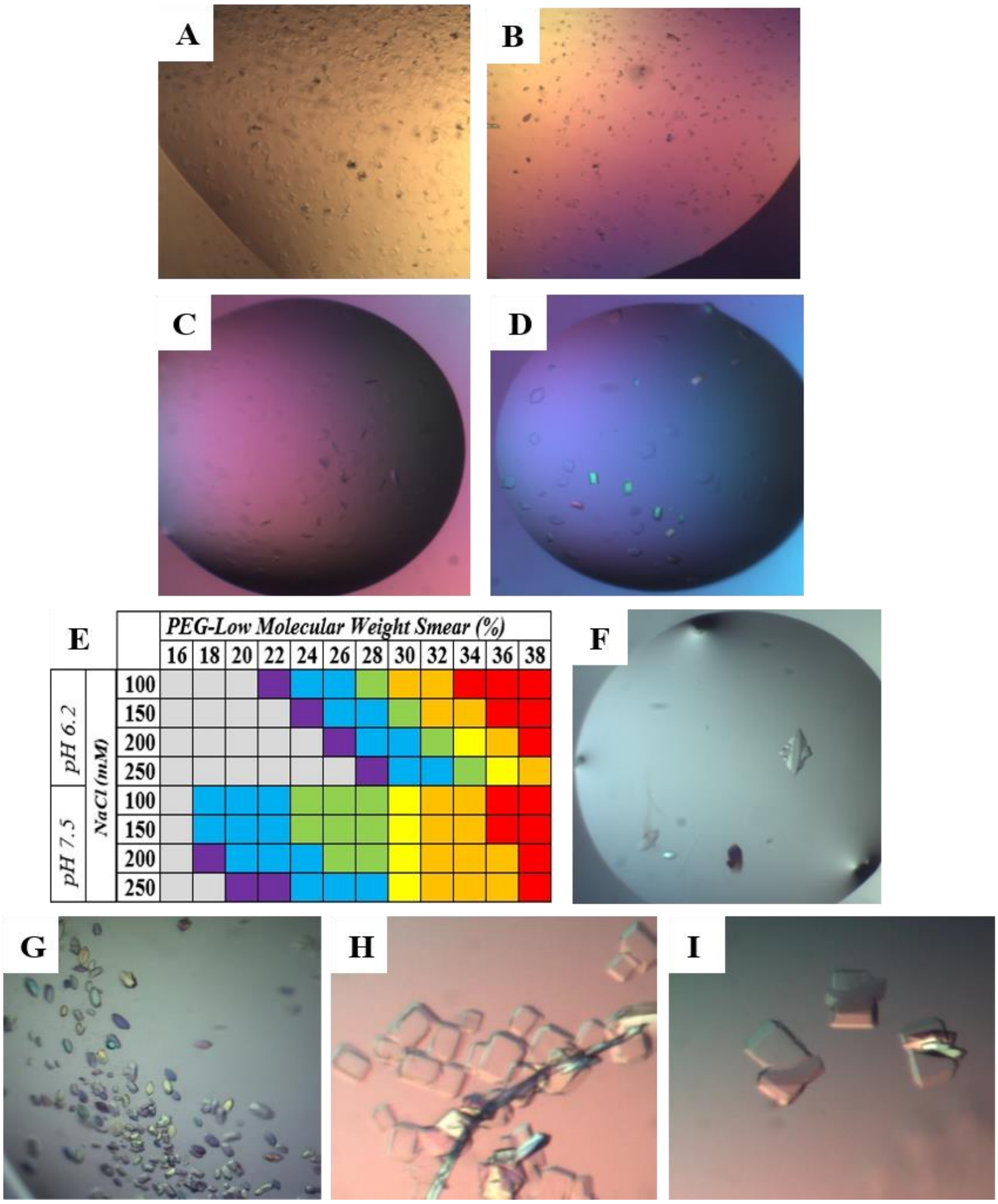
35% MPD, 0.1M Imidazole pH8.0, 0.2M MgCl$_2$

28% PEG 400, 0.1M HEPES pH7.5, 0.2M CaCl$_2$

20% PEG 8000, 0.1M Tris-HCl pH8.5, 0.2M MgCl$_2$

20% PEG1000, 0.1M Na/K-Phosphate pH6.2, 0.2M NaCl

Figure R4: *First steps in crystal optimization, from first crystalline hits to first diffracting crystals. Crystals appeared both when using junction-like DNA substrates* ***(A)*** *and 50bp linear DNA substrates* ***(B and C)*** *but they showed poor X-ray diffraction quality. DNA substrates of 18 to 28 bp profusely crystallized in magnesium-containing conditions* ***(D, E and F)****, but only crystals that appeared in the presence of 20 bp DNA showed promising X-ray diffraction* ***(G)****. Confirmation of the presence of Gcf1p in these crystals is shown in the SDS-PAGE in* ***(H)*** *1 contained protein standards, lane 2 contained a sample of 10 washed crystals, lane 3 contains 8µg of purified Gcf1p.*



			PEG-Low Molecular Weight Smear (%)											
			16	18	20	22	24	26	28	30	32	34	36	38
pH 6.2	NaCl (mM)	100	⬜	⬜	⬜	🟪	🟦	🟦	🟩	🟧	🟧	🟥	🟥	🟥
		150	⬜	⬜	⬜	⬜	🟪	🟦	🟦	🟩	🟧	🟧	🟥	🟥
		200	⬜	⬜	⬜	⬜	⬜	🟪	🟦	🟦	🟩	🟨	🟧	🟥
		250	⬜	⬜	⬜	⬜	⬜	⬜	🟪	🟦	🟦	🟩	🟨	🟧
pH 7.5		100	⬜	🟦	🟦	🟦	🟩	🟩	🟩	🟨	🟧	🟧	🟥	🟥
		150	⬜	🟦	🟦	🟦	🟩	🟩	🟩	🟨	🟧	🟧	🟥	🟥
		200	⬜	🟪	🟦	🟦	🟦	🟩	🟩	🟨	🟧	🟧	🟧	🟥
		250	⬜	⬜	🟪	🟪	🟦	🟦	🟦	🟨	🟧	🟧	🟧	🟥

Figure *R5* Optimization of the Gcf1p-Af2_20 crystals. *Crystals grown at a protein:DNA ratio 1:1.2 with DNA sequences Af2_20* ***(A)*** *or Af2_20 Overhang-1* ***(B)****, are shown. In* ***(C)*** *and* ***(D),*** *crystals grown at respective 1:1.2 and 2:1.2 protein:DNA ratios, in both cases with Af2_20 DNAs, are shown.* ***(E)*** *Optimization screening that rendered good quality diffracting crystals. Shells in grey indicate no crystals; in violet, three-dimensional single crystals from which best datasets were obtained; in blue three-dimensional multiple crystal; in yellow, multiple small crystals; in orange nuclei; and in red, protein precipitate. PEG low-molecular weight smear is a screening based on a mixture of low MW PEG. From green to red shells, crystals appeared as shown in* ***(F)*** *native crystals;* ***(G)*** *'SeMet' crystals in sitting drop set up;* ***(H)*** *and* ***(I)*** *'SeMet' crystals, in hanging drop setup.*

R3.3- Optimization of Gcf1p/Af2_20 crystals

R3.3.1- Improvement of the diffraction limit from 6 to 4 Å resolution

First crystals were scaled-up by using 24 well plates with 2 µL of initial drop volume (1 µL of Gcf1p/Af2_20 complex + 1 µL of reservoir). First optimization series consisted in small increases of the precipitant concentration in 2% steps and resulting crystals were typically growing as plates attached to the plastic well bottom like those shown in **Figure R4G** and **Figure R5A**, **B** and **C**.

The second round of crystal optimization was dedicated to generating three-dimensional single crystals. Different protein: DNA ratios were assayed and a ratio of 2 proteins per 1 DNA duplex was identified as the one that yielded crystals with best defined edges and with three-dimensional shape (see **FigureR5D**). At this point, systematic screening around the lead condition through small changes in precipitant concentration and different pH buffering systems (see section **M9.2**, **Figure M12**). Such an optimization yielded bigger crystals diffracting at 4 Å (see **Figure R5F**). After these encouraging results, further crystallizations were done with Gcf1p:DNA ratio of 2:1.2 in all cases. In parallel to the protein: DNA ratio optimization, DNA substrates Af2_19, Af2_21, Af2_20NoTract, Af2_20Shift1 and Af2_20Shift2 were also assayed (see **Table M2**). Only substrates Af2_20, Af2_20Shift1 and Af2_20Shift2 (with the adenine tract shifted along the sequence), yielded better X-ray diffracting crystals, and amongst them, crystals containing Af2_20 were the only ones that diffracted with a detectable signal beyond 3.5Å.

R3.3.2- Additives and cryocooling optimization pushed X-ray diffraction data resolution limits beyond 3 Å

In the next rounds of optimization, different crystallization additives from Additive Screen HT (Hampton Research ®) were tried. Ionic compounds such as $FeCl_3$ and $CoCl_2$ typically reduced the nucleation, thus yielded fewer yet bigger crystals that resulted in a 3.2 Å diffraction. Polar, non-charged organic molecules, specifically glycerol and D-trehalose also reduced nucleation crystals were smaller than with the ionic additives. More interestingly, crystals grown in the presence of organic additives had better defined edges and yielded usable diffraction signal up to 3.0 Å when spherical truncation was applied (as implemented in XDS_CORRECT). It is worth reporting that at this point a change on the cryocooling process of the crystals was key in pushing the resolution limit. Addition of the cryoprotectant sequentially to the crystals contained in the crystallization drop instead of transferring them into solutions of increasing cryoprotectant concentration pushed the resolution beyond 4 Å, reflecting a delicate internal crystal structure susceptible to brusque changes in the osmotic pressure. In all cases, the cryoprotectant was the precipitant (PEG Low Molecular Weight Smear) at higher concentrations, yet a concentration higher than 26% resulted in loss of diffraction quality. Best diffracting crystals showed moderate anisotropy with an overall resolution of 2.90 Å assessed by Aimless by means of the $CC_{1/2}$ indicator during data processing [135].

R3.4- Crystallization and data collection of SeMet crystals

Crystals of Se-Met Gcf1p grew similarly to native Gcf1p. We started screening crystallization conditions around the condition that yielded first Gcf1p native crystals (see **Figure R5 G-I)** and, after several optimization rounds, the crystals reached a diffraction limit comparable to those of the native protein (**Figure R6A** and **B**). Best diffracting crystals appeared in 20% PEG-LMWSmear, 0.1M Na/K-Phosphate pH 6.2, 0.2 M NaCl and 3% Glycerol, in hanging drop at 8mg/mL of SeMet-Gcf1p/DNA complex. Data collection was performed at the energy experimentally determined as the peak of the X-ray fluorescence spectra (12669.7 eV) (Figure **6C**). Prior to data collection, the omega spindle axis was aligned to one axis of the reciprocal cell. By this strategy, Friedel pairs were collected within the same frame minimizing the effect of radiation damage and maximizing the anomalous signal in the dataset. Best diffracting Se-Met crystals showed a slight anisotropy (see data processing below) (**Figure 6 D-F**). In order to enhance the anomalous differences, we designed mutations to introduce additional methionine residues (**Figure R1**). We chose to mutate those apolar residues of Gcf1pthat were aligned with a methionine in one of the mutants. Mutants **I70M**, **I84M**, **I87M**, **A170M** and **L224M** (as well as the possible combination between them), were all produced in good amounts, but only crystals of **L224M** could be grown. Notably, mutation **A170M** was found to dramatically disrupt crystal formation without affecting DNA binding (none of the following mutants: **A170M**, **I70M/I84M/A170M, I70M/I84M/A170M/L224M, A170M/L224M** could be crystallized). In order to increase the SeMet: Number of residues ratio, the shorter 53-245 Gcf1p variant was also crystallized (1 Met per 99 residues). Both L224M Gcf1p (full-length) and 53-245 variant SeMet derivatives yielded crystals that diffracted much weaker than the wild-type SeMet derivative protein and no usable datasets could be collected. A mutant that combined the 53-254 Gcf1p variant and the **L224M** substitution was also crystallized but the X-ray diffraction was even poorer than the other mutants, thus this approach was dismissed.

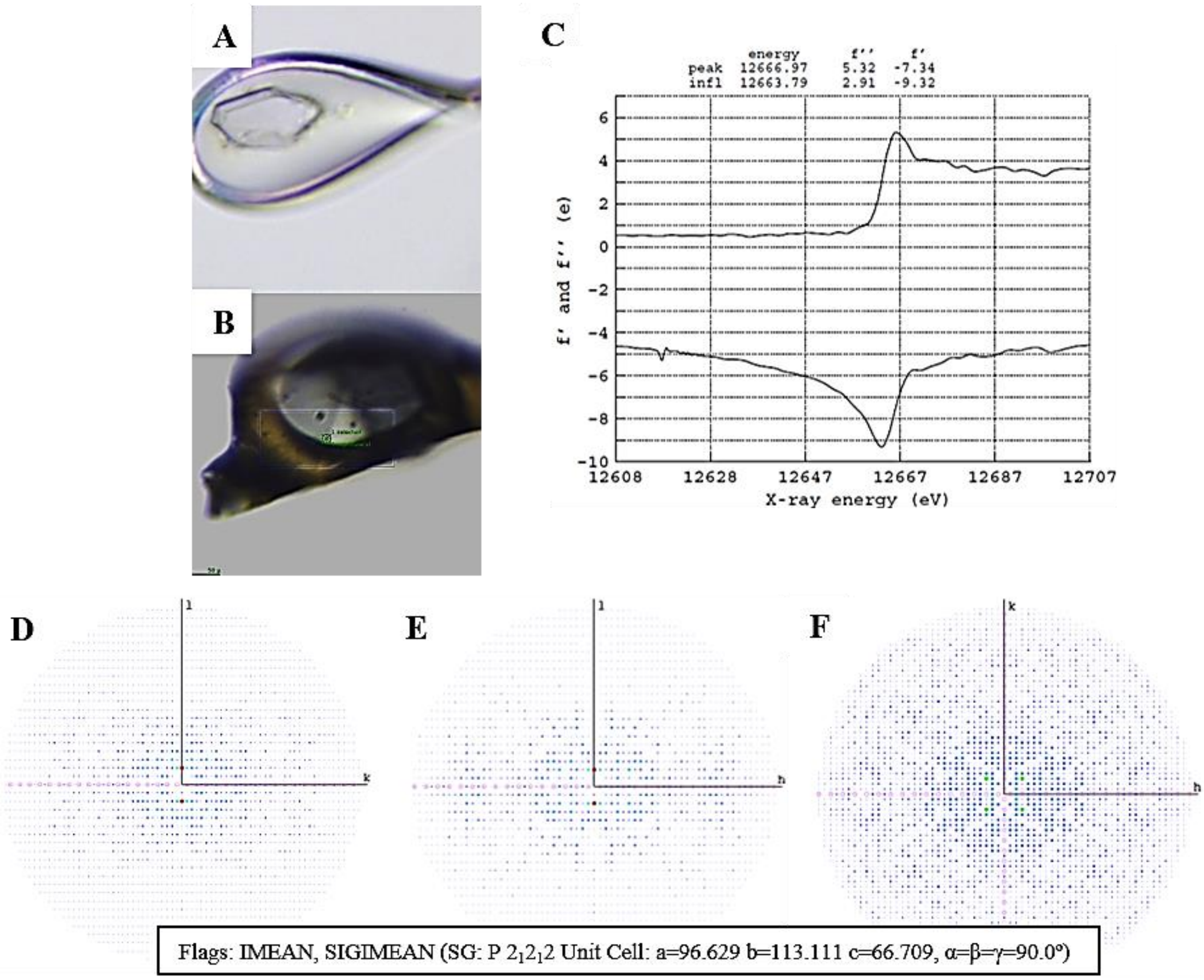


Figure R6: *Snapshots of good quality diffraction native* ***(A)*** *and Se-Met* ***(B)*** *crystals mounted in a loop and loaded to the goniometer head at the Xaloc beamline in ALBA synchrotron.* **(C)** *Plot of f' and f'' as a function of X-ray energy, both determined by an X-ray fluorescence scan. The experimentally measured f'' peak (at 12666.97 eV) at which the datasets for SAD-phasing were collected is shown on top. Panels* ***D*** *to* ***F*** *show central 2D sections of the Se-Met data, at the reciprocal space in the planes $\overline{kl}$, $\overline{hl}$ and $\overline{hk}$ respectively. Reflections show direction-dependent intensities (i.e. anisotropy) in the l axis. The dataset, indexed in the $P2_12_12$ space group, had systematic absences (expected positions of the systematic absent reflections are highlighted as pink circles) in both the h and k reciprocal axes.*

R3.5 Data processing and structure solution

R3.5.1 Preliminary data processing

Data indexing, integration and scaling was performed with the XDS package. For the best diffracting native crystal, the unit cell parameters were determined as 96.63 Å, 113.11 Å and 66.71 Å for the a, b, c axes and 90º for the α, β, γ angles, thus it belonged to the orthorhombic crystal class and Bravais lattice *oP*. Search for systematic absences with Pointless unambiguously determined presence of a screw symmetry operator in reciprocal axes **h** and **k**. For axis **l**, a marginal statistical significance for the presence of a screw component was determined, therefore prior to structure solution we kept as hypotheses both space groups **19** ($P2_12_12_1$) and **18** ($P2_12_12$).

Datasets from single crystals typically showed resolution limits of 2.9 Å for the reciprocal space axis **h**, 3.0 Å for **k** and 4 Å for **l** with an overall resolution of 2.95 Å as assessed by the $CC_{1/2}$ indicator [135]. The criteria to determine the definitive resolution limit was a balance between $CC_{1/2}$, Rmeas, completeness, multiplicity, and $I/\sigma(I)$.

An important aspect when solving a crystal structure is the quality and completeness of the diffraction data. A remarkably successful strategy to increase both aspects is the merging of independent X-ray data sets. The success of such an approach strongly relies in the isomorphicity of the diffracted regions, in our case we selected datasets with a Linear Cell Variation (LCV) within 1%. Analysis of different combinations of datasets was tested with program Blend [136] from the CCP4 suite. Four datasets from different crystals were scaled to the best dataset and combined. Merging statistics are listed in **Table R3** in which overall resolution limit of the unmerged data corresponded to that of the best diffracting axis. Statistics in **Table R3** correspond to not anisotropically corrected data, thus a systematic error on the measurement of the noise was introduced due to inherent direction-dependency on the intensities, which yielded a rather high value for the Wilson B-factor. In addition, the unmerged indicator R_{meas} was greatly improved when datasets from different crystals were combined.

Table R2: *Summary of the statistics from different unmerged data sets used for structure solution. First column corresponds to the dataset from the best diffracting native crystal, second column corresponds to the combination of 4 native datasets obtained from four different isomorphous crystals, and the third column corresponds to the combination of 4 datasets obtained from the same SeMet crystal. In parenthesis, values for the corresponding last shell.*

Description	**Native Dataset**	**4 Native Datasets**	**4 'SeMet' Datasets**
Wavelength (Å)	0.979312	0.979312	0.978960
Space group	$P2_12_12$	$P2_12_12$	$P2_12_12$
Cell axes (a,b,c) (Å)	96.900, 113.060, 66.620	96.900, 113.060, 66.620	97.460, 113.290, 66.830
Cell angles (α,β,γ) (º)	α = β = γ = 90.000º	α = β = γ = 90.000º	α = β = γ = 90.000º
Wilson B-factor (Å^2)	81.56	71.18	101.7
Resolution limit h (Å)	2.9	2.75	2.97
Resolution limit k (Å)	3.01	3.07	3.15
Resolution limit l (Å)	4.24	3.46	3.49
Resolution Range	49.39-2.9 (3.08-2.9)	73.57-2.75 (2.88-2.75)	49.56-2.97 (3.15-2.97)
Total Reflections	107837 (16505)	613961 (75128)	701786 (47702)
Unique Reflections	16803 (2653)	19644 (2571)	15680 (2313)
Multiplicity	6.4 (6.2)	31.3 (29.2)	44.8 (20.6)
Completeness (%)	99.8 (99.6)	99.9 (100.0)	98.7 (92.0)
Mean I/sigma(I)	7.9 (0.5)	7.1 (0.1)	13.6 (0.4)
Half-set correlation CC$_{1/2}$	0.997 (0.280)	0.998 (0.101)	0.883 (0.330)
Rmeas	0.174 (6.723)	0.455 (85.901)	0.246 (9.435)
Anomalous Completeness	99.8 (99.4)	100 (100)	98 (88.9)
Anomalous multiplicity	3.3 (3.1)	16.4 (14.9)	23.4 (11)
Anomalous Correlation	-0.097 (-0.031)	-0.182 (0.001)	0.528 (0.02)

Processing of diffraction data from SeMet crystals was done analogously to the native data. Several datasets were collected centred in each of the unit cell axes and thus best statistics arose when combining four datasets from the same crystal . Datasets were scaled and combined. Only reflections belonging to the same Bijvoet groups were merged with Aimless, so that the anomalous differences were kept. The final dataset showed crystal cell parameters isomorphous to the native protein (**Table R2**) and the same space group ambiguity.

R3.5.2 Determination of initial selenium atom positions

Search for selenium atoms was performed in all the possible space groups within the P222 Bravais lattice, yielding a statistically significant solution for space group **18** ($P2_12_12$).Four sites with occupancy higher than 75% were found with a significant drop to an occupancy below 30% for the fifth site (**Figure R8A**). This result indicated that a probable solution for the Se atoms positions was found so that the true space group was indeed $P2_12_12$. Moreover, considering that the primary structure of Gcf1p contains two methionine residues, these results indicated the presence of 2 Gcf1p molecules in the asymmetric unit, thus consistent with the Matthews coefficient. In a further stage, thorough analysis of the sites reflected the presence of a binary non-crystallographic symmetry axis along the asymmetric unit (**Figure R8B**). Despite the encouraging results, at that point we failed in obtaining interpretable maps. The correct Se sites, confirmed later by model building, are depicted in **figure R7**. Search for the selenium atom positions was performed at 5 *Ȧ* as it was the upper resolution limit for anomalous signal using the criteria $I_{ANO}/\sigma_{ANO}>1.3$.

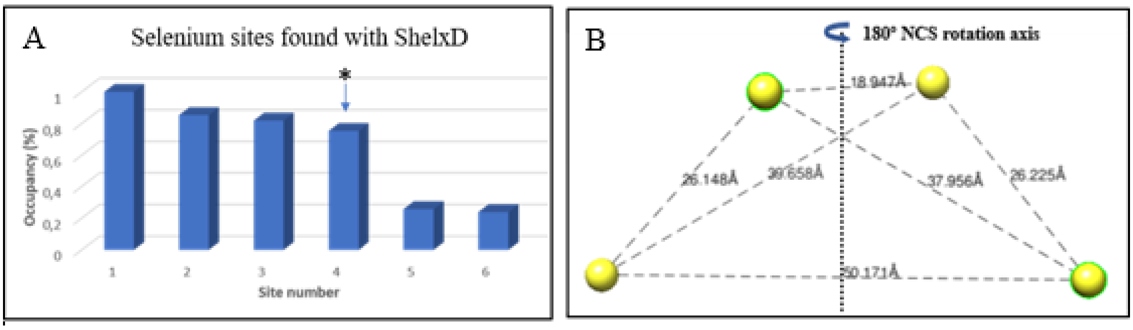


Figure R7. Determination of the Se substructure. *The occupancy of the different Se sites found for space group 18 is shown* ***(A)****. Asterisk indicates the last of the trusted positions based on the occupancy value. In* ***B****, the four selenium atomic positions with highest occupancy are represented as yellow balls. The relative distances between them indicate the presence of a two-fold NCS axis, as represented. This observation combined with the fact that only 2 methionine residues are present in the primary structure of Gcf1p suggested that two Gcf1p molecules were present in the asymmetric unit.*

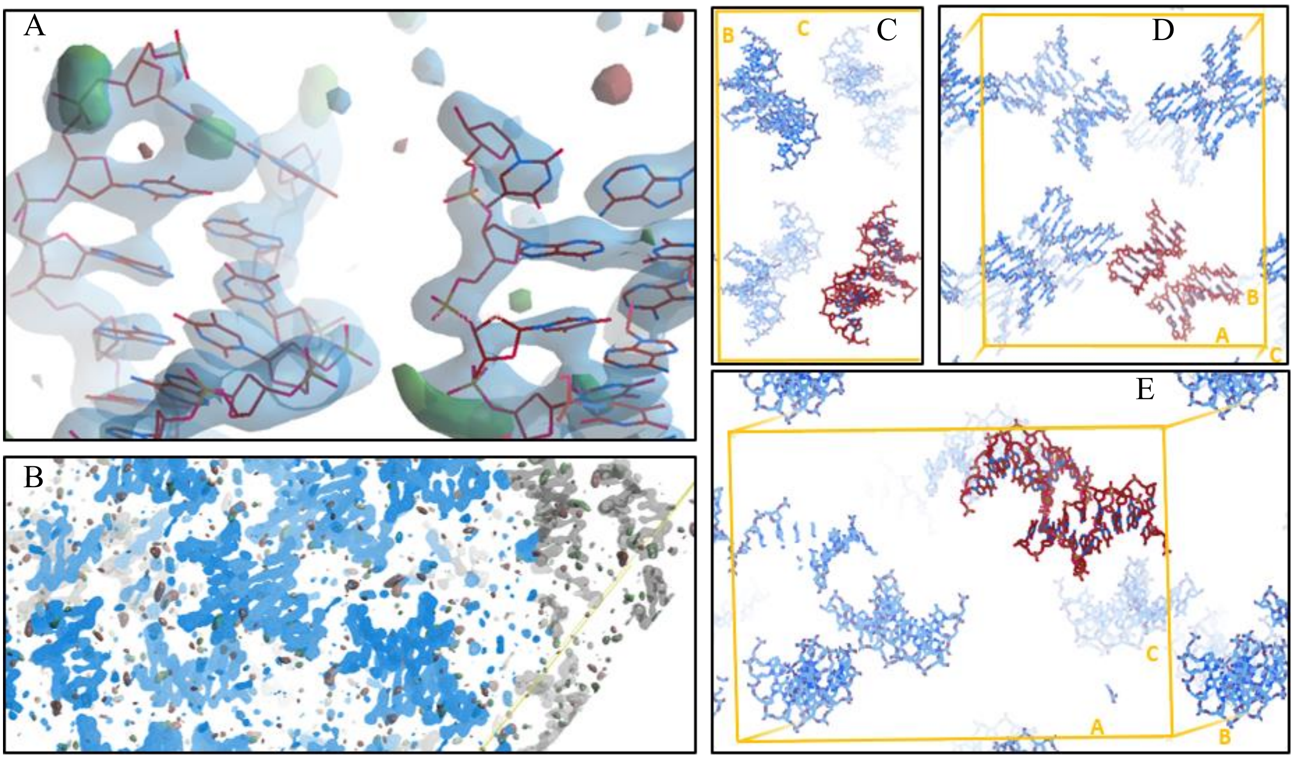
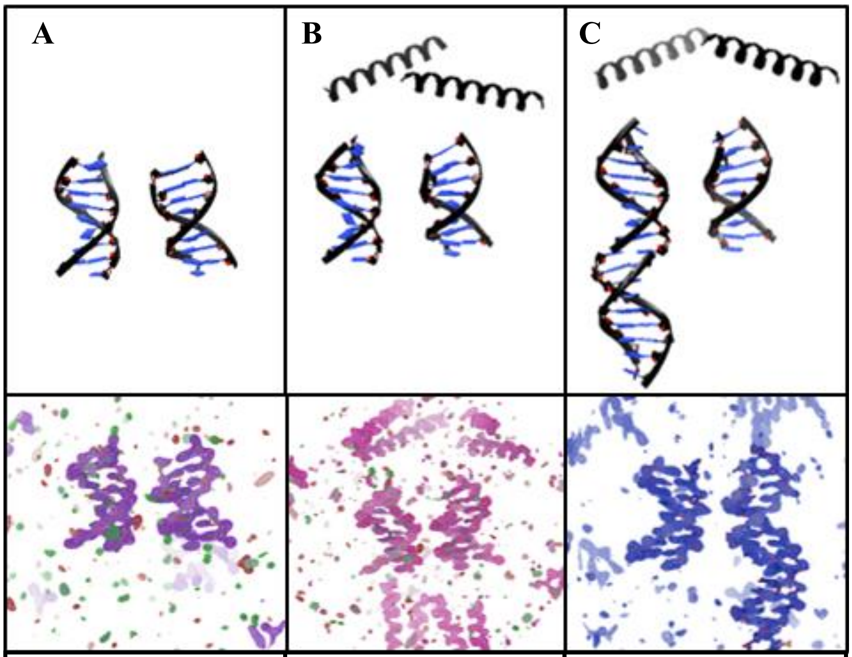


***Figure R8. Partial MR solution: (A)**, partial solution found with Phaser_MR with two DNA fragments of 8bp. The 2Fc-Fo Fourier synthesis is represented in blue, and positive and negative values for the Fc-Fo synthesis in green and red respectively. **(B)** General view of maps in **(A)**, which revealed that the density is discontinuous and only relevant where the MR search model was placed, thus making impossible to trace the rest of the molecule. **(C** to **E)** Packing of the partial solution is represented in the three different orientations, the asymmetric unit content is depicted in red, the symmetry copies in blue and the unit cell axes in yellow.*

R3.5.3 Resolution of gcf1p/DNA structure by molecular replacement

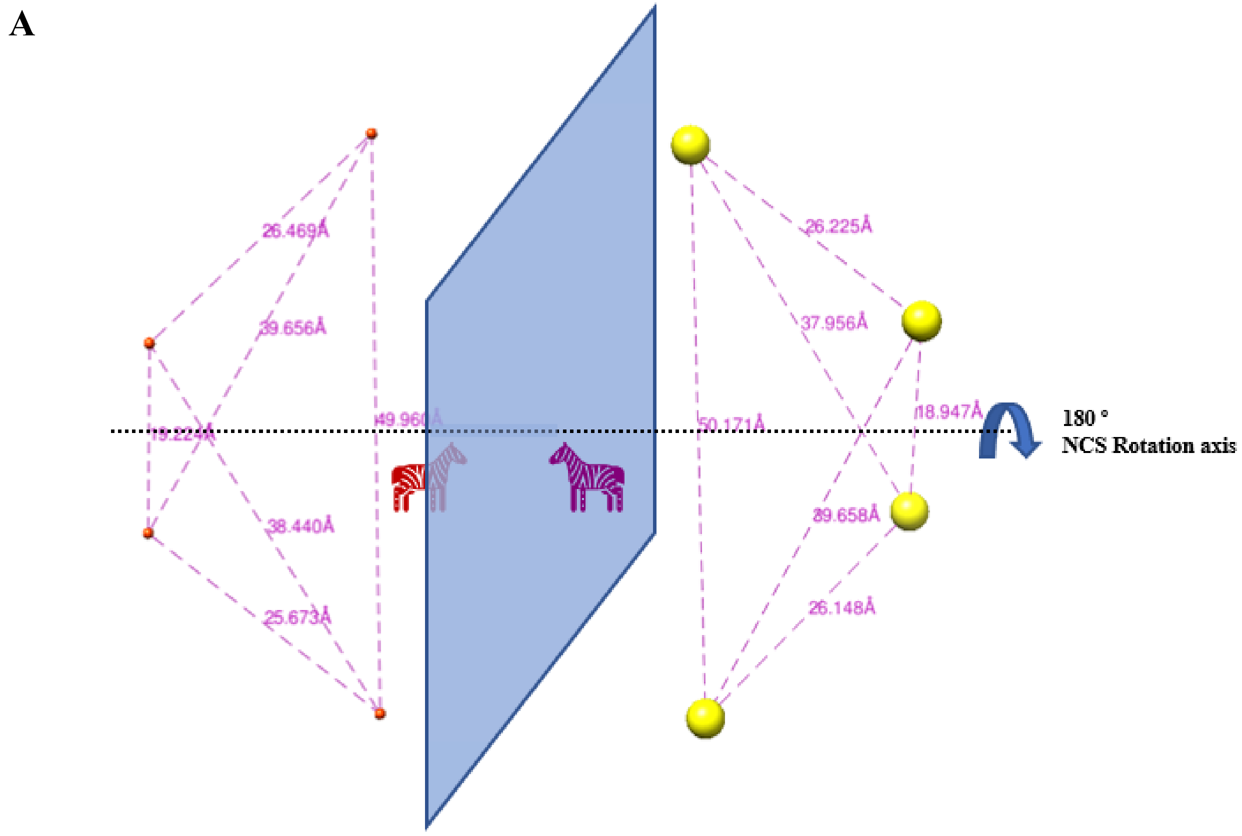
In parallel to the search of heavy atom sites, molecular replacement (MR) searches were performed in order to place both DNA and protein fragments within the asymmetric unit. MR searches were done with the four merged native dataset from different crystals at full resolution (2.18 Å at the corner of the detector at the synchrotron beamline (**Table R2**). Phaser-MR corrected the anisotropy by automatically discarding the reflections beyond 2.55 Å and keeping weak reflections in the resolution range 2.55-3.10 Å. Other approaches, like the one implemented by Staraniso [137] to correct the anisotropy by discarding such reflections, would have hampered phasing. First attempts to solve the structure by MR using the HMG-box domains of either TFAM (PDB_ID:3TMM) [39] or Abf2p (PDB_ID:5GH0) [42] as searching models, placed the models with poor scores and numerous clashes.

The initial search with DNA fragments included both ideal B-DNAs and different fragments of the Af2_22 DNA from the Abf2p/DNA structure (PDB_ID:5GH0) [42] as we used this DNA for crystallization. Two 8 bp DNA fragments, between positions 6 and 13 from the Af2_22 DNA structure (PDB_ID:5GH0_B and 5GH0_C), yielded an acceptable result (see **Figures R8, R9A**) with metrics TFZ=10.3 and LLG=115.234, which was two points higher in TFZ than 2 additional possible solutions. Importantly, this solution appeared for space group P2_12_12, in agreement with the best solution for the Se sites. Further MR searches done by fixing the DNA partial solution placed two ideal poly-Ala α-helices of 20 and 25 residues (**Figure R9B)** and a third 8 bp DNA fragment (**Figure R9C**). Placement of the different fragments over the first MR solution showed significantly good metrics and reasonable positioning of the structures.



TFZ = 10.5 LLG= 115	TFZ = 7.8 LLG= 255	TFZ = 9.3, LLG= 238
3 possible solutions	**5 possible solutions**	**1 possible solution**

Figure R9, Improvement of the partial model by successive MR searches: *First partial MR solution (A) followed by the second placement of two α-helices (B) and finally a third DNA fragment (C). Resulting maps and Phaser statistics are indicated below the images.*



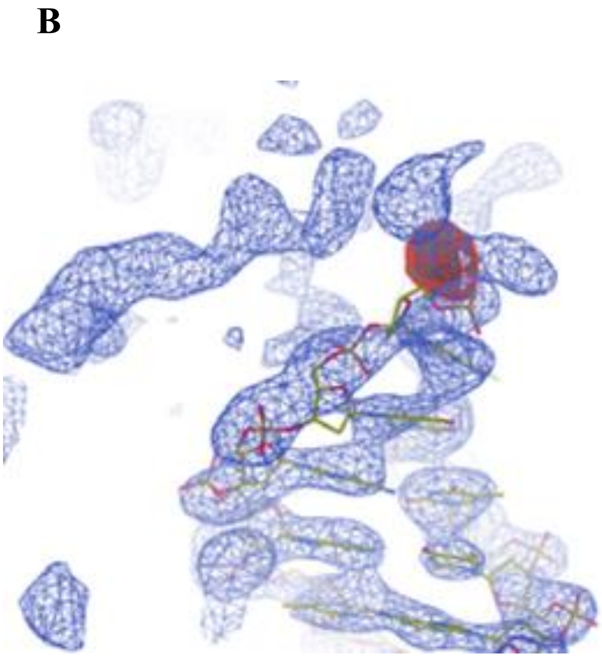


Figure R10. Structure solution of Gcf1p/Af2_20 structure. *(A) Representation of the selenium positions found by SAD with ShelxD (big, yellow spheres) vs the positions found by MR-SAD using Phaser-EP with the most complete partial MR solution (**Figure R9C**). The actual relation between solutions was stablished undoubtedly with phenix.emma [138]. (**B**) $2mF_0$-F_c Fourier synthesis from MR-SAD is shown in blue. The map around the DNA structure shows α-helix features in regions with no model, and close to a selenium atom whose red density corresponds to the anomalous difference map at 5.0 rmsd. The strong anomalous peak was very important for protein sequence assignment.*

R3.5.4 Phasing, phase extension and model building

The isomorphicity between the native and SeMet datasets (see **Table R2**) made possible to phase the SeMet data by MR-SAD using the partial model found in the native crystal (by Phaser-MR). Once the partial model placed within the SeMet derivative dataset, Phaser-EP searched for Se-sites and found four selenium atom positions with the same relative distances and symmetry relationships than those previously found by SAD with SHELXD. The positions found with Phaser-EP were indeed the mirror image of those preliminarily selected using SHELX (see **Figure R8**). By Euclidean model matching (phenix.emma [138]), taking into account space group symmetry, possible shifted origins and hand ambiguity it was confirmed that indeed both Se positions solutions were actually the same (see **Figure R10**). The Fourier synthesis calculated with Phaser-EP was clearly richer in features than that of ShelxE or Phaser-MR (**Figure R10** and **Figure R9**). Hand ambiguity was automatically solved by imposing the handedness of the partial MR solution and a single round of density modification yielded clear α-helical features surrounding the Se sites, in regions of the electron density map without model. Further rounds of density modification using phenix.resolve including non-crystallographic symmetry (NCS) averaging were performed using the pseudo two-fold symmetry axis present within the asymmetric unit. Phase extension from 5 Å (the resolution limit of the significant anomalous signal and used for the Se position search) to 2.55 Å (the full resolution of the dataset) was performed in small steps of 0.1 Å. After several density modification cycles, structural features for both DNA and protein were clearly defined even far away from the selenium sites and the partial model as shown in **figure R11**.

Phenix-Autobuild was run at this point treating the DNA as ligand and rebuilding the protein using the phase information upon density modification. The native dataset was elliptically truncated by Staraniso [137] and used as input native experimental data for automated model building. The statistics for both the spherical and elliptically truncated native datasets are summarized in **Table R3**. Initial tracing of the protein by Autobuild was completed by iterative cycles alternating manual model building (with Coot) and automatic refinement (Phenix Refine) until an R-free of 28% and Rfactor of 24% were reached. During the manual building process, the strong anomalous peaks corresponding to selenium atoms (see **Figure R13**) were crucial to assign properly the sequence, whereas bulk aromatic rings also served as a guide. Representative snapshots of initial and end states of the manual model building are shown in **Figure R12**. Staraniso anisotropic correction [137], treats the dataset as an ellipsoid and discards weak reflections, yielding a low overall spherical completeness (66% along the dataset). As an alternative approach, structure refinement was performed against different single spherically truncated datasets. From those datasets, the one that gave best results was the native dataset used as a scaling reference for merging (crystal grown in the presence of 3% L-Trehalose as additive) and whose merging statistics are summarized in **Table R2**. In order to reduce the anisotropy-related noise, the dataset was truncated at 3.2 Å and its refinement statistics are also summarized in **Table R3**. Refinement with this dataset yield overall worse indicators (higher R-work and R-free, as well as, higher B) and worse-defined map albeit richer in features as compared to the Staraniso-truncates dataset, which allowed us to trace 7 extra residues at the N-terminus of one protein chain. Refinement statistics for the model against both datasets are summarized in **Figure R14**.

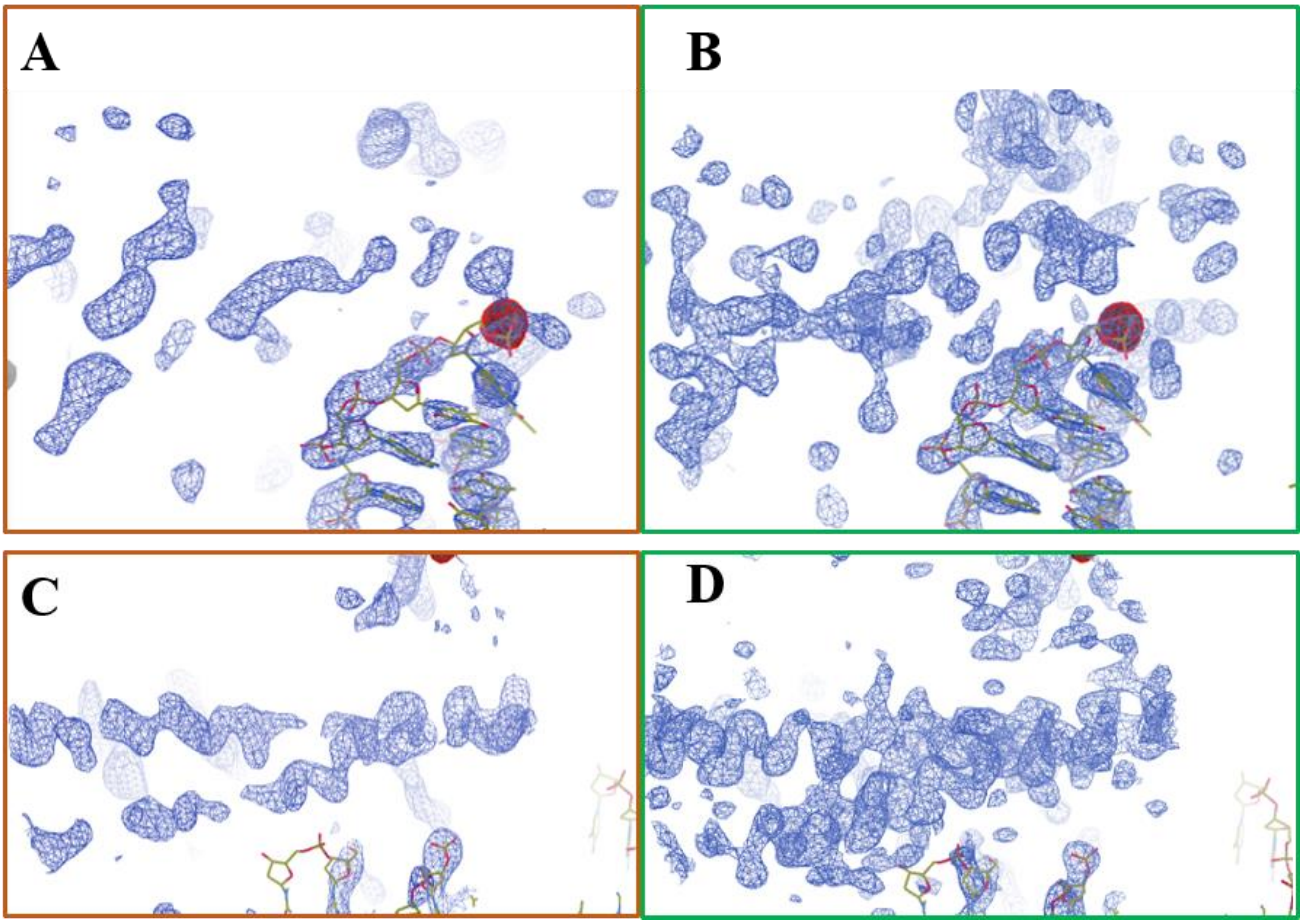
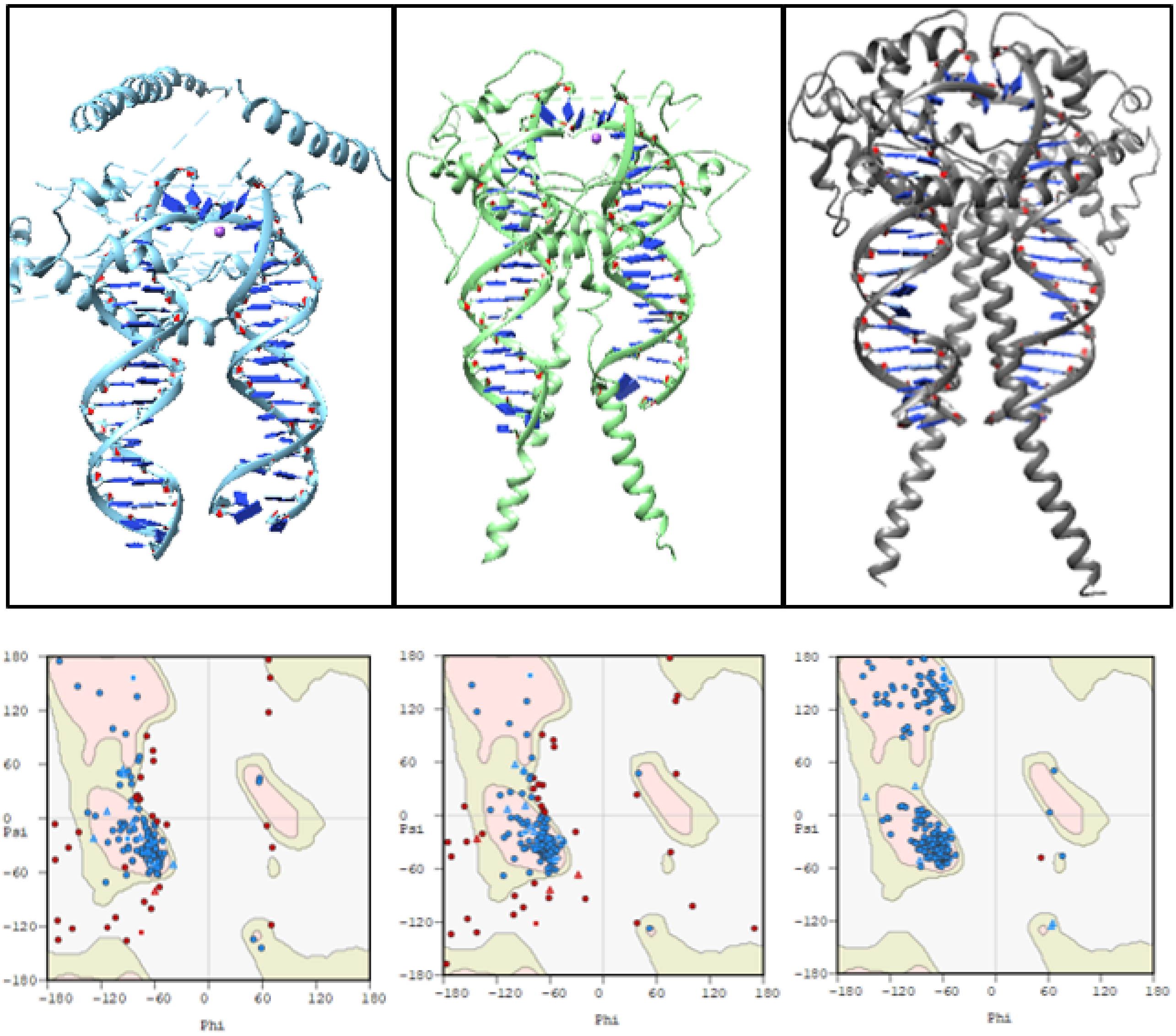


Figure R11, map improvement by density modification. *Snapshots of the 2mF$_0$-Fc maps and the atomic model after MR-SAD (A) and (C) and after density modification (B) and (D). The 2mF$_0$-Fc maps (in blue) are at 2.0 rmsd. The red density corresponds to the anomalous differences map at 5.0 rmsd*



In Preferred Regions:	146 (73.74%)	In Preferred Regions:	139 (70.20%)	In Preferred Regions:	359 (97.03%)
In Allowed Regions:	19 (9.60%)	In Allowed Regions:	19 (9.60%)	In Allowed Regions:	10 (2.7%)
Outliers:	33 (16.67%)	Outliers:	40 (20.20%)	Outliers:	2 (0.54%)
Agreement between model and data		***Agreement between model and data***		***Agreement between model and data***	
Rwork:	0.3977	Rwork:	0.3847	Rwork:	0.2415
Rfree:	0.4388	Rfree:	0.4107	Rfree:	0.2821

Figure R12, Iterative model building and refinement of Gcf1p/DNA structure. *In the top row, snapshots of the Gcf1p structure in complex with Af2_20 throughout manual model building and refinement. In the middle row, the Ramachandran plots of respective structures shown above. Improvement of both the model geometry and its agreement with the experimental data are shown in the bottom lane, by the Ramachandran statistics and R-factors Rfree and Rwork.*

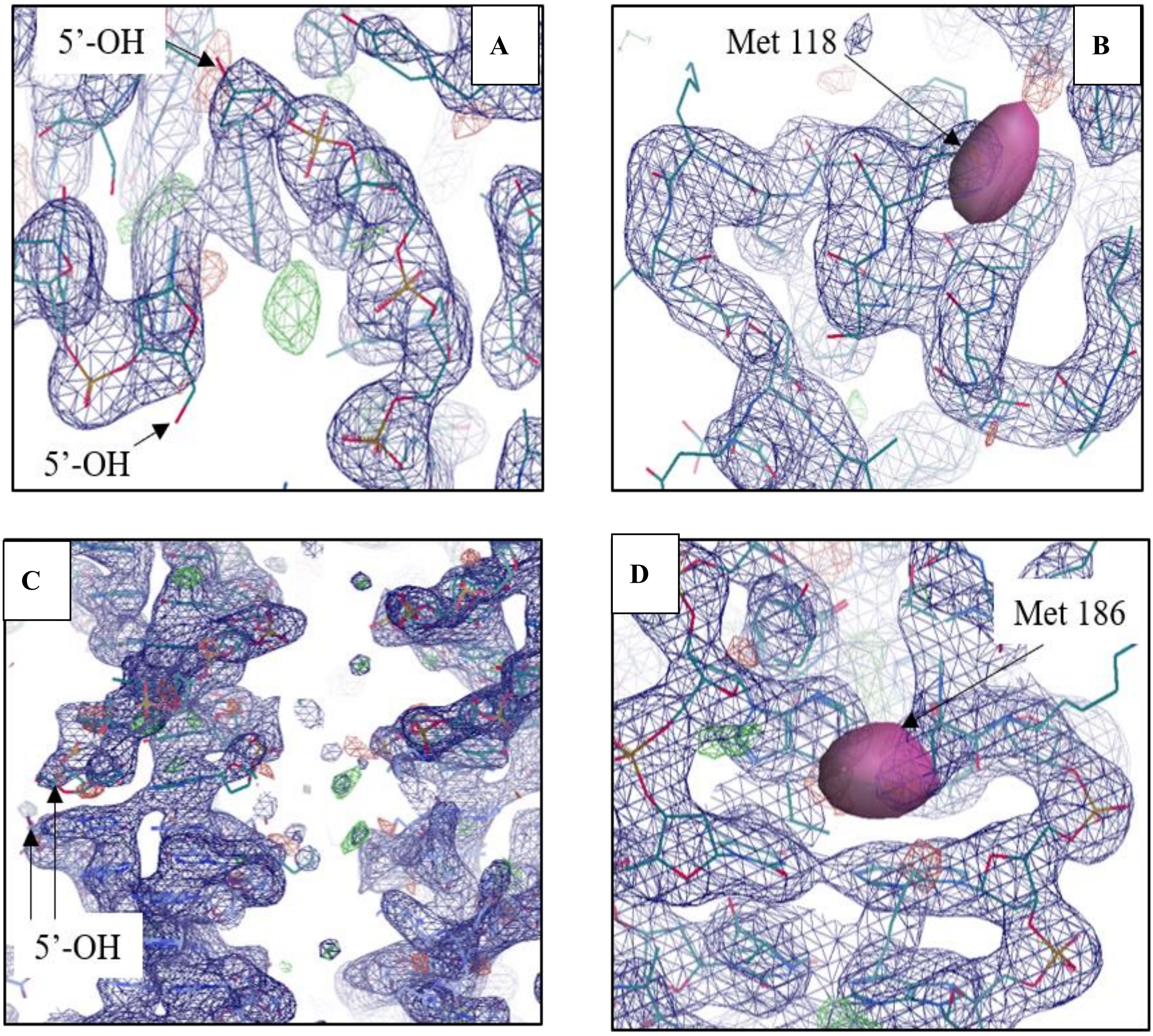


Figure R13. Confirmation of the sequence register. *(**A**) The good quality of the maps around the DNA unambiguously showed the DNA ends; due to the two-fold non-crystallographic symmetry operation along the y axis, the 5'-5' (and 3'-3') DNA ends confront. (**C**) By virtue of the crystallographic two-fold axis in z, two DNA molecules also face 5'-5' and 3'-3' ends in the other DNA end. The sequence register could be assigned in part of the protein model thanks to the strong anomalous peaks from selenium, shown in purple at 6.0 rmsd while the 2mF$_0$-F$_C$ map is shown in blue at 2.4 rmsd (**B**) and (**D**).*

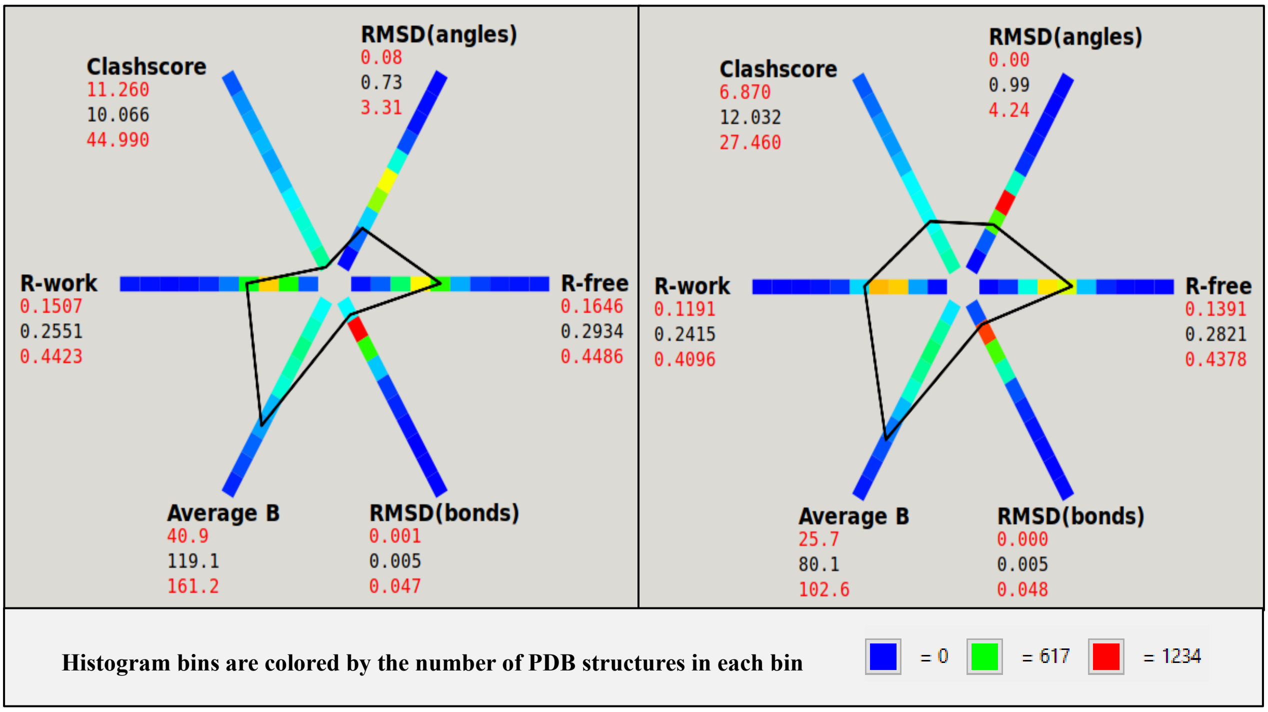


Figure R14, statistics of the refined models: *Validation statistics of the model versus available PDB structures with similar resolution. Statistics correspond to the same model either refined versus a spherically truncated dataset (left) or an elliptically truncated dataset (right). The graphics show a blue to red scale defining value ranges for each parameter. The closer to the red, the more structures report that value for the corresponding statistical parameter.*

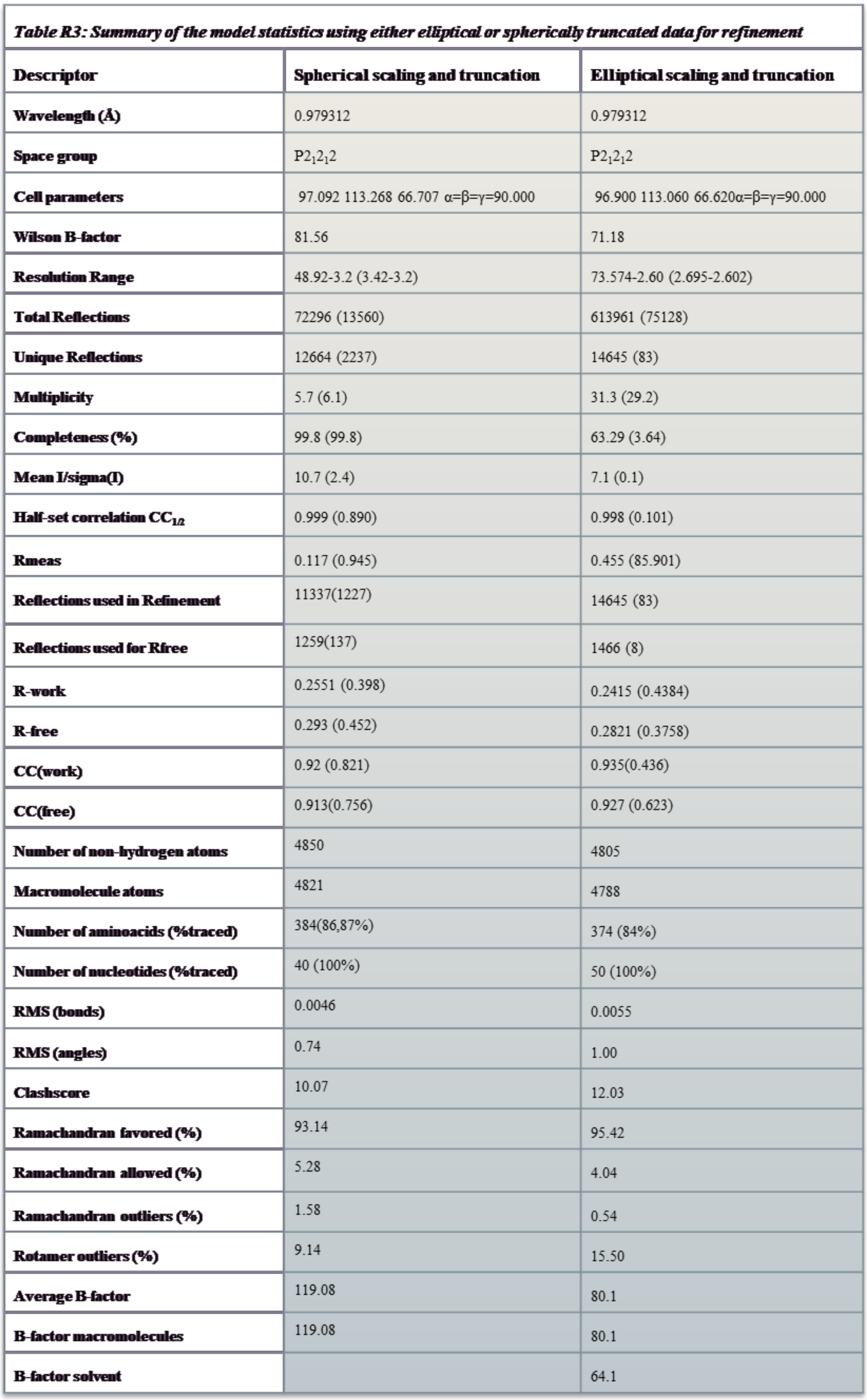


Table R3: Summary of the model statistics using either elliptical or spherically truncated data for refinement

Descriptor	Spherical scaling and truncation	Elliptical scaling and truncation
Wavelength (Å)	0.979312	0.979312
Space group	$P2_12_12$	$P2_12_12$
Cell parameters	97.092 113.268 66.707 α=β=γ=90.000	96.900 113.060 66.620α=β=γ=90.000
Wilson B-factor	81.56	71.18
Resolution Range	48.92-3.2 (3.42-3.2)	73.574-2.60 (2.695-2.602)
Total Reflections	72296 (13560)	613961 (75128)
Unique Reflections	12664 (2237)	14645 (83)
Multiplicity	5.7 (6.1)	31.3 (29.2)
Completeness (%)	99.8 (99.8)	63.29 (3.64)
Mean I/sigma(I)	10.7 (2.4)	7.1 (0.1)
Half-set correlation $CC_{1/2}$	0.999 (0.890)	0.998 (0.101)
Rmeas	0.117 (0.945)	0.455 (85.901)
Reflections used in Refinement	11337(1227)	14645 (83)
Reflections used for Rfree	1259(137)	1466 (8)
R-work	0.2551 (0.398)	0.2415 (0.4384)
R-free	0.293 (0.452)	0.2821 (0.3758)
CC(work)	0.92 (0.821)	0.935(0.436)
CC(free)	0.913(0.756)	0.927 (0.623)
Number of non-hydrogen atoms	4850	4805
Macromolecule atoms	4821	4788
Number of aminoacids (%traced)	384(86,87%)	374 (84%)
Number of nucleotides (%traced)	40 (100%)	50 (100%)
RMS (bonds)	0.0046	0.0055
RMS (angles)	0.74	1.00
Clashscore	10.07	12.03
Ramachandran favored (%)	93.14	95.42
Ramachandran allowed (%)	5.28	4.04
Ramachandran outliers (%)	1.58	0.54
Rotamer outliers (%)	9.14	15.50
Average B-factor	119.08	80.1
B-factor macromolecules	119.08	80.1
B-factor solvent		64.1

R3.6 Analysis of Gcf1p/Af2_20 crystal structure

The crystal structure of Gcf1p in complex with Af2_20 DNA reveals two Gcf1p molecules and two Af2_20 double-stranded DNAs in the asymmetric unit (a.u.). The structure was built with no gaps from residue 49 to 245, so that aminoacids 25 to 48 present in the construct were not traced due to very poor electron density. Interestingly, the three predicted helices between residues 108-155 unexpectedly fold in an unpredicted HMG domain, the HMG-box 1 (**Figure R15A)**.

R3.6.1 Overall structure of Gcf1p

In the a.u., two Gcf1p proteins (chains A and B, all-atom r.m.s.d. 0.849 Å) are related by a non-crystallographic symmetry two-fold axis and interact simultaneously with two DNAs (complementary chains WX and YZ, respectively) as displayed in **Figure R15A**. Gcf1p consists of three domains that include a long N-terminal helix (residues 59-104), followed by HMG-box 1 (residues 108-155) and HMG-box 2 (residues 159-245). Gcf1p HMG-boxes display its typical fold, consisting in an extended region preceding three helices, namely helix 1, 2 and 3. Helix 3 packs antiparallel to the extended region whereas, in between, helix 1 and helix 2-fold in a small helix bundle of two helices (see **Figure R16F**). These elements fold in an L-shape (**Figure R15**). Gcf1p folds in a hammer-like shape, in which the N-terminal helix would be the hammer handle and both HMG-boxes, situated at opposite sides of the vertical axis determined by the N-terminal helix, as a hammer head. (**Figure R15;** see a schematic representation of the secondary structure elements in **Figure R16**). At the junction between the hammer head (both HMG boxes) and the handle (long N-terminal helix), there are both the linker that connects the HMG domains (aa 156-158), and helix 3 from HMG-box 2 that runs backwards towards helix 3 from HMG-box-1, with whom interacts further sealing the connection between the two domains.

Each HMG-box contacts an independent DNA molecule, but in a surprisingly different manner. By virtue of the pseudo-symmetry two-fold axis, HMG-box1 of molecule A (HMG-box1A) and HMG-box2 from molecule B (HMG-box2B) are placed at opposite faces of the same DNA molecule (chains WX). HMG-box2B bends the end of this DNA, whereas HMG-box1A, at the other DNA side, contacts the B-DNA major groove. To the best of our knowledge, such an interaction between HMG-box 1 and the DNA major groove is unprecedented. The same applies for HMG-box2A, which bends the DNA end from chains YZ, while HMG-box1B binds at the DNA major groove.

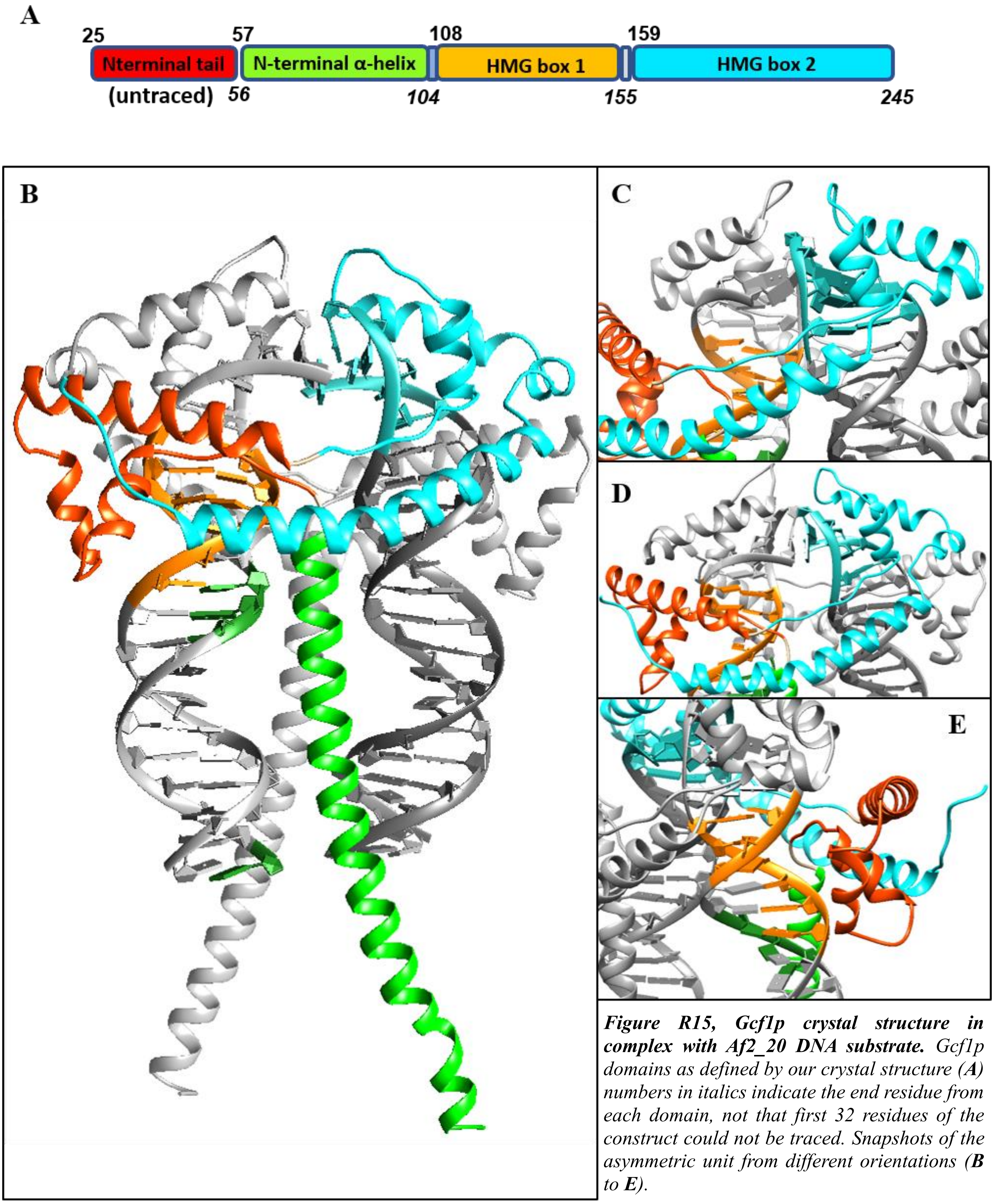


Figure R15, Gcf1p crystal structure in complex with Af2_20 DNA substrate. *Gcf1p domains as defined by our crystal structure (**A**) numbers in italics indicate the end residue from each domain, not that first 32 residues of the construct could not be traced. Snapshots of the asymmetric unit from different orientations (**B** to **E**).*

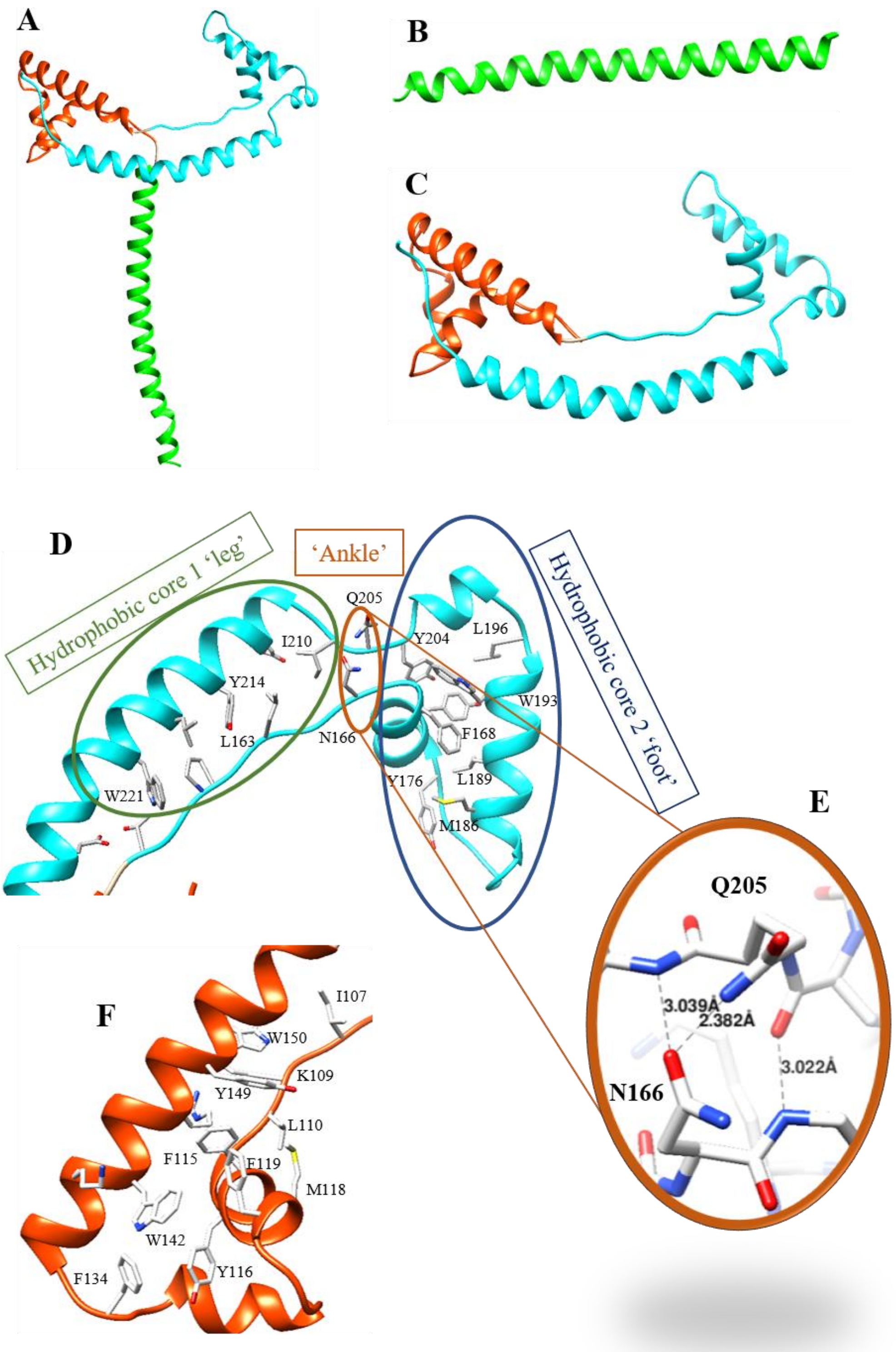


Figure R16: *Representation of the structure of one Gcf1p molecule within the asymmetric unit (**A**). In (**B**) a representation of the elongated helix in Gcf1p N-terminus (residue range 59-104). In (**C**) the globular region of Gcf1p containing the two HMG-box DNA binding domains. In (**D**) the two hydrophobic cores within HMG-box 2 and the 'ankle' in between them (zoomed in in E). In (F) The HMG-box 1 with the hydrophobic residues forming its only hydrophobic core.*

The two HMG-box domains show different dimensions but a conserved hydrophobic core

In *C. albicans* Gcf1p, HMG-box 1 (108-155, 47 aa) is smaller than HMG-box 2 (159-245, 85 aa). The elongated region of HMG-box 1 is four residues shorter than that of HMG-box 2, helices 1 and 2 from HMG-box 1 are connected by a four-residue long 'Type I' tight turn [139] while helices 1 and 2 of HMG-box 2 are connected by a six residue-long loop, mainly in β-conformation. Finally, helix 2 of HMG-box 1 is two residues shorter than that of HMG-box 2. Thus, HMG-box 1 displays a very compact small L-shape (see **Figure R16F**) that could be related with its interaction with the DNA major groove (see below). Despite the differences in size, both domains are structured around hydrophobic cores rich in aromatic rings. At HMG-box 1, Tyr116 from helix 1, Phe131 from helix 2, Phe134 at the loop between helices 2 and 3, and finally the aliphatic region of Lys 139 together with Trp142 at helix 3, make a highly conserved hydrophobic core at the elbow of the L-shape. This core is critical for the architecture of [140] HMG-boxes, as illustrated in **Figure R16**. Note that π-stacking interactions occur between Tyr 116, Phe131 and Phe134, which are lying parallel to each other at approximately 5 Å distance.

For HMG-box 2, two hydrophobic cores were defined before and after an interruption in helix 3 (195-236) by an elongated conformation of three residues (aa 202-204), which perform a β-strand interaction with the HMG-box elongated region, specifically between the α-carbonyl of Tyr204 and the α-amino of Asn167, which is stabilized by a second interaction between Ile206 α-amino and the Asn 166 δ-amide oxygen atom, forming a sort of "ankle" not observed in other HMG-box domains (see **Figure R16 D,E**). At both sides of the 'ankle' two hydrophobic cores forming a 'leg' and a 'foot' region appear (**Figure R16D**). The first core, above the "ankle" is formed by Leu163 from the elongated region, and Ile 206, Tyr 214 and Leu218 from helix 3, together with the aliphatic regions of both Lys 160 (elongated region) and Lys211 (helix 3). The second core is mainly participated by aromatic residues belonging to the three helices of HMG-box 2. Residues involved are Phe 168, Tyr 171* from helix 1; Trp 193* and Leu 196 from helix 2; and the aliphatic region of Lys 201* and Tyr 204* from helix 3 (indicated with an asterisk are the residues conserved from HMG-box 1). This hydrophobic core extends with additional residues towards the end of the short L-arm, i.e. inside the coiled coil between helices 1 and 2. Thus, the "ankle" sustained by hydrophilic contacts, a feature not present in other two-HMG-box domain proteins such as TFAM (3TMM) and in Abf2p (5GH0), could add structural flexibility to this otherwise highly compact part of the domain.

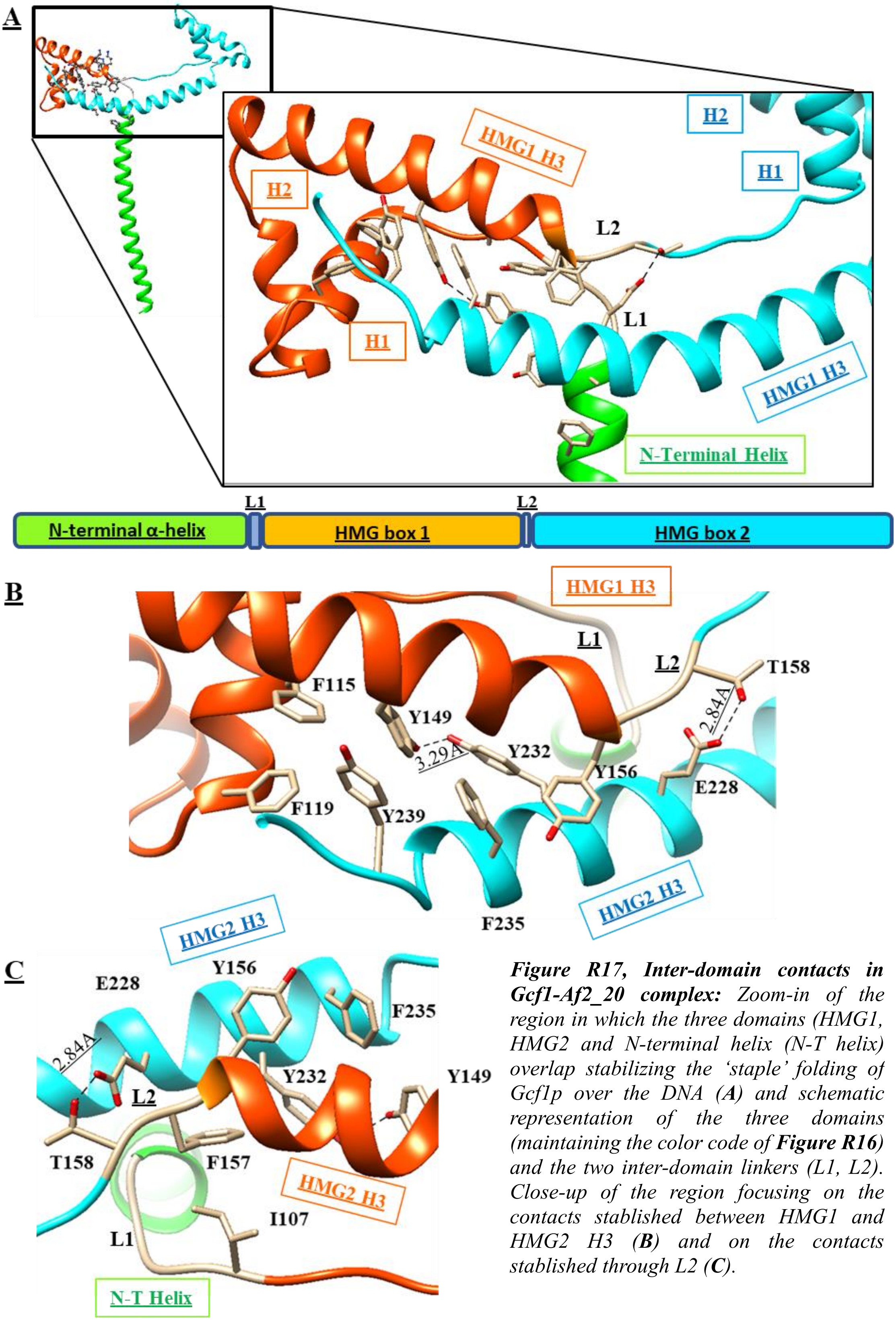


Figure R17, Inter-domain contacts in Gcf1-Af2_20 complex: *Zoom-in of the region in which the three domains (HMG1, HMG2 and N-terminal helix (N-T helix) overlap stabilizing the 'staple' folding of Gcf1p over the DNA (**A**) and schematic representation of the three domains (maintaining the color code of **Figure R16**) and the two inter-domain linkers (L1, L2). Close-up of the region focusing on the contacts stablished between HMG1 and HMG2 H3 (**B**) and on the contacts stablished through L2 (**C**).*

Inter-domain interactions

In the Gcf1p/DNA complex, the three protein domains interact right at the junction between the handle and the head of the hammer-like structure of the protein (see **Figure R17**). At this junction, the end of long N-terminal helix, together with the linker between HMG-box domains (linker L2) and helix 3 from HMG-box 2, contact each other (**Figure R17A**). In addition, helix 3 from HMG-box2 extends beyond the junction by contacting in an antiparallel manner helix 3 from HMG-box 1, with which creates a hydrophobic core that fuses with the core of HMG-box 1 (see **Figure R17B**). Eventually, HMG-box 2 helix 3 reaches HMG-box 1 Helix 1. Such a hydrophobic core between domains is formed by residues Phe 99 (N-terminal α-helix), Phe 115 and Phe 119 (HMG-box 1 Helix 1), Tyr149 (HMG-box 1 Helix 3), and Tyr 232 (HMG2 Helix3, and which forms an hydrogen-bond with Tyr 149), Phe 235, and Tyr 239 (HMG2 H3 and C-terminal tail, respectively) (**Figure R17B**).

The linker between HMG-box 1 and HMG-box 2 (L2), formed by residues Tyr 156, Phe 157 and Thr 158 is also pivotal in the inter-domain interactions at the junction point. L2 establishes a hydrophobic core with Ile 107 from the linker L1 (between N-terminal α-helix and HMG-box 1), and with the aromatic residues Tyr 232 and Phe 235 from HMG-box 2 Helix 3 (see **Figure R17C**). Thr 158 establishes further hydrogen bonds with the carboxyl of Glu 228 lateral chain from HMG-box 2 Helix 3. Besides the hydrophobic cores, additional hydrogen bond interactions are stablished at the very end of the N-terminal α-helix, between the from the guanidinium group of Arg 104 (N-terminal α-helix) and the oxygen of the carbonyl of Lys 222 (Helix 2 from HMG-box 2). Moreover, identical hydrogen bond is stablished between Arg225 from HMG-box 2 helix 3 and carbonyl oxygens of Ala 103 and Ser 105 (see **Figure R18**)

These combination of local hydrogen bonds and extensive hydrophobic contacts suggest a stable lock that seem to fix the disposition of the domains when bound to DNA. This lock might be reversible as in Abf2p [42] but the lack of X-ray diffraction data from the protein alone prevented us from extracting further conclusions at this point.

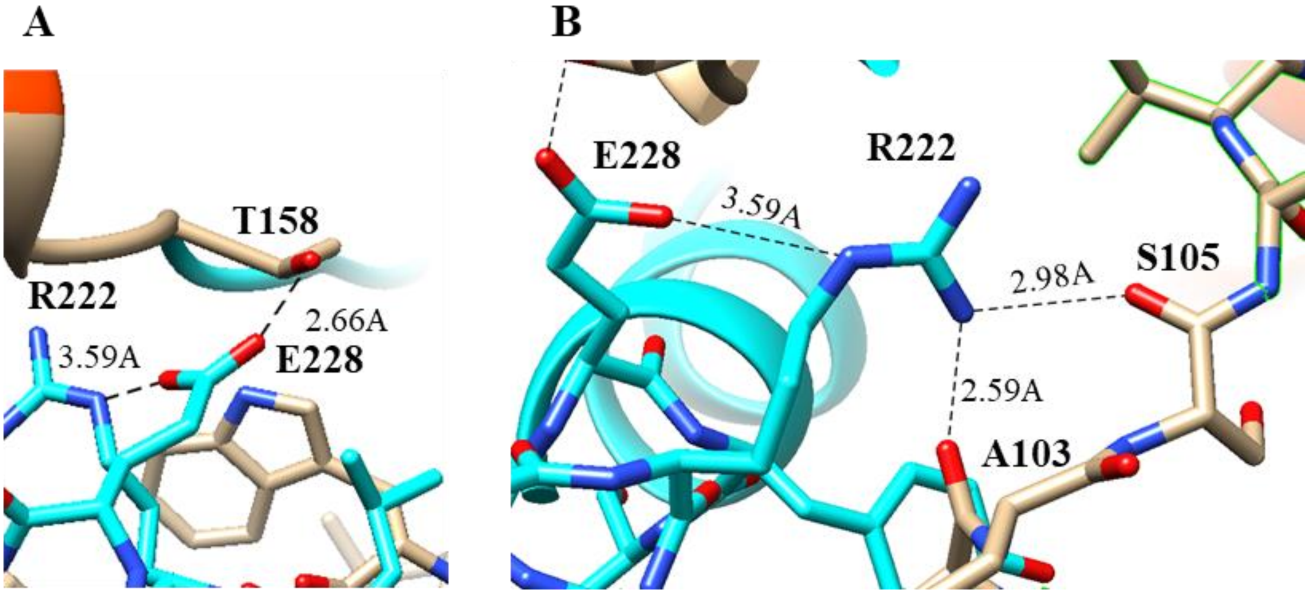


Figure R18, inter-domain hydrogen bonding and salt bridge interaction at Gcf1p inter-domain junction point. *Interaction stablished between L2 and HMG2-H3 (**A**) and between L1 and HMG2-H3 (**B**).*

R3.6.2 Protein-DNA interactions by the HMG-boxes

By means of the non-crystallographic two-fold axis, the two Gcf1p molecules A and B contact two independent DNA molecules (pairs WX, YZ) by the DNA binding domains HMG-box 1, HMG-box 2, and the long N-terminal helix, as shown in **Figure R15**. The four HMG-boxes, from chains A and B, completely cover base pairs 11 to 20 of both DNA molecule ends, like a cap (as illustrated in **Figure R19**).

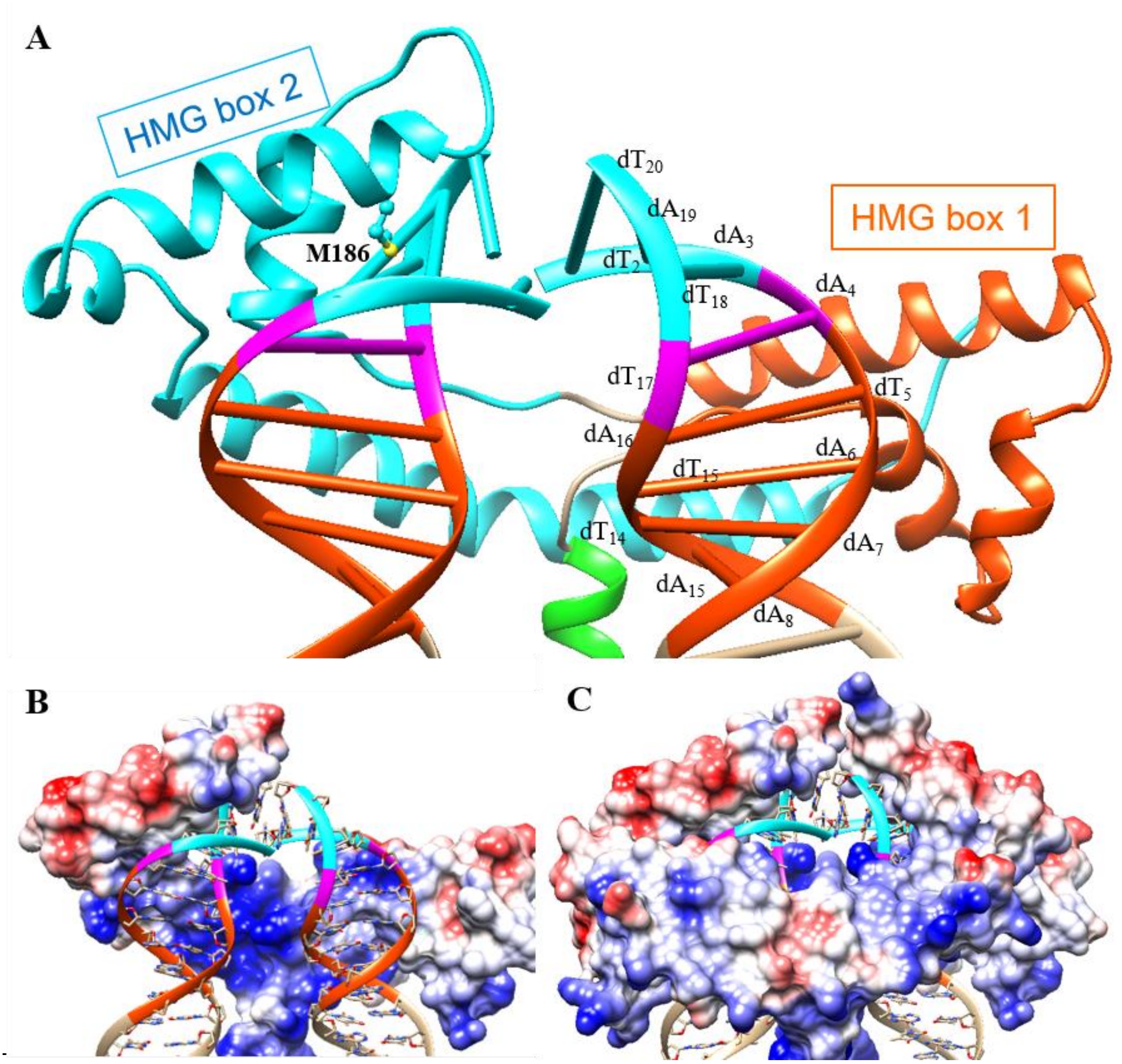


Figure R19, interaction of Gcf1p with 2 double-stranded DNA molecules. *In (**A**) binding of one Gcf1p chain (chain B of our model) to two DNA molecules through HMG-box 1 (chain YZ of our model) and HMG-box 2 (chain WX of our model), chain A of our model is not shown. The insertion residue from HMG-box 2 (Met 186) is shown. In (**B**) a surface, atomic Van der Waals radii, model representation with chain B and in (**C**) with chains A and B of our model.*

The surface of Gcf1p that faces the solvent shows an equivalent distribution of positively and negatively electrostatic patches, whereas all regions from the HMG-box domains and the N-terminal α-helix that contact the DNA phosphate backbone are widely positively charged and compensate the strong negative charge of the nucleic acids (**Figure R19**). As commented above, contacts of the HMG-box domains with DNA are radically different, since HMG-box 2 performs a canonic interaction with the minor groove, while HMG-box 1 unprecedentedly interacts with the DNA major groove by an unexpected novel mechanism.

We first describe the contact between HMG-box 2 (molecule B) and the DNA (chains WX) so it serves as a reference for the canonical contacts. The L-shape concave surface of HMG-box 2 contacts the minor groove of the DNA, between base pairs 17 to 20 (chain W numbering). This involves both the main and side chains of residues from Leu 163 (at the N-terminal elongated region within HMG-box 2) to Lys 215 (HMG2-Helix3 beyond the "ankle" region). Whereas the phosphate backbone is contacted by polar residues, all interactions with base pairs are performed by hydrophobic residues. A critical interaction is performed by Met186 (HMG-box 2 Helix 2) that inserts at base-pair step $T_{18}A_{19}/T_2A_3$ (strands WX), completely disrupting the base stacking and opening the roll angle between both base pairs by 61 degrees (**Figure R20**). Therefore, Met186 is identified as the insertion wedge typically found in HMG-box domains [141]. The previous base pair ($T_{17}T_{18}/A_3A_4$) shows DNA shearing in the x-axis induced by Phe168 and Ala169 (HMG-box 2 Helix 1). These events facilitate minor groove widening and induce, by means of the insertion, an abrupt bending of almost 90 degrees at the base pairs contacted by HMG-box 2. Interestingly, by virtue of the two-fold axis, both DNA duplexes termini are distorted and make an irregular stacking interaction between the 3' ends of chains Y and W (incompatible with a pseudo-continuous DNA). The two HMG-box 2 that contact such DNA ends also come close and face respective loops between helix 1 and 2. However, these loops show different conformations (see **Figure R21**) and do not make contacts between them. Molecule B shows weaker density but a clear reorientation of the loop, for example Gly179 changes both signs of its angles phi and psi (rotations around N-Cα and Cα–C bonds, respectively) from +105.1º and -3.5º (α-helix conformation or α-region in the Ramachandran plot) to -111.0º and 43.9º (left-handed helix, or Lα-region), and so does Asp180 (-79.1º and 149.6º to 54.3º and 5.85º, respectively). This results in a much flatter loop in molecule A. Apparently, both conformations could co-exist simultaneously, and deeper inspection shows that a contact of Arg178 with Glu71 from a symmetry mate induces the different conformation. This conformational change affects the partial insertion of Leu182 (first turn of helix2) between bases A_{19}-T_{20} , which is visible in molecule A but, in molecule B, the first turn of helix 2 is displaced, the electron density at Leu182 is poorly defined, and Thr183 approximates to the DNA bases (chain X). Even if induced by symmetry contacts, the few contacts between both loops and the poorly defined density suggests that the weak interaction between the tips of the HMG-boxes is induced by the DNA during crystallization. It is conceivable that a slight movement between the two Gcf1p molecules could position properly the 5' and 3' ends in an appropriate orientation for a pseudo continuous U-turn.

In contrast to HMG-box 2, HMG-box 1 faces the DNA major groove and it contacts, like a tweezer, the phosphate backbone of DNA chain X at three different locations (contacts summarized in **Figure R21**). One contact is a salt bridge between Arg 123 guanidinium group (helix 2) and the backbone phosphate of thymine from the T_{10}/A_{11} bp. This interaction is aided by the amide of neighboring Ala 124 that contacts the following phosphate of adenine from A_{11}/T_9 bp. The same phosphate backbone is contacted some positions further, at T_{14} by both Ser105 hydroxyl and the imidazole of His 108 from the elongated region. At the other side of the major groove, Lys 109 (also from the elongated region) contacts the adenine A_4 from the complementary backbone. Unlike HMG-box 2, the L-shape concave surface of HMG-box 1 does not contact the strands or the DNA bases. In addition, the main axis of HMG-box 1 is not parallel to the one of the grooves but rather orthogonal. Such mode of binding suggests that the dimensions of this HMG domain is adapted to the dimensions of the major groove. Indeed, the residue that inserts DNA in other HMG-boxes (e.g. Met186 in Gcf1p HMG-box 2 helix 2), in HMG-box 1 is at the end of a truncated helix 2 and substituted by an alanine residue, whose very short side chain cannot insert between base pairs (see **Figure R22**). The structural comparison of both boxes suggests that such a short HMG-box is not able to bend the DNA by the narrow groove but rather face the major groove and grasp both strands.

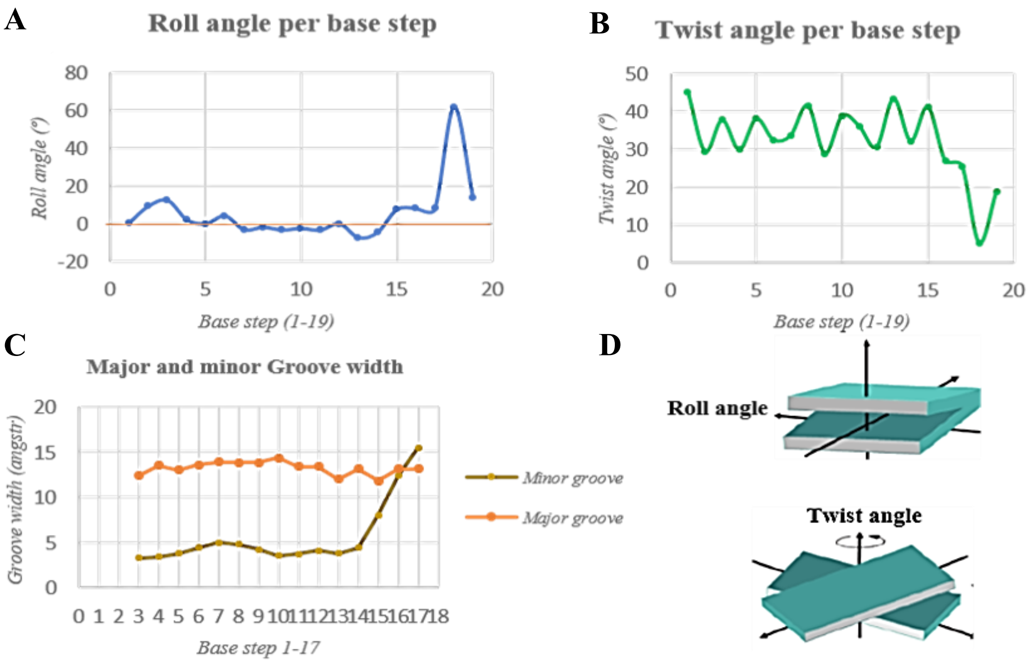


Figure R20. Structural parameters of the Af2_20 substrate when bound to Gcf1p. *Plots representing the roll angle (**A**), the twist angle (**B**) and the major and minor groove width (measured distance between phosphates of complementary base pairs) (**C**). (**D**) graphical representation of a roll and a twist angle within a base pair step.*

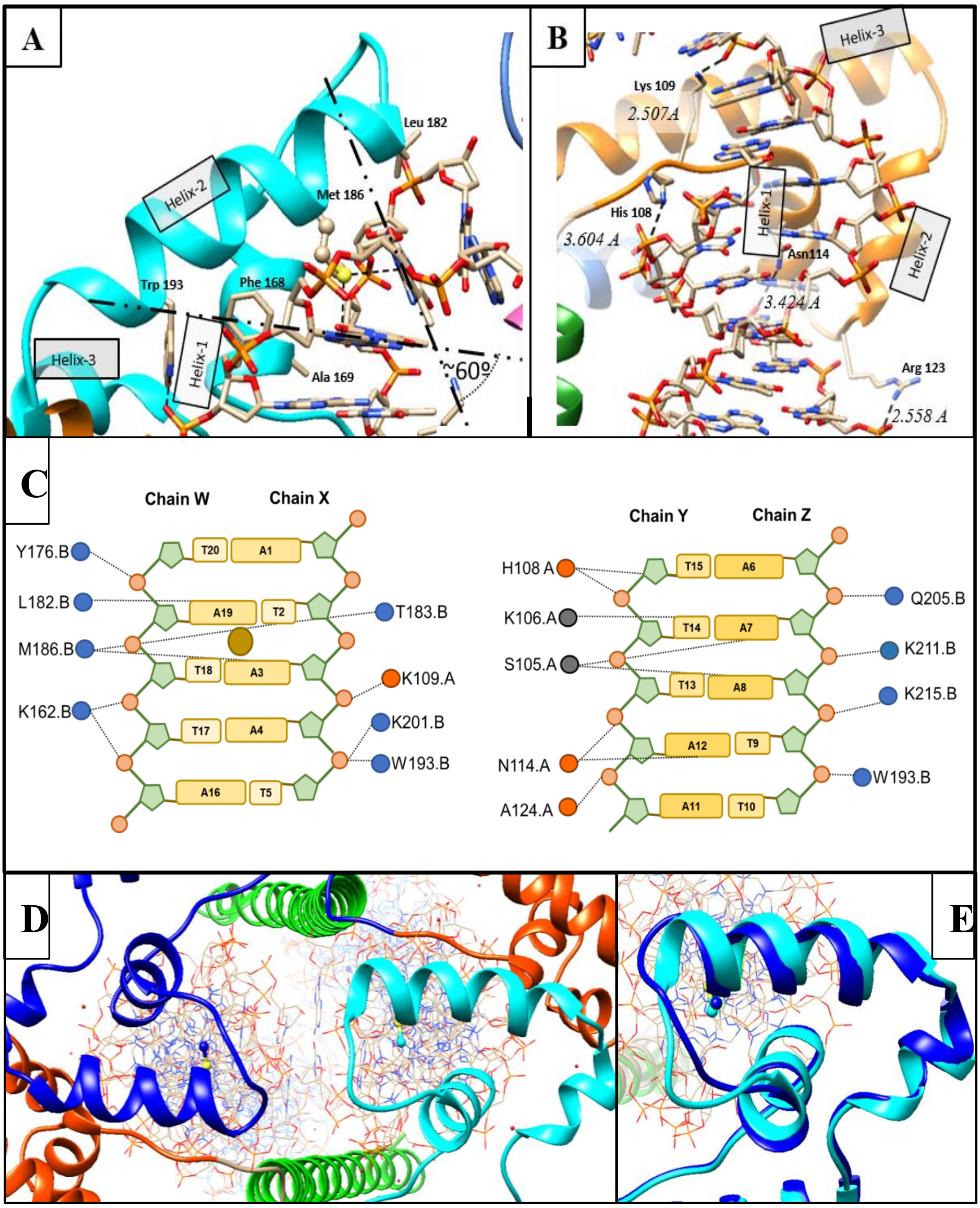


Figure R21, Gcf1p contacts DNA major groove through HMG-box 1 and DNA minor groove through HMG-box 2 displaying conformational variability. *Close-in images of HMG-box 2 of chain B binding to the DNA minor groove (A) and of HMG-box 1 of chain A binding to the DNA major groove (B) with the list of protein-DNA contacts respectively. Conformational variability is displayed between the two HMG-box 2 meeting head-to-head the crystal structure (C and D).*

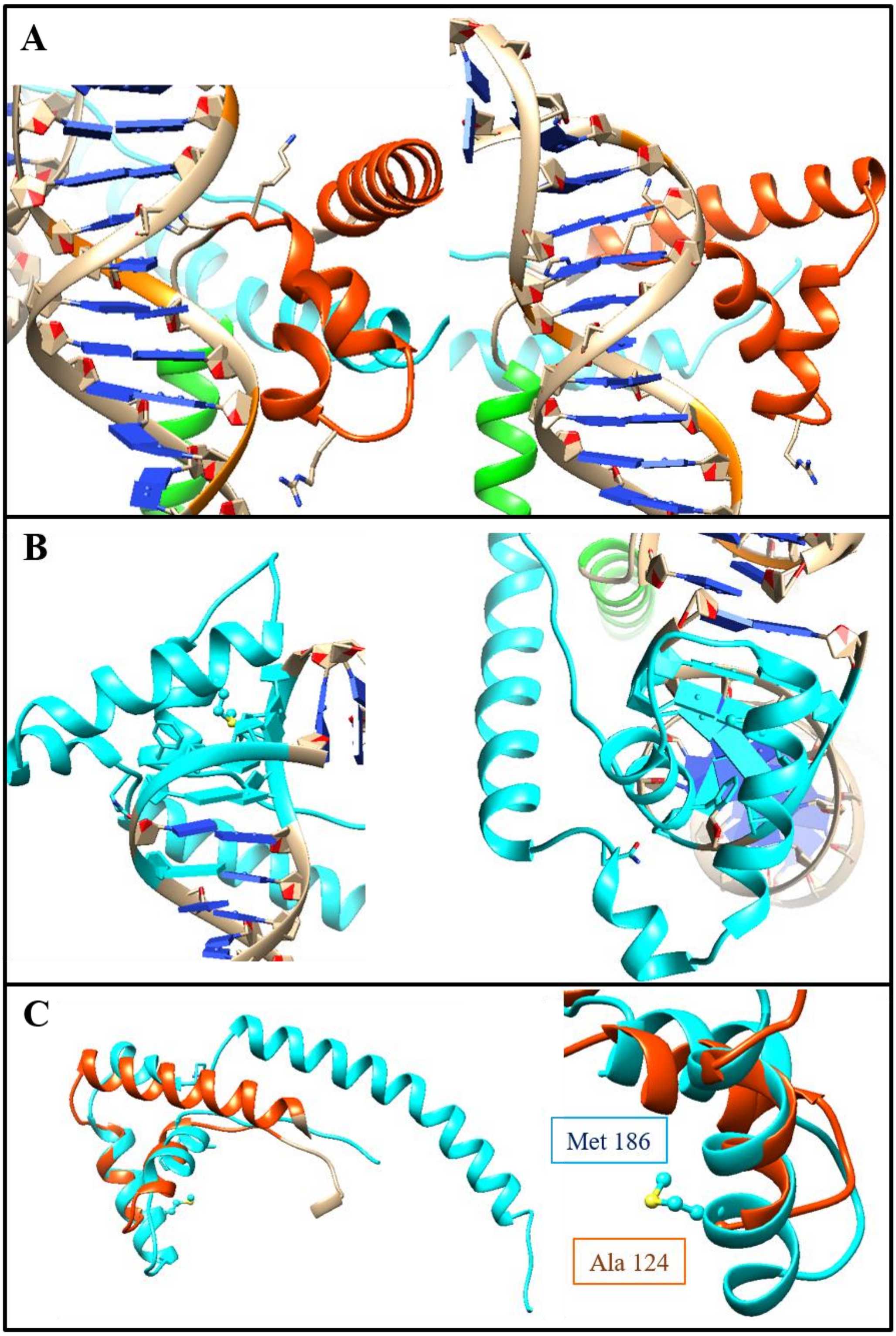


Figure R22, HMG1 and HMG2 domains of Gcf1p display different DNA binding features and geometry. *Close-in of HMG1 (**A**) and HMG2 (**B**) contacting DNA (contacted DNA positions are coloured as the protein domain). Alignment between HMG1 and HMG2 highlighting the position of the intercalating methionine residue (M186)*

R3.6.3 Structure of the DNA in complex with Gcf1p

The protein contacts (see previous section) induce strong deformations to the DNA molecules in our asymmetric unit (a.u.) (**Figure R22**), as a result, the two DNA molecules in the a.u. are arranged forming a splitted U-turn (**Figure R23 A**). Molecules are straight for most of their structure except for the last three base pairs 18-20, which are distorted by HMG-box 2A or 2B and thus respective 3' ends contact each other, as explained in the previous section (**Figure R22B**). Instead, the remaining of the two DNA molecules are almost parallel to each other, both structurally and in sequence. The interaction site with the HMG-boxes involves bp 12 to 20 (yet the contact is mainly to bases 1 to 9 from the complementary chain). Towards the 5' end of chain X, the long N-terminal helix of one protein contacts two regions (1 and 2) of the DNA backbone, separated by one DNA helix turn, of one DNA molecule. At region 1, the ε-amino of both Lys 102 and Lys 98 contact the phosphates from Thy 13 from chain X and Ade 11 from chain W, respectively. At DNA region 2, ε-amino of Lys 91 and γ-hydroxyl of Ser 87 contacts phosphate of Ade 4 from chain W, whereas Lys 84 contacts phosphate of Thy 3 of the same chain. Despite the contacts, the distance between the DNA backbones of DNA molecules WX and YZ is rather constant, e.g. the distance between Ade 20 phosphates from chains X and Z (complementary to Thy 2 from chains W and Y, respectively) is 10.2 Å, and one turn ahead (Ade 11 from chains W and Y) is 9.7 Å. In contrast, the binding of the HMG-1 domains at the major groove brings the straight DNA region closer, to a minimal distance between phosphates of 7.7 Å at position Thy 15 from chains W and Y (**Figure R23C**). Noteworthy, the two negatively charged phosphate backbones are packed at such a short distance with no intervention of positive side chains that would stabilize the DNA electronegative repulsion, which is thus compensated by the overall arrangement of the two HMG-box domains of Gcf1p.

The crystallization of Gcf1p required a DNA with specific properties, i.e. the presence of an adenine-tract conferring local rigidity [142] and a DNA length of exactly 20 bp. The adenine tract requirement might be related to the positioning of Gcf1p onto DNA, as it also occurred with *S. cerevisiae* Abf2p [42]. The fact that a region of the DNA is rigid, may systematically position the protein molecules at a precise location on the DNA oligo used for crystallization. This led to a crystal structure in which both DNA are distorted and interact by the same 3' end, which has no meaning for a continuous DNA at the molecular level. However, the 5' and 3' ends are not far away, at 10 Å on a vertical plane. It is conceivable that if the two HMG-boxes 2 (which perform very weak crystallographic interactions) separate from each other, the 5' and 3' ends (see above) could get closer and built a continuous DNA. This would lead to the formation of a narrow U-turn by binding of two Gcf1p molecules to a long DNA duplex. Our model would also point out to the fact that, if possible, the combined action of two Gcf1p molecules would be needed to form a U-turn.

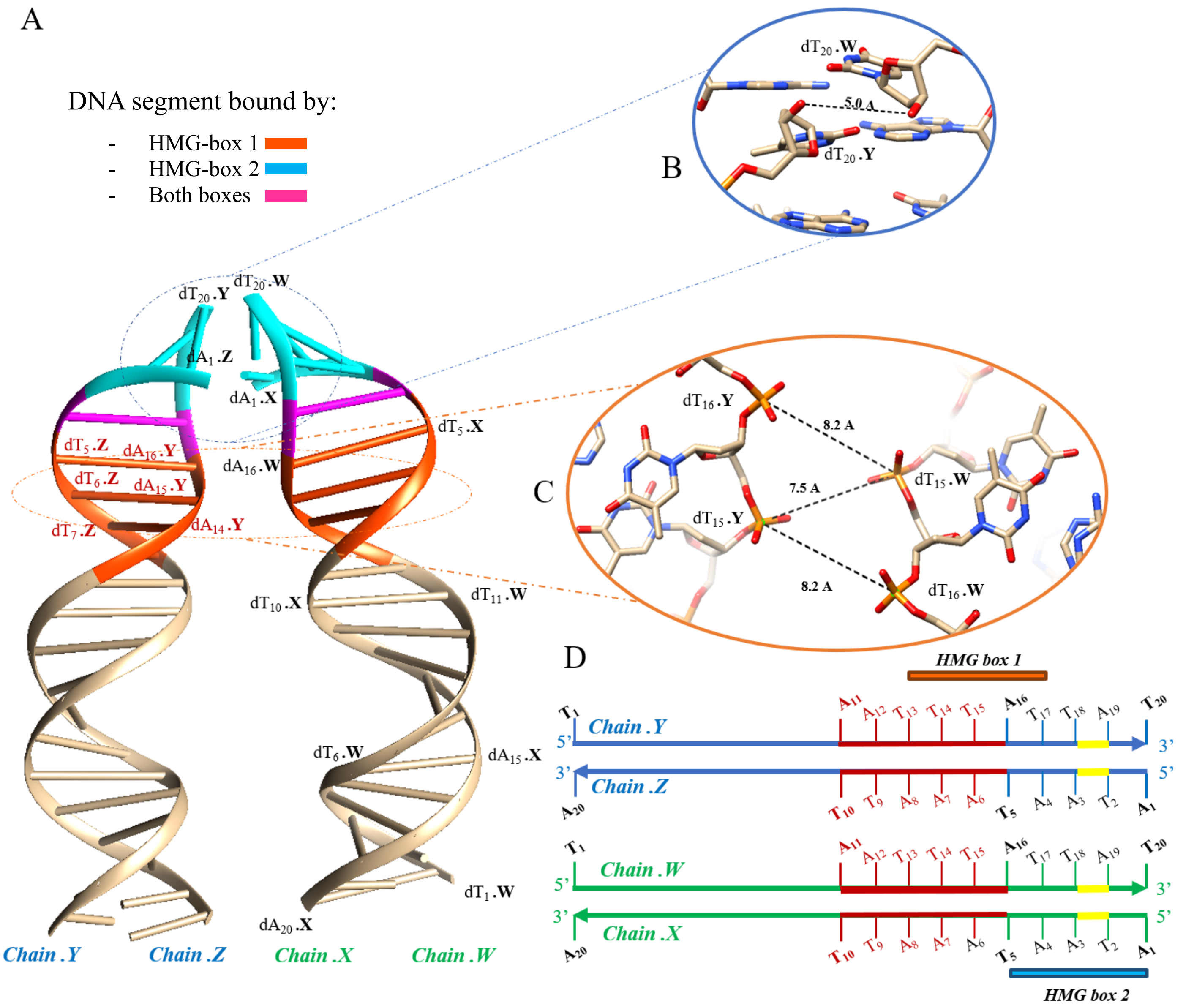


Figure R23. Detailed view of the DNA substrate Af2_20 when bound by Gcf1p. *(**A**) Front view of the two double-stranded DNA molecules as appear in our crystal structure (Gcf1p is not represented). Numbering in red (chains Y and Z) highlight the presence of the A-tract. (**B**) Zoom-in of the region in which the two molecule ends face each other as a result of bending by HMG-box 2. (**C**) Zoom-in of the region in which the distance between DNA phosphate backbones is reduced to 7.5 Å due to the binding of HMG-box 1. (**D**) Scheme of both double-stranded DNA molecules present in the asymmetric unit (YZ and WX). The presence of the A-tract is highlighted in red in both molecules. In yellow there is highlighted the insertion point of the Met 186. Orange and blue cassettes represent the region covered by HMG-box 1 and HMG-box 2 respectively.*

R3.6.3 Gcf1p performs a coiled coil between asymmetric units.

The 45 residue-long amphipathic alpha helices at the N-terminus of Gcf1p form a coiled-coil with a symmetry related protein from an adjacent asymmetric unit, with which the 5' ends of the DNA strands W and Y and the 3' ends of strands X and Z make a stacking interaction. The N-terminal helix has heptad periodicity BxBBxxx (B stands for buried residue). The two N-terminal α-helices intertwine with each other in a levogyre, antiparallel coiled-coil similar to a canonical leucine-zipper [143], but in this case, also buried lysine residues establish hydrophobic interactions through their side chain aliphatic region with hydrophobic residues such as valine, isoleucine and leucine as depicted in **Figure R24**. In order to assess the stability of this interface, the PDBe server PISA that analyses interactions in the crystal [144] was used.

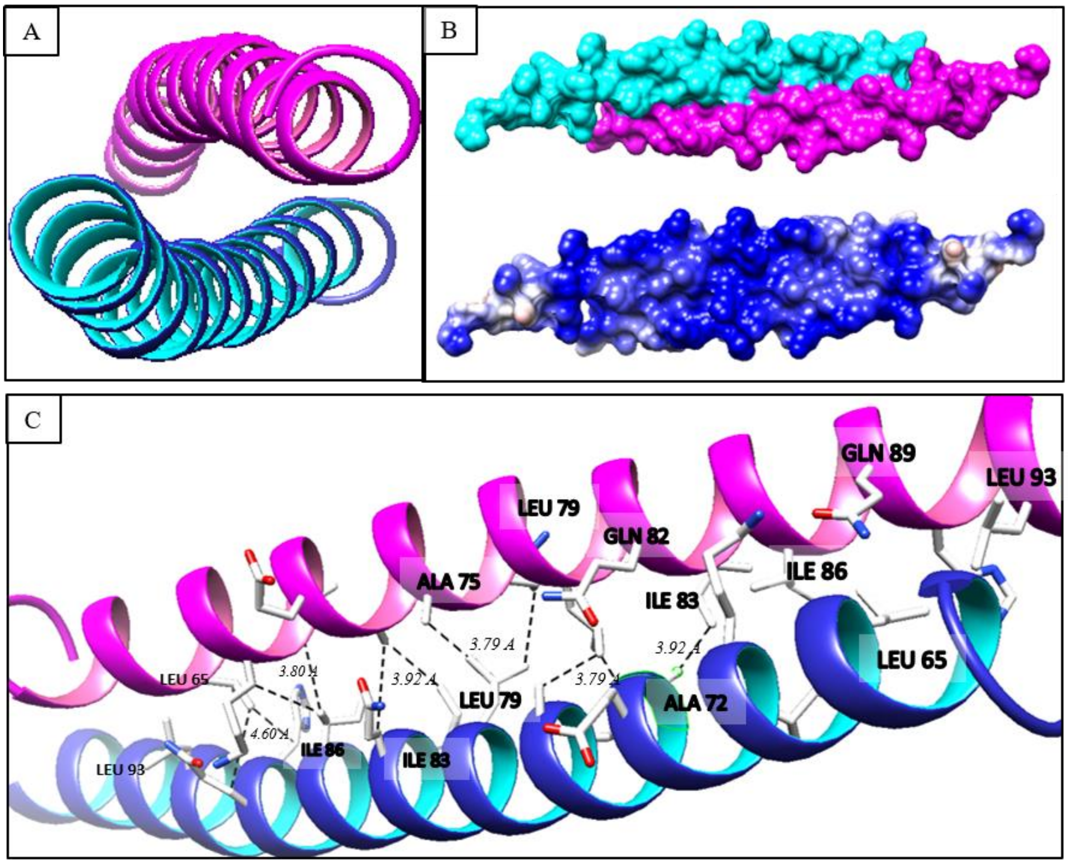


Figure R24. Interactions of the coiled coil between symmetry partners. (A) *Two symmetry-related α-helices intertwine in a levogyre superhelix.* ***(B)*** *the surface representation of the fragment shows a continuous electrostatic positive charge on the surface that contacts the DNA.* ***(C)*** *Buried residues involved in the coiled coil formation. Residues lie within Van-der-Waals attractive range as shown in this figure.*

Amongst all inter-molecule interfaces present in the crystal, the one forming the coiled coil was the one with better stability statistics. The server showed an average value of 994.1 $\dot{A}^2$ interaction surface and a free Gibbs energy of -17.3 kCal/mol for the coiled-coil interface. Furthermore, the Complex formation Significance Score (CSS), which indicates the probability that a given interface is involved in complex formation, yielded a value of 1, strongly suggesting that the coiled coil is a meaningful interaction from the thermodynamic point of view. The formation of such a coiled coil would imply that Gcf1p in complex with Af2_20 DNA assembles in a supramolecular complex formed by four Gcf1p monomers and four DNA fragments of 20 bp, as shown in **Figure R25**. Thermodynamic parameters for such an assembly were also calculated by PISA yielding a total surface area of 67630 $\dot{A}^2$ for a total buried area of 27610 $\dot{A}^2$, a solvation free-energy gain of $\Delta G^{int} = -140.1\ kCal\ /mol$ and a dissociation free energy of $\Delta G^{diss} = 59\ kCal/mol$. All in all, calculations emphasize the thermodynamic stability of the Gcf1p-Af2_20 tetramer hence indicating that this assembly could be biologically relevant. Nevertheless -as stated in PISA server- it is important to emphasize that stability predictions and any conclusions extracted from them must be considered in the light of other experimental evidences. In the following sections, results of other biophysical techniques will be summarized and discussed in order to understand to which extent our crystal structure can explain the biological role of Gcf1p.

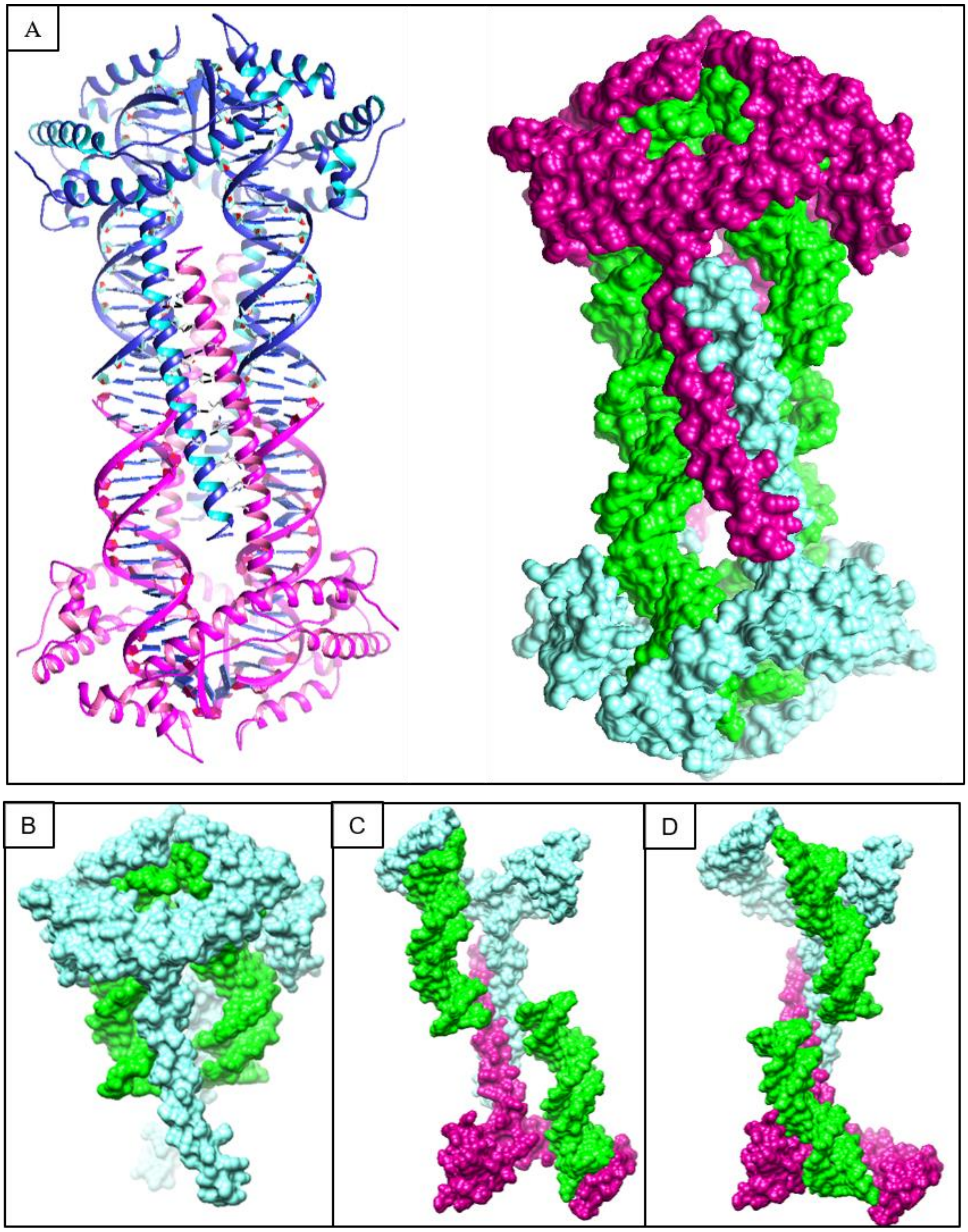


Figure R25. Energetically favored supramolecular assemblies according to PISA server *[144].* ***(A)*** *The tetrameric protein-DNA complexsolvation free energy of -140.1 kCal/mol). Complexes in panels* ***(B)****, (****C****) and (****D****) show a solvation free energy of -50.3, -63.6 and -49.5 respectively.*

R4. Multimerization state analysis by SEC-MALLS

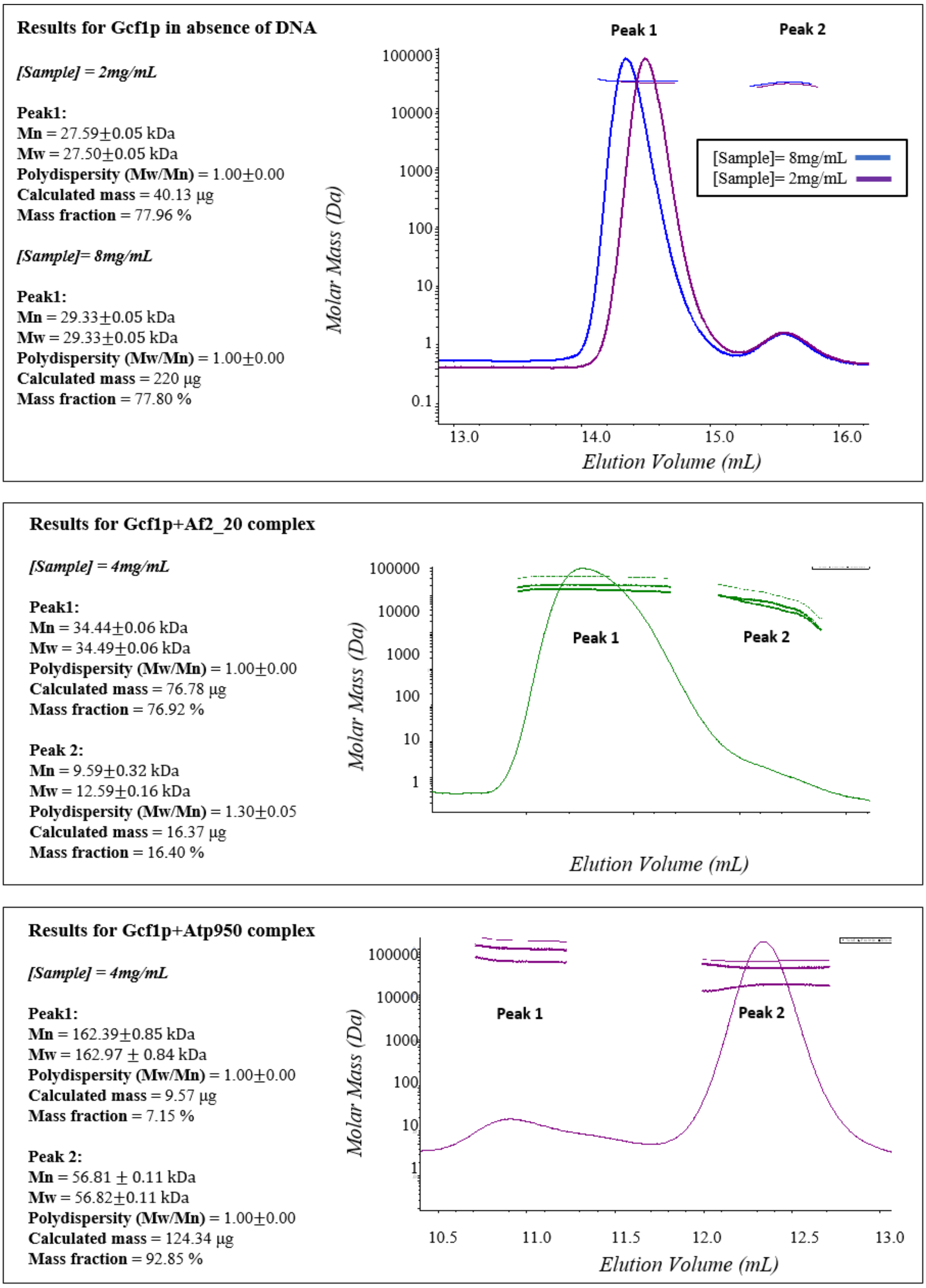
In order to assess the multimerization state of Gcf1p in solution both in the absence and presence of DNA, determination of the absolute molecular weight (MW) for both protein and protein:DNA complexes was performed. Size Exclusion Chromatography coupled to a Multiple Angle Laser Light Scattering detector (SEC-MALLS) was the technique used to perform such analysis.

R4.1. Oligomeric state of Gcf1p in the absence of DNA

Protein samples of 50 µL were run in a size-exclusion Superdex 200 10/300^{TM} (GE-Healthcare®) column coupled to a scattering DAWN-HELEOS-II-detector (Wyatt Technology ®), system was equilibrated with 750 mM NaCl, 50 mM Tris-HCl pH8.0 buffer and scattering measurements were performed at 664.3 nm wavelength. In such a column and buffer conditions, Gcf1p showed an elution volume of 14.5 mL, corresponding to a 50 kDa protein according to column calibration, i.e. a Gcf1p dimer. Nevertheless, scattering measurements confirmed a molecular weight of 27.59±0.05 kDa for 2 mg/mL sample and 29.33±0.05 kDa for 8 mg/mL sample (**Figure R26A**).

Results indicated that Gcf1p was a monomer in the absence of DNA at the protein concentrations and buffer conditions of the assay. The fact that elution occurs at an elution volume corresponding to a molecular weight twice as big as the actual particle size indicates that the free protein has a hydrodynamic radius of a bigger size compared to the globular markers used for calibration and suggests an extended conformation. Note that the protein at higher concentration elutes sooner from the column and the corresponding estimated MW is slightly higher than the protein at lower concentration. This behavior could be indicating a tendency of the protein to multimerize upon concentration, nevertheless polydispersity indicator for both measurements (M_W/M_N) yielded a value of 1.00±0.00, indicating a monodisperse peak, with very small or no contribution from oligomeric species

.



Results for Gcf1p in absence of DNA

[Sample] = 2mg/mL

Peak1:
Mn = 27.59±0.05 kDa
Mw = 27.50±0.05 kDa
Polydispersity (Mw/Mn) = 1.00±0.00
Calculated mass = 40.13 μg
Mass fraction = 77.96 %

[Sample]= 8mg/mL

Peak1:
Mn = 29.33±0.05 kDa
Mw = 29.33±0.05 kDa
Polydispersity (Mw/Mn) = 1.00±0.00
Calculated mass = 220 μg
Mass fraction = 77.80 %

Results for Gcf1p+Af2_20 complex

[Sample] = 4mg/mL

Peak1:
Mn = 34.44±0.06 kDa
Mw = 34.49±0.06 kDa
Polydispersity (Mw/Mn) = 1.00±0.00
Calculated mass = 76.78 μg
Mass fraction = 76.92 %

Peak 2:
Mn = 9.59±0.32 kDa
Mw = 12.59±0.16 kDa
Polydispersity (Mw/Mn) = 1.30±0.05
Calculated mass = 16.37 μg
Mass fraction = 16.40 %

Results for Gcf1p+Atp950 complex

[Sample] = 4mg/mL

Peak1:
Mn = 162.39±0.85 kDa
Mw = 162.97 ± 0.84 kDa
Polydispersity (Mw/Mn) = 1.00±0.00
Calculated mass = 9.57 μg
Mass fraction = 7.15 %

Peak 2:
Mn = 56.81 ± 0.11 kDa
Mw = 56.82±0.11 kDa
Polydispersity (Mw/Mn) = 1.00±0.00
Calculated mass = 124.34 μg
Mass fraction = 92.85 %

Figure R26. MW calculations of Gcf1p alone and in complex with two DNA substrates by SEC-MALLS. *(**Top**) Gcf1p alone at 2 mg/mL and 8 mg/mL. (**middle**) Gcf1p in complex with Af2_20 DNA substrate (20 bp) and (**bottom**) Gcf1p in complex with Atp950 substrate (50 bp).*

R4.2. Oligomeric state of Gcf1p in complex with DNA

In order to analyse the oligomerization state of Gcf1p in complex with DNA, we performed SEC-MALLS analysis of Gcf1p in complex with Af2_20 and Atp950 which are double-stranded DNA substrates of 20 and 50 base pairs respectively. Protein:DNA complexes were prepared at 1:1 ratio following the three-step dialysis procedure that stabilizes complex formation (*see SEC-MALLS sample preparation section,* **Materials and Methods 7.2**). In both cases, 60 µL sample at 4 mg/mL was injected into a Superdex 200 10/300^{TM} (GE-Healthcare®) column coupled to a MALLS system, so that the absolute MW of the complexes could be determined. In this case, both the columne and the MALLS system were previously equilibrated in a 20 mM NaCl, mM Tris-HCl pH 8.0 buffer. The Gcf1p:Af2_20 DNA complex yielded a major peak corresponding to a 34.5kDa specie (peak 1 in **Figure R26B**), which is highly consistent with a 1:1 protein:DNA complex (25.9kDa protein +12 kDa DNA). A second overlapping peak is also observable, it is polydisperse (M_W/M_N =1.30±0.05) and, considering its molecular weight (9kDa), it may correspond to free DNA molecule (MW=12 kDa) from dissociated protein:DNA complex.

Results with the Gcf1p:Atp950 yielded a more complex pattern, with a main peak of 56.82 kDa mass and an earlier eluting minor peak of 162.97 kDa. The main peak can only correspond to 1:1 protein:Atp950 complex (25.9kDa protein + 25kDa DNA). The earlier, minor peak can only correspond to a 4:2 protein:DNA complex, as it is shown in **Figure R26C**. To verify the presence of such complexes, a second aliquot of the concentrated protein was loaded onto another available Superdex 200 10/300^{TM} column, elution fractions were later loaded to native acrylamide gel. The elution form this second SEC showed the same pattern of two peaks as the MALLS column in this case, the smaller peak appear as a shoulder fused to the main peak, see **Figure R27A**. Eluted fractions were loaded in a native PAGE that additionally contained control samples of free DNA and a sample of a Gcf1p:Af2_20 complex at 10 mg/mL, the concentration at which crystal plates are set up. The gel was stained with SYBR Safe 1X in order to reveal the presence of DNA. It showed that the first fraction loaded into the gel, corresponding to an elution volume of 10.5-11.0 mL, contained a band with slow mobility suggesting a big complex, whereas the following fractions showed this band and a heterogeneous population of protein:DNA complexes with faster mobility, probably reflecting different stoichiometries due to progressive decomposition of the biggest protein/DNA complex along the gel filtration run. Surprisingly, the band with the slowest mobility showed the same electrophoretic mobility as the Gcf1p:/Af2_20 complex loaded as a control at high protein concentration (0.5µL at 10 mg/mL), see **Figure R27B**.

The earliest eluting peak of the Gcf1p/Atp950 complex has a molecular weight compatible with a 4 protein:2 DNA$_{(50\ bp)}$ stoichiometry. Results suggest that this complex follows the multimeric assembly described in our crystal structure, although in this case it would contain two 50 bp DNA molecules instead of four 20 bp DNA establishing stacking interactions. In such arrangement, the coiled coil interface could be involved in the stabilization of the multimeric form. Complexes formed with Af2_20 DNA yield a band of exactly the same electrophoretic mobility when loaded into a gel after complex formation, nevertheless this band disappears after a gel filtration run (see **Figure R27B**), suggesting that the tetrameric complex observed in our crystal structure is present in solution, although it can be disrupted by gel filtration. Results indicate that this multimerization behaviour is only present when Gcf1p is bound to DNA. Furthermore, results suggest a concentration-dependent phenomenon since the complexes are dissociated during the Size Exclusion Chromatography (which dilutes the sample into a bigger volume and exerts a pressure on them). Such a concentration dependency would explain why the tetrameric complex formed with 20 bp DNA substrate (partially maintained through weak stacking interactions) is not detectable by SEC-MALLS.

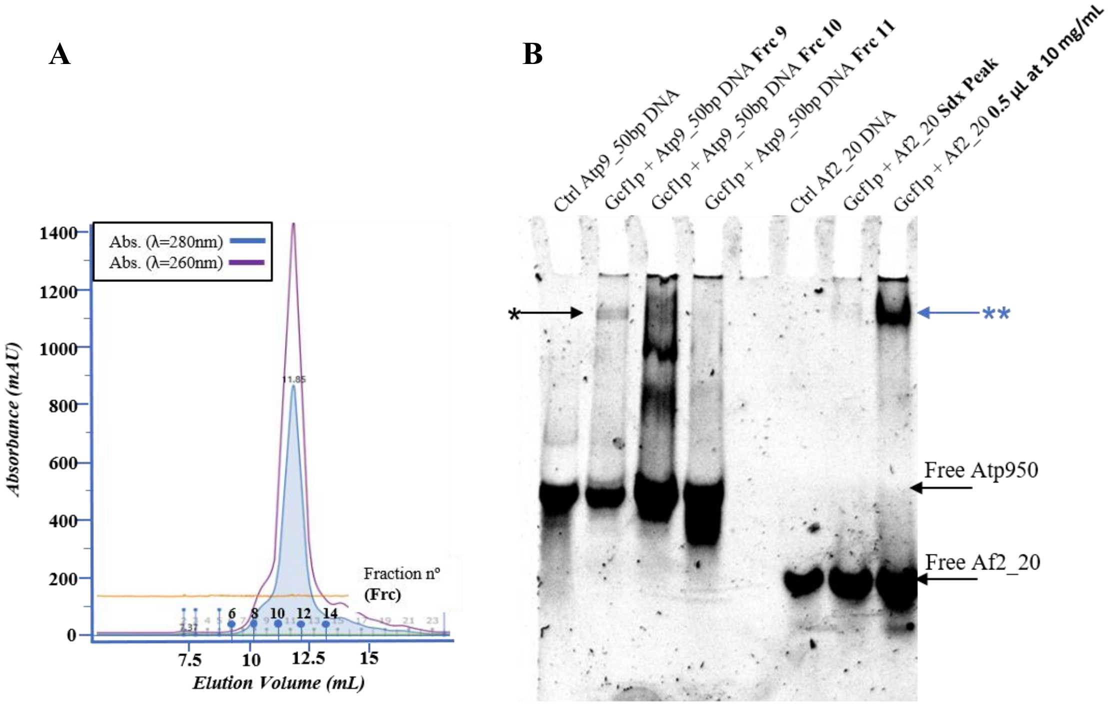


Figure R27: Analysis of the Gcf1p/ Atp950 complex. *(**A**) Elution profile of Gcf1p:Atp9_50bp substrate in a Superdex 200 10/300^{TM} different than that used for SEC-MALLS. (**B**) Fractions eluting from the column were directly loaded in a native acrylamide gel (10%). By the gel, arrow indicate the mobility of both DNA substrates when not bound to the protein, as well as the Gcf1p:Atp9_50bp complex of slower mobility (*) and the complex Gcf1p:Af2_20 prior to gel filtration (**).*

R5. Structural analysis of Gcf1p and its DNA complexes in solution

R5.1 Analysis of Gcf1p in solution in the absence of DNA

In order to further characterize the multimerization state of Gcf1p on the DNA in solution, we analysed the complex by Small-Angle X-ray Scattering (SAXS). Two variants of Gcf1p were tested, the full-length construct Gcf1p25-245 and the Gcf1p53-245 in which the N-terminal disordered tail not visible in the structure was removed. Furthermore, considering that the crystallized structure is traced from residue 58 on due to intrinsic disorder of aminoacids 25-57, SAXS was an ideal option for the characterization of this region.

Full-length Gcf1p samples were analysed at the BioSAXS beamline BM29 at ESRF (Grenoble, France). Measurements of the protein in absence of DNA were performed at high salt buffer (750 mM NaCl, 50 mM Tris-HCl pH 8.0). In order to characterize the apparent concentration dependent multimerization that previous SEC-MALLS analysis suggested, concentration series (1.25 mg/mL, 2.5 mg/mL, 5.0 mg/mL, 10 mg/mL) were measured. Superposition of the corresponding scattering curves showed no concentration-dependent oligomerization occurred for the free protein (see **Figure R28A**). However, whereas the part of the curve at lower angles ($q < 0.5$ nm^{-1}) is well defined at both low and high concentrations, at high concentration this region may reflect undesired inter-particle interactions due to a concentration effect. So, for this part of the curve, is preferable to analyse the low concentration samples. Instead, at higher angles (the curve region corresponding to 0.5 nm^{-1} < q < 5 nm^{-1}) the low concentration sample may become noisier while at high concertation the signal contrasts better. An approach to solve this problem is to select and combine distinct regions from low and high concentration curves. Thus, the Guinier region of the 1.25 mg/mL sample curve at low angles ($q < 0.5$ nm^{-1}) was combined with the higher angles portion of the 10 mg/mL curve (0.5 nm^{-1} < q < 5 nm^{-1}) in order to minimize the effect of inter-particle interaction whilst maximizing the signal at higher angles.

Consensus Bayesian estimation of the molecular mass from this curve [145] yielded a MW interval of 22750-29950 kDa (Credibility Interval Probability: 90.72%). SAXS results therefore confirmed that Gcf1p is a monomer in solution in the absence of DNA as detected by SEC-MALLS. Furthermore, Gcf1p does not show concentration-dependent oligomerization in the 1.25-10 mg/mL concentration range in absence of DNA. The calculated Radius of Gyration (R_g = 3.76 nm) is much larger than the expected (1.8 nm) from the Gcf1p protein sequence if assuming a globular protein, according to Flory equation ($R_g = 3N^{0.33}, where\ N = number\ of\ residues$), suggesting an extended conformation in the absence of DNA. The corresponding Kratky plot (**Figure R28B**) showed a flat profile at higher q values typical for extended proteins with conformational variability. Moreover, the pair-wise distribution function, or $P_{(r)}$ plot, shows a maximum inter-particle distance of 16.20 nm with two histogram maxima and a strong positive skew, consistent with an extended particle with two main globular domains connected by a flexible linker [149] (*see SAXS theoretical background section,* **Materials and Methods 8.1.3**).

In order to assess the contribution of the disordered N-terminal tail to the solution scattering curve in the free full-length protein, the Gcf1p53-245 mutant was analysed and both types of samples compared. One sample concentration at 2.0 mg/mL, was measured at DESY Synchrotron beamline I-04. MW estimates using Bayesian inference estimated a MW interval of 20850-24640 Da (Credibility Interval Probability = 94.75%) consistent with the 23070.3 Da weight of the deletion construct. The deletion mutant showed a $R_g = 3.28$ nm, slightly shorter than the R_g of full-length Gcf1p ($\Delta R_g = 0.48$ nm) but still much larger than the one expected from the MW of the construct if globular. Although the Kratky and $P_{(r)}$ plots of the Gcf1p53-245 mutant is less defined due to the lower concentration, it shows features similar to that of the full-length protein (**Figure R28**), indicating it has still highly disordered regions and an extended conformation. In the light of this results we can infer that, besides the N-terminal tail, other flexible regions are present in Gcf1p resulting in an extended conformation of the protein when not bound to DNA.

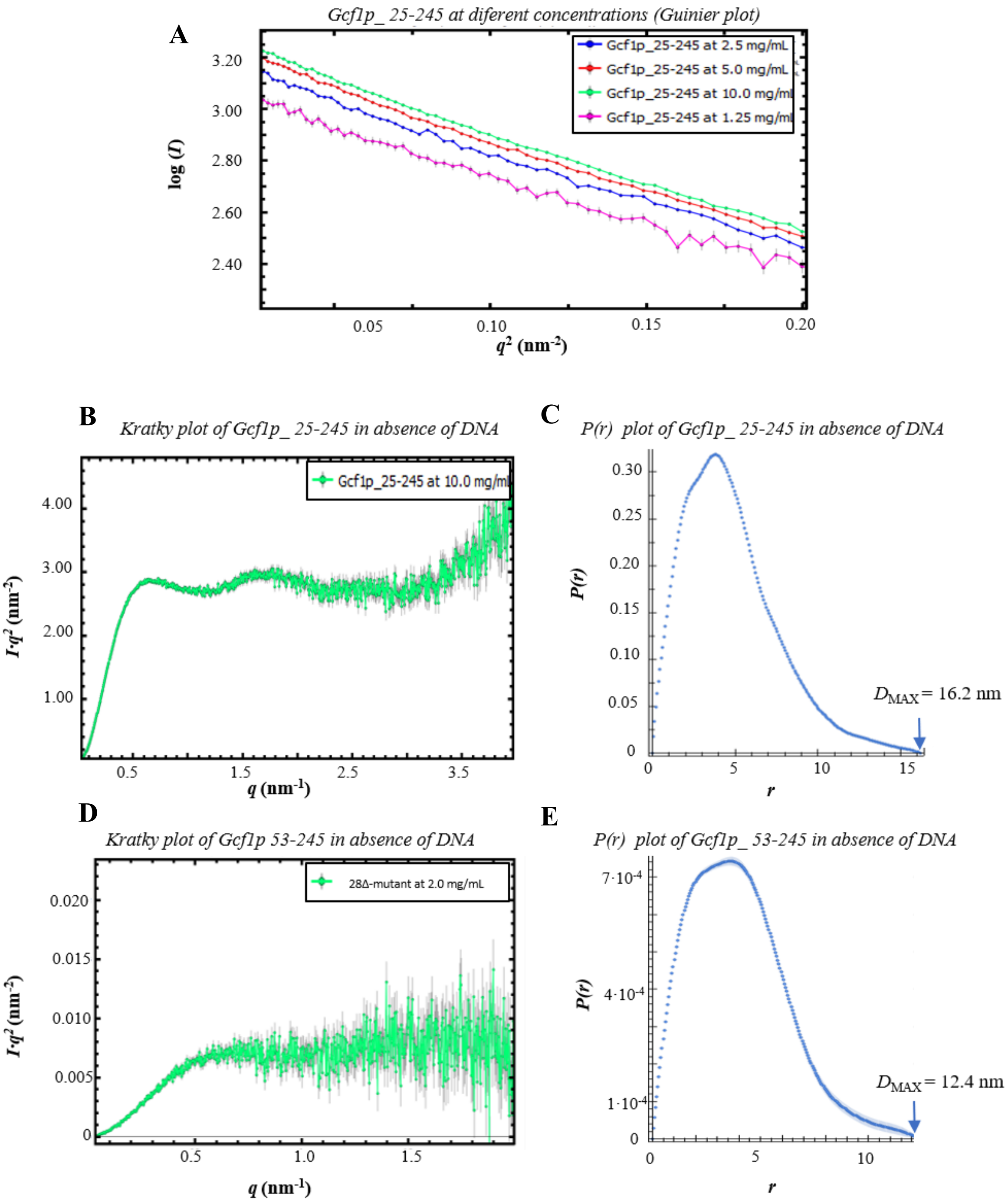


***Figure R28. SAXS data analysis of free Gcf1p. (A)**, the Guinier representation of Gcf1p full-length at different concentrations for q<0.5. **(B)** full-length and **(D)** Gcf1p53-245 mutant show features of a protein with flexible regions (Kratky plot) and/or with extended conformation (pair-wise distribution), respectively **(C)** and **(E)**.*

R5.2 Analysis of Gcf1p in solution in complex with DNA

In order to detect the effect of the DNA on the protein, protein/DNA complexes were analysed at the BioSAXS beamline at DESY (P12). In order to make a thorough characterization, five different available dsDNA substrates were assayed: Af2_20 (20bp), Y22 (22bp, [146]), Af2_21NoT (21 bp and absence of an A-tract), Atp950 (50 bp) and J3.12 (Holliday junction with 12 bp branches) (sequences are listed in **Table M2**). Both full-length Gcf1p and the Gcf1p53-245 mutant were assayed along with TFAM as a control since its characterization in solution by SAXS is known and available in the literature [39].

R5.2.1 Gcf1p compacts upon DNA binding

Scattering curves were measured for protein-DNA complexes as well as for free DNA. In order to establish the appropriate concentration of DNA for SAXS measurement (i.e., good signal-to-noise ratio and reduced inter-particle interaction) samples of Af2_20 at different concentrations by dilution in water were measured at 0.26 mg/mL, 1.30 mg/mL and 6.50 mg/mL. Amongst them, 1.30 mg/mL was selected as the concentration to be used for all the DNA samples mentioned above.

Protein:DNA complexes were prepared at 1:1.2 ratio following the previously explained three-step dyalisis procedure (see *SEC-MALLS sample preparation section,* **Materials and Methods 7.2**). Protein:DNA complexes were measured at concentrations of 0.25 mg/mL, 0.5 mg/mL, 1.0 mg/mL and 2.0 mg/mL. Only complexes formed with Af2_20 DNA substrates will be discussed in this subsection, parameters for all the complexes are listed in **Table R4**.

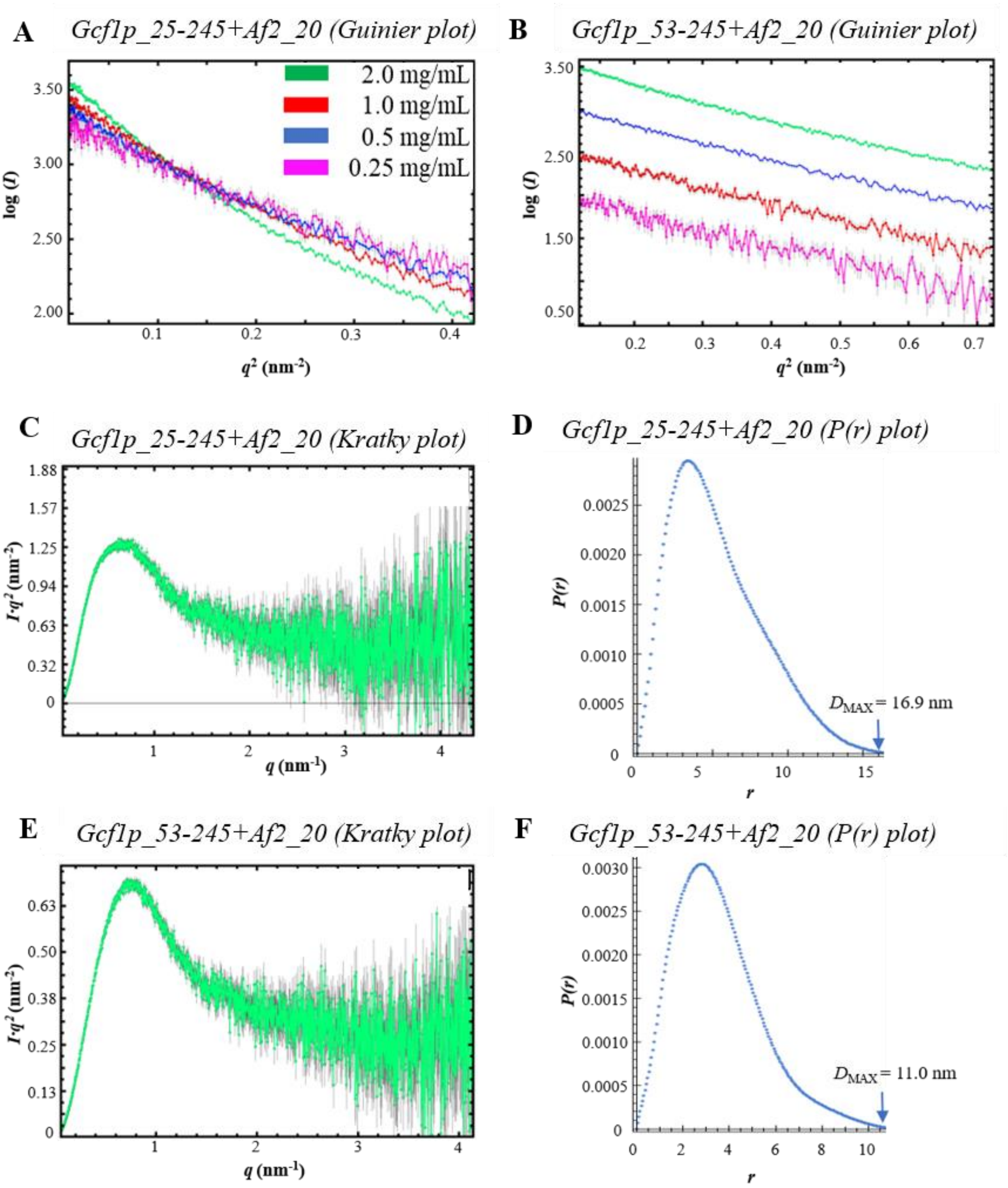


Figure R29. SAXS data analysis of Gcf1p bound to DNA. (A) *Guinier representations for the complexes Gcf1p full-length and (**B**) of the Gcf1p Gcf1p53-245 mutant at different concentrations showing a concentration-dependent change of curve slope only when the N-terminal tail is present. The bell-shape Kratky plot in both cases (**C** and **E**) shows a less flexible and more globular particle as compared to the same proteins in absence of DNA, and the P(r) plots (**D** and **F**) indicate that the complexes are more compact than the free protein.*

The complexes of Gcf1p25-245/Af2_20 showed a strong concentration dependence, evidenced by non-parallel Guinier regions as depicted in **Figure R29A** and notable increase in the molecular weight (**Table M4**). Consequently, extrapolation to zero concentration could not be performed. Molecular weight calculations for each concentration using the Porod volume approach $\left(\frac{Volume}{1.5} > MW > \frac{Volume}{2.0}\right)$ yielded values of 32885-43486 Da (at 0.25 mg/mL), 40550-54060 Da (0.5 mg/mL), 44514-59352 Da (1.0 mg/mL), 65254-87005 Da (2.0 mg/mL) MW ranges. Features of the Kratky plot and the $P_{(r)}$ plot revealed that the complexes had a more globular and less extended conformation than free Gcf1p, as shown in **Figure R29C** and **Figure R29E**.

In contrast with the native protein, the complex of Gcf1p53-245/Af2_20 did not show symptoms of concentration dependence in the measured range. Therefore, for this sample, combination of curves was performed as described for the case of free Gcf1p. A molecular weight range of 27927-37236 Da was calculated with the Porod volume approach. R_g was reduced from 3.76 to 2.88 nm upon DNA binding, indicating compaction. Both the bell-like Kratky plot and $P_{(r)}$ plot also showed features of a globular compact complex. Parameters of the protein/DNA complexes support the notion that DNA reduces the conformational space of the protein.

R5.2.2 The concentration-driven multimerization in presence of DNA depends on the disordered N-terminal tail.

Values for Rg, D_{MAX}, Porod volume and MW (calculated by Porod volume or Bayesan inference methods) of all the complexes assayed are summarized in **Table R4**. All the values shown in the table are directly extracted from single subtracted curves corresponding to different concentration values and denote that the concentration-dependent increments in the R_g occur for most of the samples. The discrimination between positive interparticle interaction or a true multimerization phenomena was performed based on visual inspection of the curve (multimerization causes differences over all the curve while inter-particle interaction causes differences only in the Guinier Region).

Complexes formed with the Gcf1p53-245 mutant showed a maximal MW of 34950 Da, which agreed with one protein and one DNA (37000 Da). In contrast, complexes formed with full-length Gcf1p25-245 showed a MW that systematically increased with concentration, up to 87052 Da in our range of concentrations, compatible with 2 proteins and 2 DNAs (74000 Da), 3 proteins and 1 DNA (90000 Da) or 2 proteins and 3DNA (88000Da) or, more probably a combination of these different stoichiometries. Our SAXS results show that the presence of the disordered N-terminal tail is essential for Gcf1p multimerization in complex with DNA at the concentration range tested in our assay (see **Table R4**).

Table R4: *Summary of particle parameters for all the complexes assayed. Correlation between Rg and concentration was detected in different cases. Calculation of molecular weight either by means of the Porod volume approximation (Porod MW) or by Bayesian Inference (Bayesian MW), as described in [145], were used to discern between positive inter-particle interaction and actual multimerization*

(mg/mL)	Gcf1p25-245 + Af2_20					Gcf1p53-245-mutant + Af2_20				
	Rg (nm)	*Dmax (nm)*	*Porod Vol (A^3)*	*Porod MW (Da)*	*Bayesian MW (Da)*	*Rg (nm)*	*Dmax (nm)*	*Porod Vol (A^3)*	*Porod MW (Da)*	*Bayesian MW (Da)*
0,25	3,19	13,6	65627	32813-43751	29250-32750	2,67	12,04	48259	24129-32172	31300-34950
0,5	3,75	16	82934	41467-55289	25250-34200	2,82	14,64	52554,8	26277-35036	33450-37300
1	3,81	18,3	89460	44730-59640	54950-64650	3,09	20	54510	27255-36340	27900-31300
2	4,33	17,1	130578	65289-87052	61600-69650	3,08	14,3	59042	29521-39361	32000-34950
	Concentration-dependent multimerization					**Positive inter-particle interaction**				
	Gcf1p25-245 + Af2_21 NoT					**Gcf1p53-245-mutant + Af2_21 NoT**				
	Rg (nm)	*Dmax (nm)*	*Porod Vol (A^3)*	*Porod MW (Da)*	*Bayesian MW (Da)*	*Rg (nm)*	*Dmax (nm)*	*Porod Vol (A^3)*	*Porod MW (Da)*	*Bayesian MW (Da)*
0,25	3,79	16,5	61411	30705-40940	23350-29250	2,47	7,98	46570	23285-31046	31300-34950
0,5	3,97	19	75730	37865-50486	34200-38100	2,65	10,2	44037	22018-29358	20250-24000
1	3,99	17,59	88739	44369-59159	54950-64650	3,02	14	50821	25410-33880	27900-30600
2	4,37	20,58	119542	59771-79694	52500-60200	3,33	15	58057	29028-38704	33450-37300
	Concentration-dependent multimerization					**Positive inter-particle interaction**				
	Gcf1p25-245 + Y22					**Gcf1p53-245-mutant + Y22**				
	Rg (nm)	*Dmax (nm)*	*Porod Vol (A^3)*	*Porod MW (Da)*	*Bayesian MW (Da)*	*Rg (nm)*	*Dmax (nm)*	*Porod Vol (A^3)*	*Porod MW (Da)*	*Bayesian MW (Da)*
0,25	3,16	12,55	54564	27282-36376	25250-29250	2,58	8	50400	25200-33600	32750-37300
0,5	3,24	14,88	55209	27604-36806	20250-23350	2,67	9,15	47545	23772-31696	25250-29250
1	3,66	19,05	64339	32169-42892	32750-38100	2,8	13	52161	26080-34774	33450-38950
2	3,81	15,5	74071	37035-49380	34200-39750	3,02	12,9	58190	29095-38793	32000-34950
	Concentration-dependent multimerization					**Positive inter-particle interaction**				
	Gcf1p25-245 + Atp950					**Gcf1p53-245-mutant + Atp950**				
	Rg (nm)	*Dmax (nm)*	*Porod Vol (A^3)*	*Porod MW (Da)*	*Bayesian MW (Da)*	*Rg (nm)*	*Dmax (nm)*	*Porod Vol (A^3)*	*Porod MW (Da)*	*Bayesian MW (Da)*
0,25	4,42	17,7	77707	38853-51804	40650-46150	4,4	15,58	67135	33567-44756	43300-50300
0,5	5,37	23,19	97204	48602-64802	NA	4,36	16,59	64502	32251-43001	43300-48200
1	4,92	21,5	99784	49892-66522	51450-60200	4,17	16,58	60382	30191-40254	40650-51450
2	4,65	24,4	89611	44805-59740	48200-54950	NA	NA	NA	NA	NA
	Concentration-dependent multimerization					**Negative Inter-particle interaction**				
	Gcf1p25-245 + J3					**Gcf1p53-245-mutant Gcf1p + J3**				
	Rg (nm)	*Dmax (nm)*	*Porod Vol (A^3)*	*Porod MW (Da)*	*Bayesian MW (Da)*	*Rg (nm)*	*Dmax (nm)*	*Porod Vol (A^3)*	*Porod MW (Da)*	*Bayesian MW (Da)*
0,25	2,95	NA	NA	NA	NA	2,97	9,17	78013	39006-52000	42400-47150
0,5	2,99	9,95	77937	38968-51958	45200-54950	3,02	10,3	83028	41514-55352	33450-37300
1	2,97	10	75144	37572-50096	43300-48200	2,96	9,43	78208	39104-52138	34950-39750
2	2,9	8,9	72511	36255-48340	43300-49250	2,87	8,63	71968	35984-47978	37300-43300
	Fix oligomerization State					**Fix oligomerization State**				

R5.3 Modelling Gcf1p structural heterogeneity in solution using SAXS data

R5.3.1 Modelling Gcf1p in absence of DNA using the Ensemble Optimization Method

The Ensemble Optimization Method (EOM [147]) allowed us to generate a library of Gcf1p models from which sub-ensembles of conformations were selected based on its agreement to the experimental data via genetic algorithms. Data used was a combination of two different curves taken from the same concentration series. As it has been explained before, the lower angle portion (q < 0.5 nm^{-1}) of the 1.25 mg/mL curve was combined with the higher angle portion (0.5 nm^{-1} < q < 5 nm^{-1}) of the 10.0 mg/mL curve.

In order to generate the ensemble of conformations, three domains were defined as rigid bodies, based on our crystal structure: HMG1, HMG2 and the 45-residue long N-terminal helix. Residue range 25-59, corresponding to the N-terminal tail, was created assigned as random coil. However, with this flexible region only, any ensemble was found to describe the SAXS experimental data ($\chi^2 = 1.497$). In a second trial, residue ranges 106-114 and 152-170, corresponding to the extended regions of HMG1 and HMG2 respectively, were additionally assigned as flexible, as well as the residue range 238-245 corresponding to the 7 residue-long C-terminal tail. EOM selected a sub-ensemble with an overall fit of χ^2 = 1.033 to the experimental data (**Figure R30A**). EOM produced a sub-ensemble of four models consistent with the experimental scattering data (**Figure R30C**). These structures did not represent all conformations adopted by the protein in solution but overall provided an intuitive idea of the conformational space explored by the protein. The distributions of R_g values from the initial (blue curve) and selected (red curve) conformational pools are outputted by EOM, as depicted in **Figure R30B**. The R_g values of the initial and selected conformational ensembles are 3,85 nm and 3,63 nm, respectively. The selected sub-ensemble shows a distribution of distances with a maximum at 3,0 nm, thus more compact than the generated conformational pool.

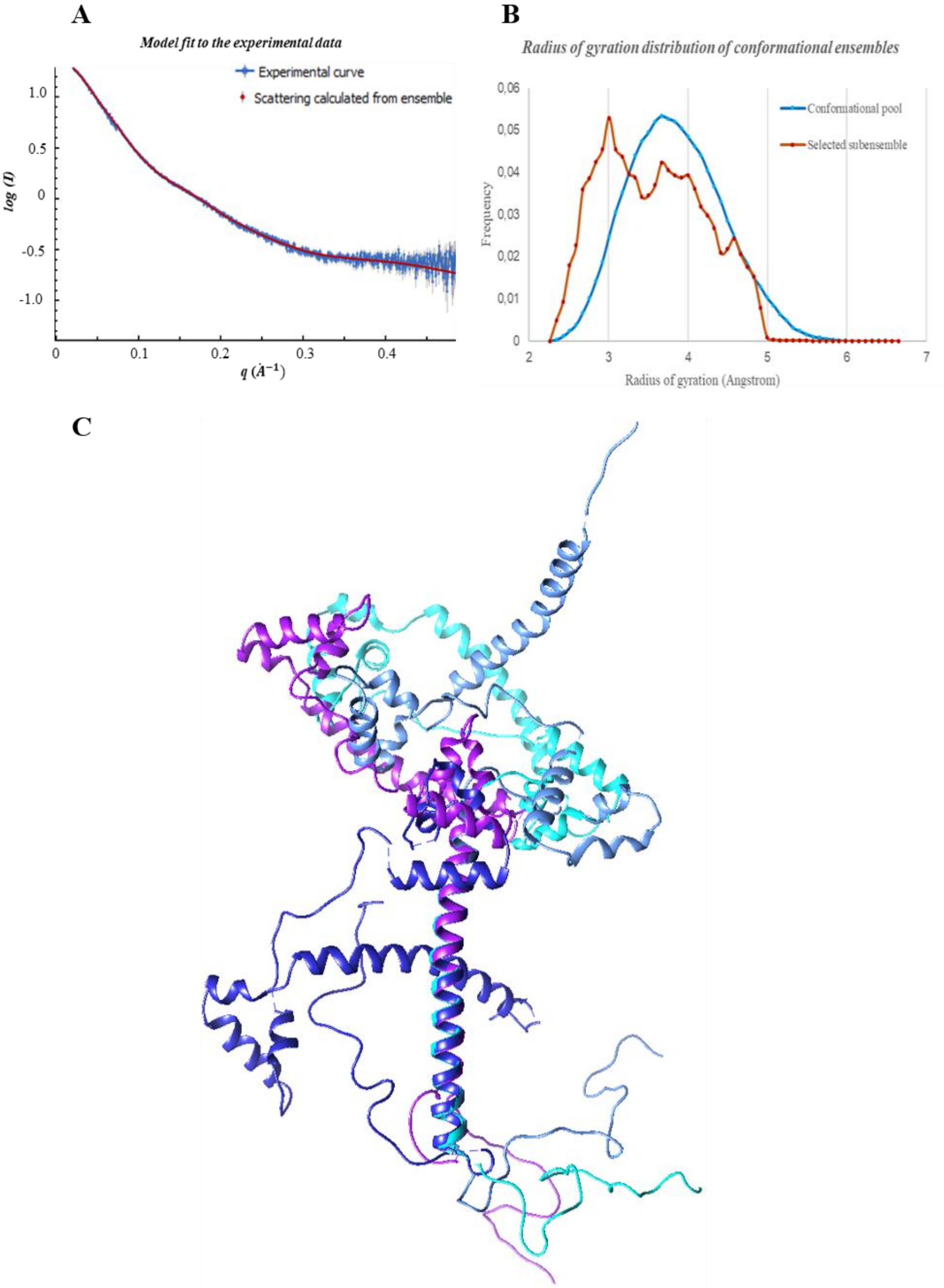


Figure R30. Modeling Gcf1p flexibility in absence of DNA. *(**A**) fit of the selected sub-ensemble to the experimental data, (**B**) Radius of gyration distribution for the conformational pool (blue) and for the selected sub-ensemble (red). (**C**) ensemble of structures representative for Gcf1p different conformations in solution in absence of DNA.*

R5.3.2 Modelling Gcf1p complexes DNA using computed scattering factors

CRYSOL (Svergun, Barberato, & Koch, 1995) is a program to fit macromolecular structures to experimental SAXS curves. Therefore, it is a powerful tool to validate structural models obtained from crystal structures.

Different models were prepared exploring all the possible protein and DNA combinations up to a maximum 4:4 protein:DNA complex as observed in our crystal structure. Data from the Gcf1p 25-245 + Af2_20 complexes at each concentration was used for fitting. Due to its concentration-independent behaviour, for 28Δ-mutant + Af2_20 complexes only the curve at the maximum concentration, XX mg/mL, was used.

Significantly good fit (χ^2 = 1.099) was found between the Gcf1p25-245/Af2_20 data and the 1:1 protein:DNA model in which the DNA is bound by the non-canonical HMG-box 1 (**Figure R31A**). Flexible refinement of the high resolution model using normal mode analysis with SREFLEX (Panjkovic & Svergun, 2016) improved the fit to the data (χ^2 = 0.999) (**Figure R31B**). Comparison between structures before and after normal mode refinement revealed small discrepancies in the conformation of HMG box 2 which appears to be more flexible and closer to the DNA in the refined structure (**Figure R31C**). Moreover, the DNA shows a less sharp bend in the refined structure (**Figure R31C**). SREFLEX also produced other 8 refined structures in good agreement with the data ($1.00< \chi^2 < 1.05$). Such structures, superimposed in **Figure R31D**, strongly suggest that 1:1 protein DNA complexes keep a certain degree of flexibility.

Deconvolution of the scattering curves measured for the Gcf1p25-245/ Af2_20 complex was more difficult, not only due to the presence of the disordered N-terminal tail but also to different oligomerization states probably coexisting at each concentration assayed (suggested by the MW increase at increasing concentrations, see **Table R4**). Curves at 0.25 mg/mL and 0.5 mg/mL could not be used due to their poor signal to noise ratio at high angles. Protein models showed poor fit to the curve at 1.0 mg/mL ($\chi^2 \geq 3.550$). Curves were better explained as a combination of the calculated scattering from different models. OLIGOMER software was used to fit a combination of different calculated form factors to the experimental curve (Konarev P. V., Volkov, Sokolova, Koch, & Svergun, 2003) yielding an overall fit to the curve of $\chi^2 = 1.430$ (**Figure R31A**). OLIGOMER revealed three major species in solution at different relative concentrations (**Figure R31A**). The models selected for the curve at 1.0 mg/mL have a molecular weight of 52000 Da and 88000 Da, discordant with the 47300-59640 Da range calculated from the raw data. Nevertheless, if a weighted average value is calculated using the relative proportion of each stoichiometry calculated by OLIGOMER, a value of 54880 Da is obtained. Curves at 2.0 mg/mL were also modelled using OLIGOMER and a reasonably good fit was obtained (χ^2 = 1.506) as illustrated in **Figure R32B**. OLIGOMER-selected stoichiometries for the 2.0 mg/mL curve are of higher order than those selected for the 1.0 mg/mL curve, and are consistent with a concentration-dependent multimerization mediated by coiled-coil formation through the N-terminal α-helix of Gcf1p (**Figure R32C**). The weighted MW average of complexes selected by OLIGOMER (83530 Da) is consistent with the MW value (65000-87000) calculated from the corresponding experimental curve at 2 mg/ml (**Table R4**). In both cases, computed scattering curves used for Oligomer analysis were obtained using the WAXSiS server (Chen & Hub, 2014), (Knight & Hub, 2015) which performs scattering curves based on explicit-solvent all-atom molecular dynamics simulations.

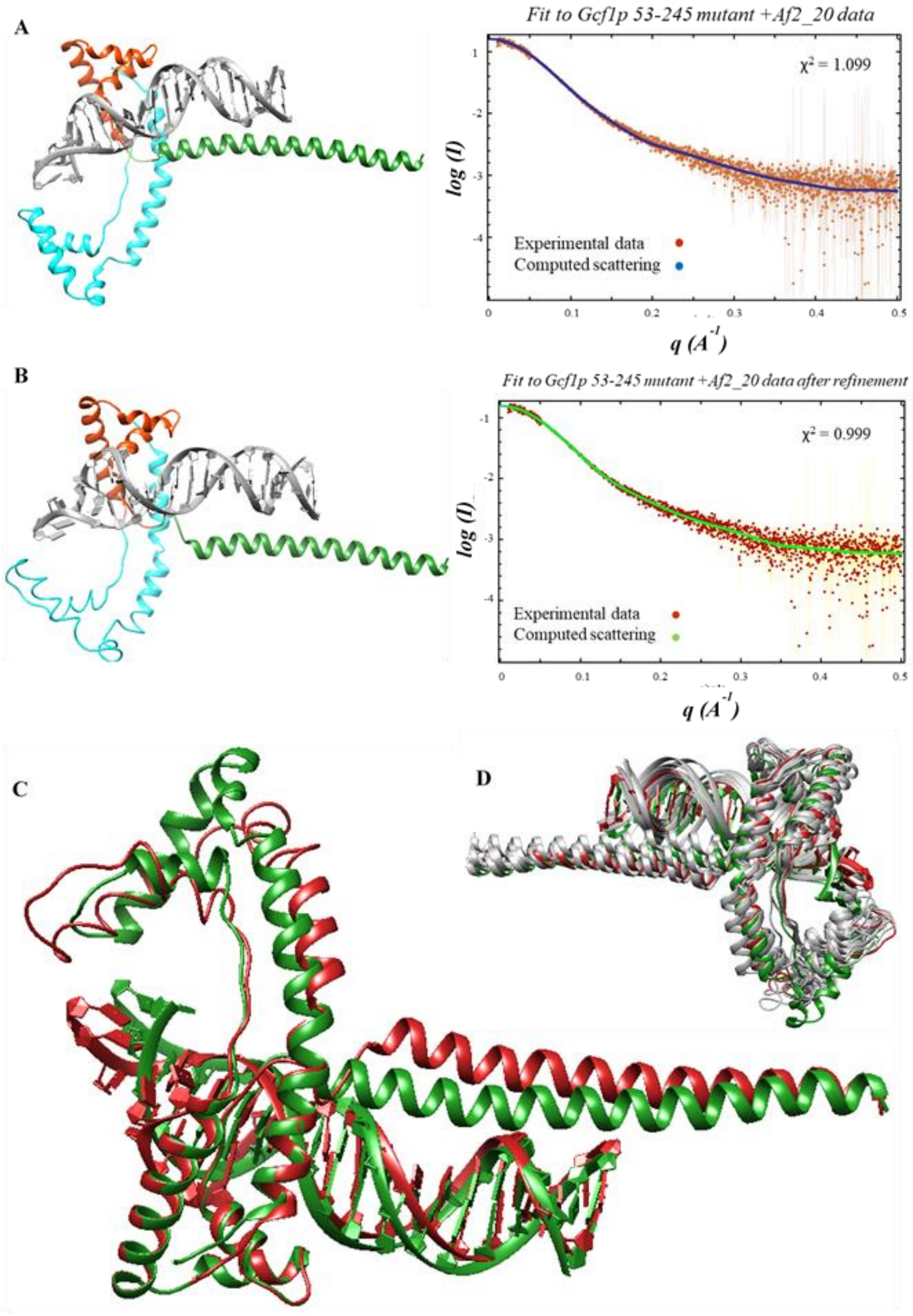


Figure R31. Modelling Gcf1p53-245 mutant+Af2_20 SAXS curves using computed form-factors from high-resolution crystal structures. *(**A**) Fit to data of HMG-1 bound DNA 1:1 Gcf1p+Af2_20 structure performed with CRYSOL). (**B**) Best fit to data of the models generated by normal mode refinement with SREFLEX. (**C**) Superposition of the structure before (in green, χ^2=1.099) and after normal mode refinement (in red, χ^2=0.999). (**D**) Superposition of both structures shown in (**C**) together with 8 alternative structures generated by SREFLEX (in gray, 1.05>χ^2>1.00).*

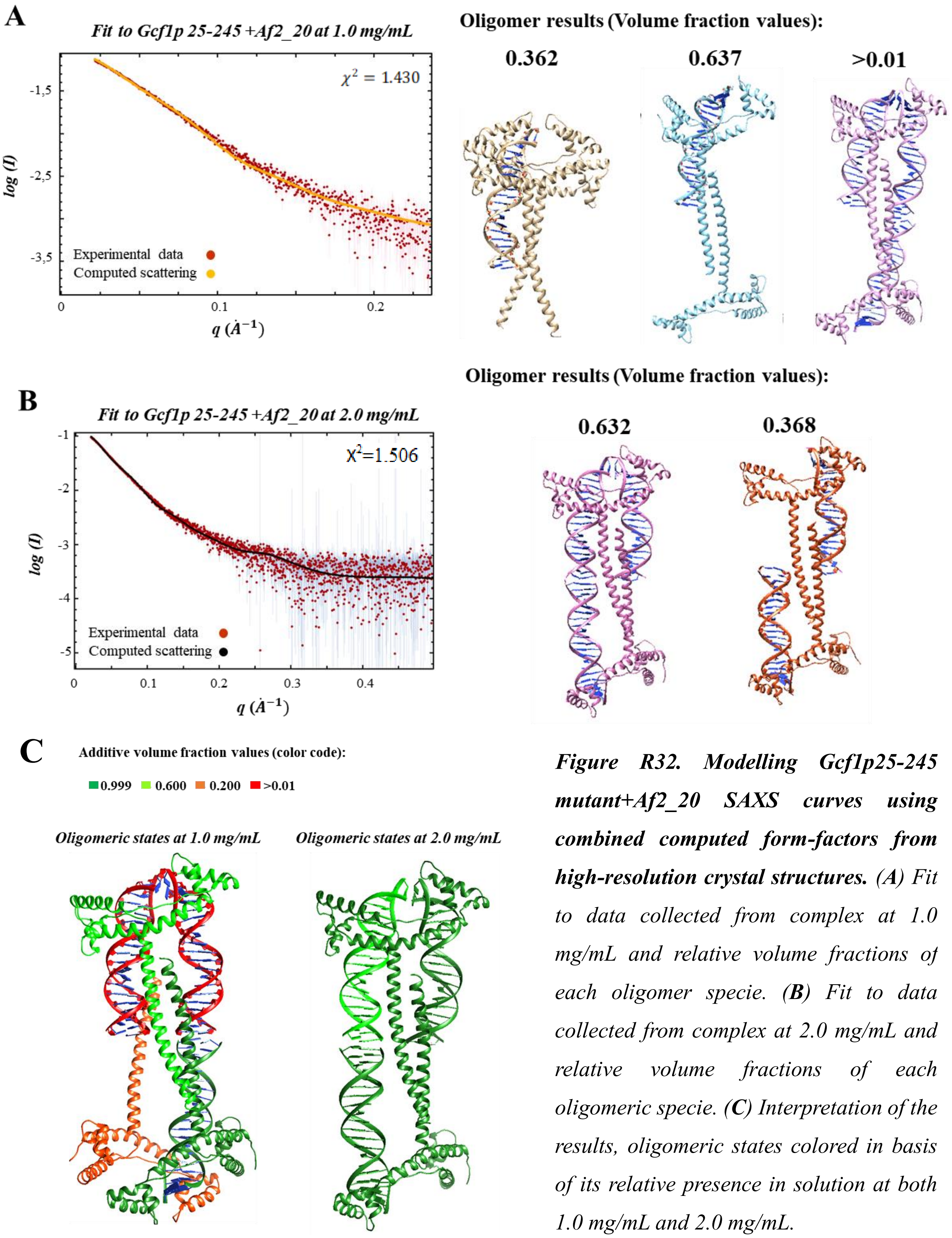


Figure R32. Modelling Gcf1p25-245 mutant+Af2_20 SAXS curves using combined computed form-factors from high-resolution crystal structures. *(A) Fit to data collected from complex at 1.0 mg/mL and relative volume fractions of each oligomer specie. (**B**) Fit to data collected from complex at 2.0 mg/mL and relative volume fractions of each oligomeric specie. (**C**) Interpretation of the results, oligomeric states colored in basis of its relative presence in solution at both 1.0 mg/mL and 2.0 mg/mL.*

Our SAXS results strongly suggest that the interactions present in our crystal also occur in solution. These include the coiled coil interaction between the N-terminal helices, the DNA binding through HMG-box 1 and the DNA binding through HMG-box 2. They also suggest an interaction between Gcf1p and Af2_20 substrates in which the "tetramer" complex (see **Figure R25**) ensembles gradually in a concentration dependent manner. Proteins would bind first to DNA through HMG1 and, to a minor extent, through HMG2. Once 1:1 protein:DNA complexes are formed, homodimers start to form mediated by the inter-protein coiled coil. This process is hampered when the disordered positively charged N-terminal tail is not present, and we did not observe assemblies of more than one protein and one DNA in the assayed concentration range for the Gcf1p 53-245 mutant. The disordered N-terminal tail, which is not traced in our crystal, is nonetheless involved in the early stages of complex formation. The disordered N-terminal tail is rich in lysine residues (12 out of a total of 33 residues) and therefore positively charged. We propose that this positive tail can establish interactions with both DNA and protein and would bring close in space two independent 1:1 protein:DNA complexes that would then form a stable multimeric assembly mediated by the coiled-coil region. Nevertheless, it is possible that the Gcf1p53-245 mutant would form stable multimeric assemblies at concentrations of complex higher than the ones of the assay. In consequence, we cannot affirm that the N-terminal tail is needed for the proper formation of the coiled coil, albeit its presence is favouring it.

5.3.3 Ab initio models from SAXS data

Prior to modelling using our crystal structure, dummy residues models were generated from our SAXS data. This approach is much less relevant than the computed curves from a high-resolution crystal structure by CRYSOL, nonetheless it is a visual way to represent low resolution envelopes and it was used to model some of the non-crystallized complexes from which a high resolution model was not available: Gcf153-245 + Y22, Gcf1p25-245 + Y22, Gcf1p25-245 + Atp950 and Gcf1p + J3. Models were built using 4.20 *Å* radius dummy atoms in all cases, for each curve 10 independent *ab initio* models were generated, refined with Dammin and averaged using Damaver. Fitting to the experimental curves yield adjustments of χ^2 = 1.072, χ^2 = 1.007, χ^2 = 1.050 and χ^2 = 1.130. Overall shape determination by this procedure outputs the low-resolution envelopes summarized in **Figure R33**. Overall dimensions of the Gcf1p53-245 + Y22 complex are consistent with a 1:1 protein DNA complex whereas those of the Gcf1p25-245 + Y22 are consistent with the formation of inter-protein coiled-coil. Overall dimensions of the Gcf1p25-245 + Atp950 envelope, were also consistent with our crystal structure, suggesting that Gcf1p assembles in a similar manner when bound to DNA substrates of a length in the 20 – 50 bp range. Low-resolution envelopes of the Gcf1p in complex with the J3 DNA junctions evidence a rather flat, compact particle, without the protuberances that are observed in the other envelope corresponding to a 1:1 particle, i.e. that computed from the 28Δ-mutant + Y22 curve. This different overall shape, added to the fact that this complex does not multimerize in the assayed concentration range even in presence of the N-terminal tail, suggests that Gcf1p adopts when bound to this substrate, a different conformation than the one observed in our crystal structure, i.e. bound to linear 20 bp DNA substrate.

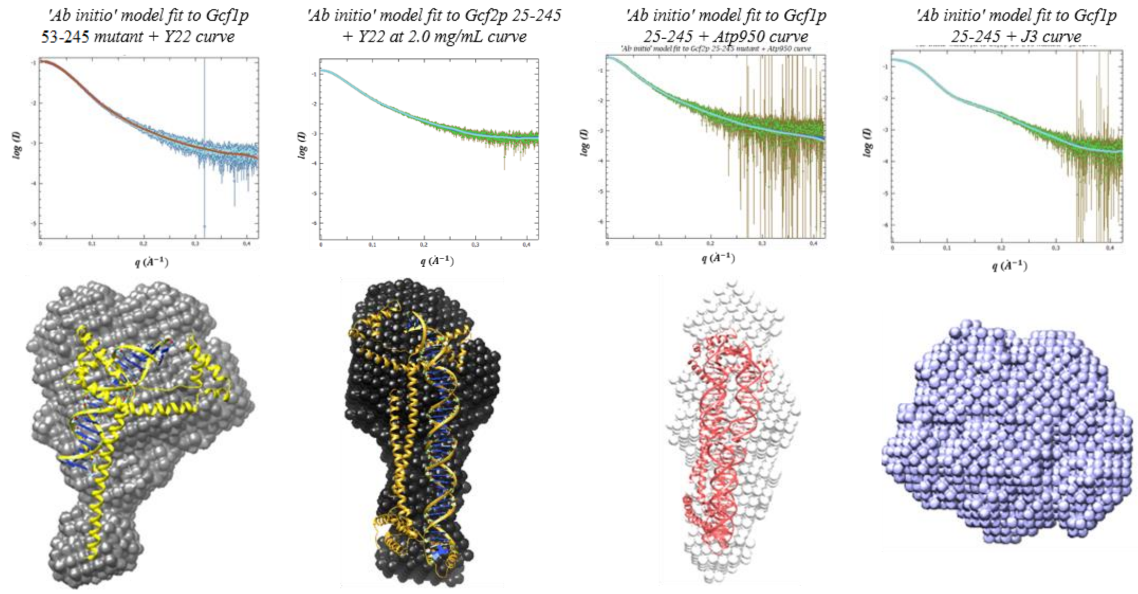


Figure R33: *Representation of the low-resolution 'ab initio' models for complexes without a high-resolution model available. From left to right correspond to: Gcf1p Δ28 mutant + Y22, Gcf1p 25-245 + Y22, Gcf1p 25-245 + Atp950 and Gcf1p 25-245 + J3. Fitting to each of the independent models to the data is displayed on top of each averaged model. Bead models are calculated at 5 Angstrom resolution. The high-resolution models are just for representative means, no fitting of the high-resolution structure to the bead model was performed in any case.*

R6. Electron microscopy of Gcf1p bound to long DNA fragments

Gcf1p, as the major organizing protein of *Candida albicans* mitochondrial genome (mtDNA), interacts with DNA molecules longer than 40000 bp. DNAs of such a length are not manageable for structural analysis by macromolecular crystallography nor by SAXS (see previous sections). In order to analyse the effects of Gcf1p binding to DNA molecules longer than 500 bp, transmission electron microscopy was performed.

R6.1 Gcf1p binding to long DNA substrates

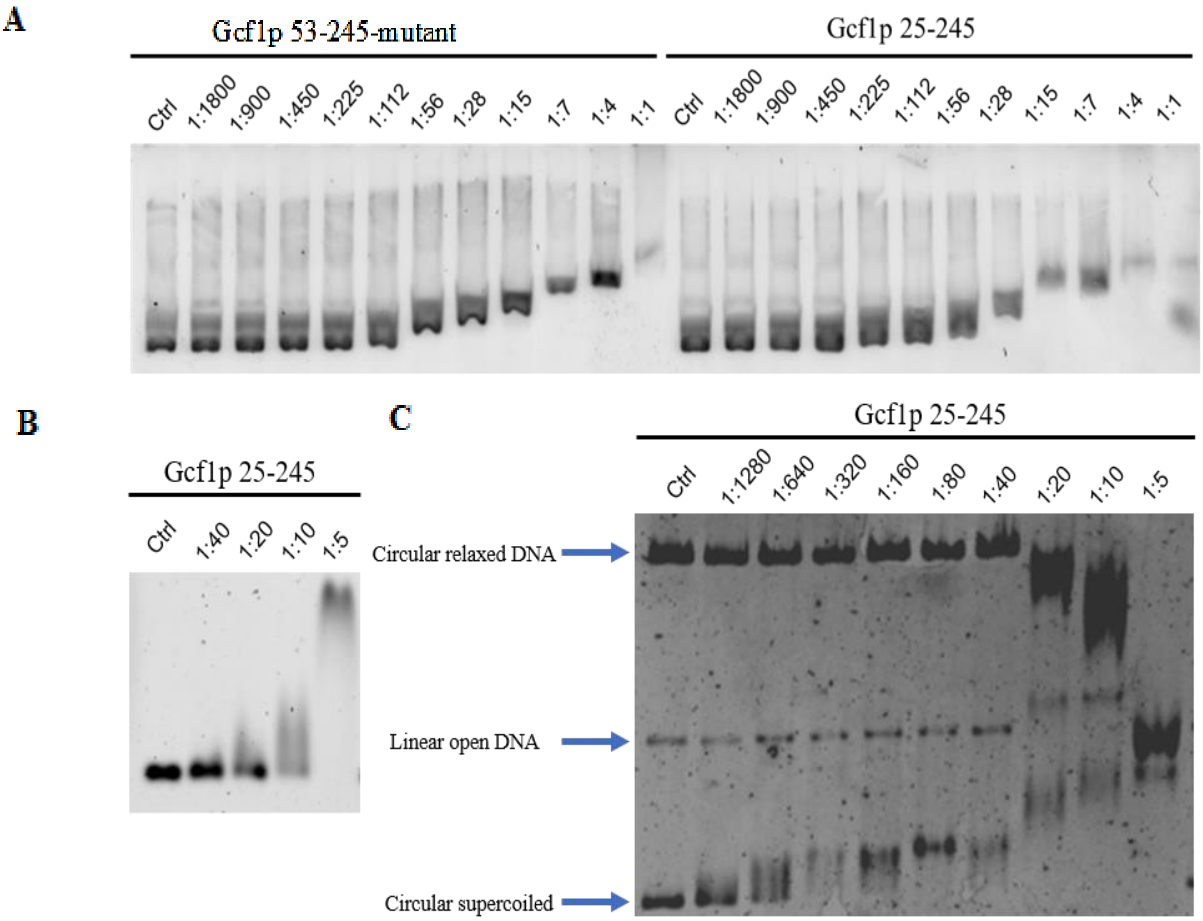


Figure R34. Gcf1p binding to long DNA substrates. *(**A**) native agarose gel (1%) with increasing amounts of either Gcf1p 25-245 or Gcf1p53-245 mutant, incubated with constant amounts of supercoiled DNA. (**B**) binding of DNA to the linear PBR1000 (1000 bp). (**C**) EMSA performed with a heterogeneous sample of PBR322 plasmid that contained three different topologies: circular relaxed, open linear and circular supercoiled. The ratio of protein molecules per DNA base-pair is indicated above.*

Gcf1p binding to longer DNA substrates was tested by EMSA. The chosen substrate was PBRC7, a 1131 bp DNA fragment amplified from the PBR322 plasmid between positions 2576 and 3707 with specific DNA oligos

(**Table M3)**. Protein at increasing concentrations was incubated with PBR1000 substrate at a constant concentration of 1 ng/µL, for 30' at room temperature as indicated in **Materials and Methods**, **M5.3** section. Samples were loaded in a native agarose gel and stained with SYBR Safe 1X. The results show a progressive electrophoretic retardation of the DNA at increasing protein concentrations, evidencing binding of both Gcf1p25-245 and Gcf1p53-245-mutant to the DNA, see **Figure R34**.

R6.2 Optimization of sample preparation for EM studies

R6.2.1 Optimization of protein-free DNA samples

In order to characterize the effect of protein binding on the DNA conformation, images from negative controls, with only DNA and without any protein, were taken for reference. The electron microscopy images of a good DNA control display specific features that are going to be compared with the sample containing the protein. Good quality DNA EM samples are characterized by the presence of single DNA molecules of homogeneous length, absence of aggregation and a relaxed 'worm-like chain' conformation when deposited onto the functionalized carbon films.

The initial trials yielded images that did not show such characteristics, so that optimization of the DNA substrate production underwent as follows (*see optimization of the DNA substrates for EM section,* **Materials and Methods 6.2.1**). Phusion DNA polymerase was chosen over Taq and Herculase, as it yielded a more homogeneous DNA sample as illustrated **Figure R35A**. DNA purification by ethanol precipitation was not enough to remove impurities and, in addition, aggregates were visible in the first micrographies (**Figure R35C**). Both impurities and aggregates in PCR products were eliminated by anionic exchange purification using a MiniQ^{TM} column (GE-Healthcare®) (**Figure R35B**). Following anionic exchange chromatography10X dilution in MilliQ water, thus decreasing salt concentration to 65 mM NaCl. Following this protocol, we could obtain a control DNA sample suitable for structural analysis by EM (**Figure R35D**). Other DNA substrates used in our EM studies (pUC19-linear, pUC19-relaxed and pUC19-supercoiled) were prepared by our collaborator Sonia Baconnais from Prof. Éric le Cam laboratory at the Institut Gustave Roussy (Villejuif).

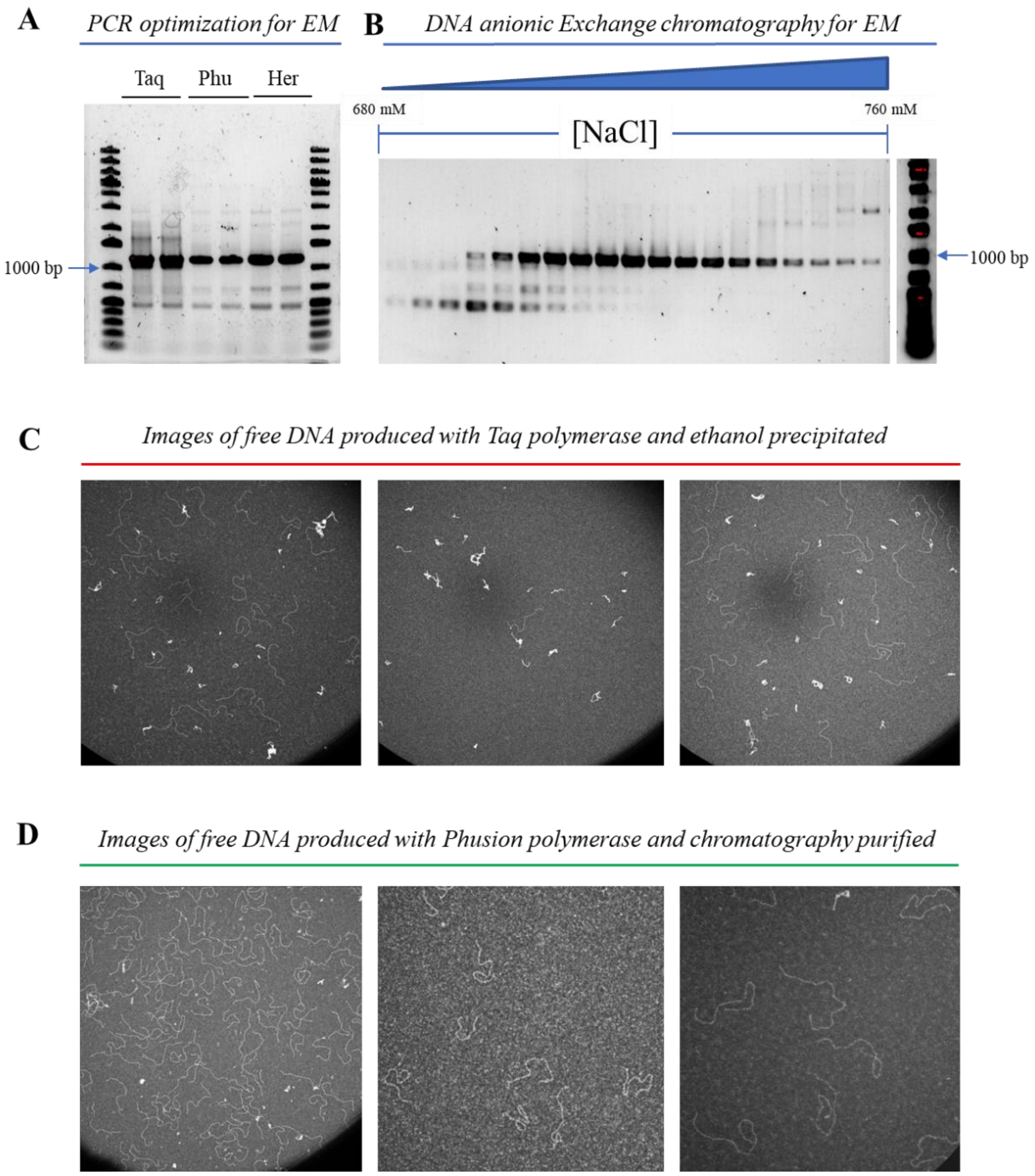

A
PCR optimization for EM
Taq
Phu
Her
1000 bp
B
DNA anionic Exchange chromatography for EM
680 mM
[NaCl]
760 mM
1000 bp
C
Images of free DNA produced with Taq polymerase and ethanol precipitated
D
Images of free DNA produced with Phusion polymerase and chromatography purified

Figure R35. Optimization of DNA samples for EM studies with Gcf1p. ***(A)*** *Gel electrophoresis with EM substrates (pBR_C7 product, see* ***Table M3****) produced with polymerases, Taq, Phu and Herculase (Her). The marker for 1000 bp is indicated.* ***(B)*** *Anionic exchange chromatography samples loaded on an agarose gel (1%).* ***(C)*** *EM micrographies of the free DNA produced following the original protocol based on ethanol precipitation.* ***(D)*** *Free images of the DNA after protocol optimization.*

R6.2.2 Optimization of protein/DNA samples

Our first images of protein/DNA complexes showed that the samples had a strong tendency to form extensive protein/DNA aggregates (**Figure R36A**). Thus, after incubation of the protein/DNA complexes, a gel filtration was performed using a Superose 6 ™ (GE-Healthcare®) column, yielding images with lower background noise and reduced presence of aggregates. This step was necessary only for the complexes formed with the linear and relaxed DNA substrates, whereas the supercoiled DNA samples showed a much inferior presence of aggregates. At a further step, we discovered that the addition of supercoiled DNA significantly reduced the presence of aggregates in protein-DNA mixtures containing linear DNA. Alternatively, the use of centrifugal filters also reduced the absence of aggregates (*see optimization of protein-DNA complexes for EM section, **M6.2.2***). Images of analysable protein-DNA complexes are illustrated in **Figure R36B**.

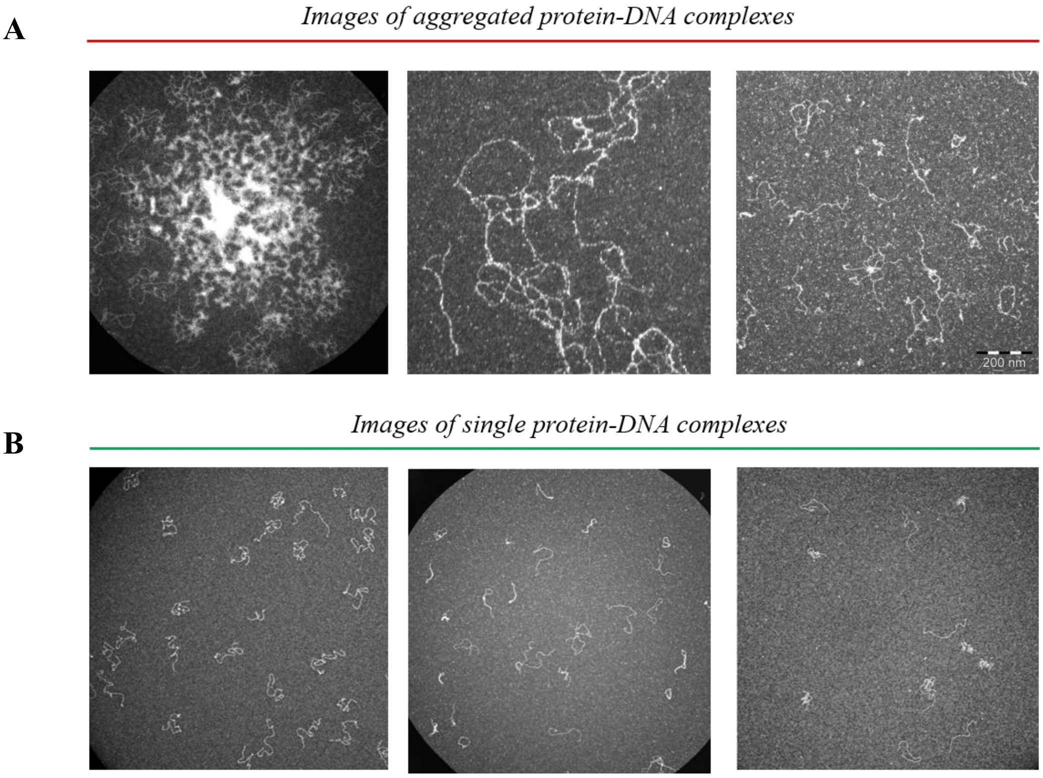

A
Images of aggregated protein-DNA complexes
200 nm
B
Images of single protein-DNA complexes

Figure R36. Optimization of protein-DNA samples. *(A) Examples of EM image fields that are not usable for structural analysis. (**B**) Images of usable images of protein-DNA complexes. From left to right aggregates were decreased by centrifugation with 30kDa MW Cut-off centrifugal filters, by addition of supercoiled DNA -which appears as white and extremely compacted particles- and by gel filtration.*

R6.3 Gcf1p compacts DNA following a bridging mechanism

Gcf1p induces local distortions on the DNA structure that are not present on the free DNA images (**Figure R37A** image 2). Furthermore, cross-strand binding and DNA wrapping events are observable upon Gcf1p binding (**Figure R37A** images 3 and 4). As compared to the free DNA, Gcf1p-bound DNA molecules show a more compacted structure (**Figure R37A** image 1). Protein-covered compacted DNA molecules coexist with naked DNA molecules (**Figure R37A** image 5), suggesting that Gcf1p binds to previously bound DNA molecules with more affinity than to naked DNA, i.e. Gcf1p shows cooperativity of binding. Protein-DNA complexes observed by electron microscopy showed a significant number of hairpin-like structures of different length that were not present in the micrographies of free DNA (**Figure R37B**). These structures showed that the presence of protein brought independent DNA stretches close in space, both in parallel and anti-parallel arrangements.

Such features observed for Gcf1p-DNA complexes are characteristic of DNA bridging, a DNA compaction mechanism observed in bacterial nucleoids of different species [153], [154], [155]. These protein-mediated DNA bridging events require the polymerization of protein on the DNA and the ability to generate cross-strand binding events. Finally, accumulation of bridging events can involve more than two DNA segments, thus inducing DNA compaction (**Figure R37C**).

In addition to Gcf1p 25-245, protein-DNA complexes were formed using the Gcf1p 53-245 mutant. TFAM, used as a control, also induced DNA bridging as previously described in the literature [90]. This suggests that the N-terminal coiled coil between proteins detected in our crystal structure is not required for DNA bridging and suggests that HMG-box binding would be enough to compact DNA. Our results do not allow us to perform the proper statistical analysis to determine whether the N-terminal tail increases or decreases the prevalence of this phenomena.

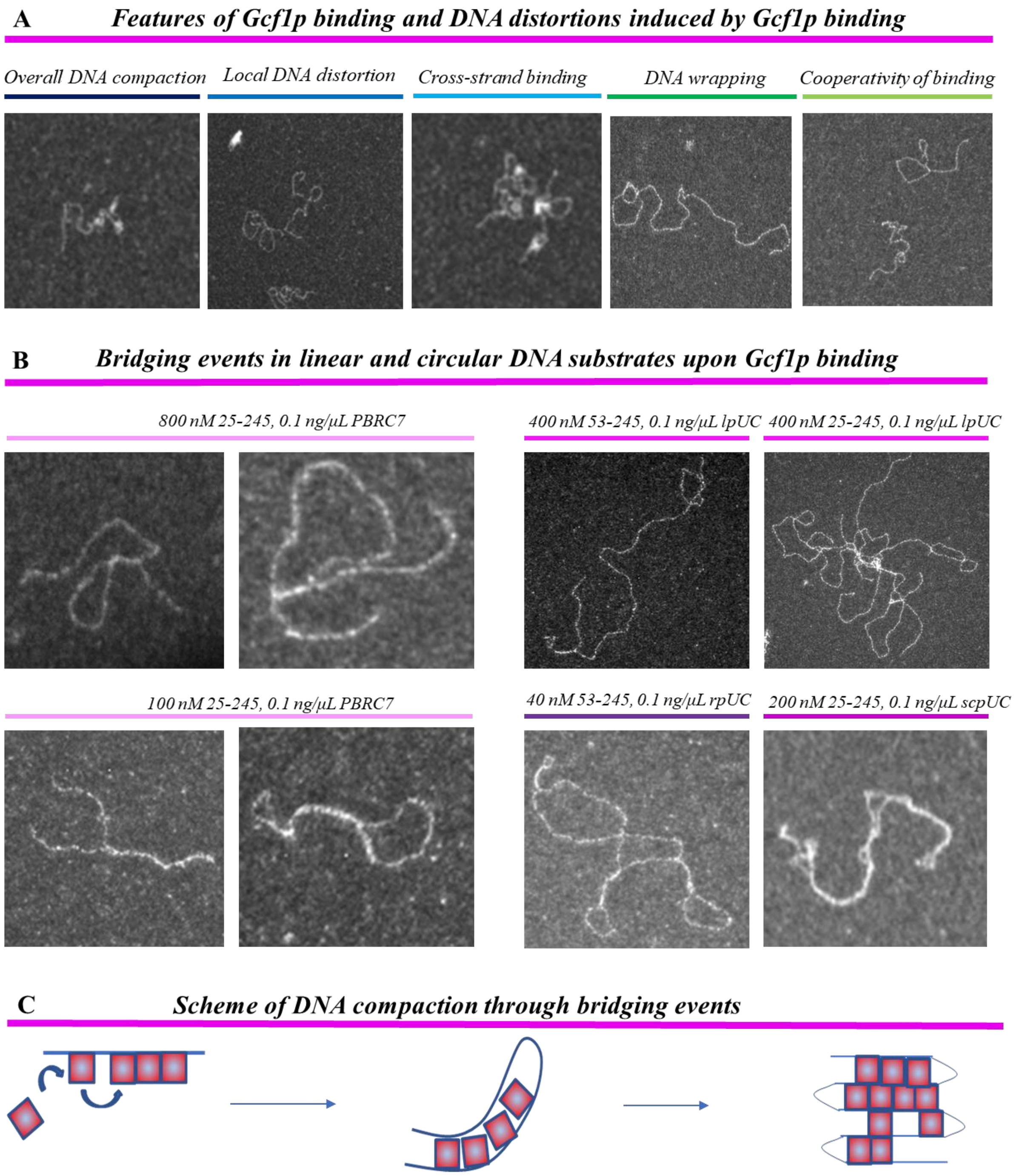

A
Features of Gcf1p binding and DNA distortions induced by Gcf1p binding
Overall DNA compaction
Local DNA distortion
Cross-strand binding
DNA wrapping
Cooperativity of binding
B
Bridging events in linear and circular DNA substrates upon Gcf1p binding
800 nM 25-245, 0.1 ng/μL PBRC7
400 nM 53-245, 0.1 ng/μL lpUC
400 nM 25-245, 0.1 ng/μL lpUC
100 nM 25-245, 0.1 ng/μL PBRC7
40 nM 53-245, 0.1 ng/μL rpUC
200 nM 25-245, 0.1 ng/μL scpUC
C
Scheme of DNA compaction through bridging events

*Figure R37. **Overall features of Gcf1p/DNA complexes observed in by EM.** (A) Features induced to DNA due to Gcf1p binding. **(B)** Evidences of DNA bridging at different protein concentrations. Different DNA substrates assayed highlight that Gcf1p induces DNA bridging regardless of sequence, length or topology. (**C**) General scheme for protein-mediated DNA bridging. From left to right, cooperative binding and protein polymerization on the DNA. The ability of the protein to induce local distortions and cross-strand binding allows the formation of double-stranded DNA 'hairpins' resulting in intra-molecule bridging. In a final step bridging events can involve more than two DNA segment. This can result in more compacted nucleoprotein particles as observed in some EM images.*

DISCUSSION

Gcf1p crystal structure presents striking differences when compared to previously available structures of mitochondrial HMG-box proteins [39], [40], [42], [157], [41]. Starting from N-terminal, Gcf1p presents a long N-terminal helix of 45 residues that is not present in TFAM [39], [40], [157], [41] nor in Abf2p [42]. Abf2p presents a short N-terminal helix of 10 residues, it could be argued that both helices are necessary for a function essential in yeast mitochondrial DNA. Nevertheless, its actual role in the crystal structure is completely different, whilst N-terminal helix of Gcf1p is involved in protein dimerization, that of Abf2p is involved in the formation of a hydrophobic core that stabilizes protein-DNA complex. This last function is performed in Gcf1p by the last 15 residues of HMG-box 2, a feature which is also not present in the previously available structures. Another particularity of Gcf1p HMG-box 2 is the presence of a potentially flexible point within Helix 2 that we have named as 'ankle' (**Figure R16**). This 'ankle' could be responsible for a certain degree of freedom when bound to DNA in solution, indeed our results show how HMG-box 2 subtly unfolds when DNA is bound by HMG-box 1 (see **Figure R31**). The most striking feature of Gcf1p is a presence of an HMG-box domain (HMG-box 1) that binds DNA by the major groove, which to our knowledge is reported for the first time. Such a particularity can be due a different mechanism for mtDNA compaction in *Candida* genus in relation to *Saccharomyces*. It is also the first time that a coiled-coil dimerization surface is reported in a mitochondrial HMG-box crystal structure. Such particularities together provide a new target for biomedical applications. Given that the coiled-coil dimerization surface is not presence in human TFAM, a drug that would target such interaction could affect *Candida albicans* viability without impairing TFAM function. Nevertheless, a thorough *in vivo* analysis must be done in order to characterize the role of such a coiled coil in *Candida albicans* viability. Besides the impact our findings may have in biomedicine, our results can open a new paradigm in yeast mtDNA maintenance. Sequence analysis has identified that such an HMG-box protein with a coiled coil forming domain and with a shortened or absent HMG box 1 is not uncommon in yeast (**figure I19A**). One hypothesis could be that the evolution to a smaller HMG-box 1 unable to bend the DNA somehow is compensated by the acquisition of this dimerization surface that allows to have to HMG box 2 domains in tandem.

The crystal structure of Gcf1p bound to DNA shows the formation of an imperfect (or distorted) U-turn. A perfect U-turn was observed in other mtDNA compacting proteins, yet Gcf1p induces it by a unique mechanism. Human TFAM/DNA complexes [39], [40], [85], [146] show intertwining between both protein and DNA, which is facilitated by the long flexible linker between HMG-boxes. Instead, Gcf1p contacts the DNA U-turn by one side, thus the intertwining between protein and DNA observed for TFAM might be specific for mtDNA compaction in mammals, as suggested but the high homology of TFAM in this taxonomic class (Rubio-Cosials, NSMB 2011). Abf2p, instead, as for Gcf1p contacts one side of the DNA U-turn by its two HMG-boxes, but both boxes contact the minor groove. Gcf1p is also distinct in this, since HMG-box 1 contacts the DNA major groove and the U-turn depends on the contact of two Gcf1p proteins simultaneously. As it has been mentioned before, the contacts between HMG boxes 2 belonging to the two crystallized chain seem spurious and resulting from the crystal packaging forces. Therefore, they are susceptible to be disrupted thus allowing for the allocation of extra bases between the two DNAs leading to an eventual formation of a U-turn. This would necessarily imply additional flexibility of the protein, in order to change the arrangement between HMG-box 1 and HMG-box 2. Experiments in solution show that Gcf1p is a flexible protein in solution that rather adapts to the DNA conformation, therefore it is not unfeasible that conformational changes in the protein allow for the formation of a U-turn. Such hypothesis would need further experimental evidence, for instance by single-molecule FRET [85].

TFAM, shows a progressive binding of the protein domains to the DNA, so that the protein wraps the nucleic acid and bends it [39], [85]. Whereas such progressive binding is no required for proteins binding to one side of the DNA (as for Gcf1p and Abf2p), it is noteworthy that in both TFAM and Abf2p the binding is always led by HMG-box 1, whilst HMG-box 2 has very low affinity. In both TFAM and Abf2p the DNA binding promotes the formation of new interactions between protein domains. Thus, it is reasonable to expect that the interactions between protein regions may change during complex formation also in Gcf1p. Our results indicate that Gcf1p explores multiple conformations in solution and compacts upon DNA binding, similar to TFAM and Abf2p [39], [42]. Characterization of the protein-DNA complexes in solution allowed us to elaborate a model in which HMG-box 1 of Gcf1p contacts the DNA and starts to dimerize through the formation of coiled coil structures through the N-terminal helix of Gcf1p in a concentration dependent process that is enhanced by the presence of positive charges at the N-terminal disordered tail. Due to the restrictions imposed by the crystal packing forces, we have been limited to crystals formed with 20bp DNA. Nevertheless, our analysis in solution reflects that the coiled coil mediated complex can be formed when binding DNAs of 50 bp. This could reflect a capability of the helices forming the coiled coil to slide respect each other thus allowing a greater separation between the globular part (the 'hammer-head') of Gcf1p. More feasibly this could be reflecting a DNA complex with unbound DNA segments. In both cases, our results in solution strongly suggest that the coiled coil interaction is not a crystal artefact but rather a relevant interaction potentially involved in DNA compaction.

Gcf1p multimerization behaviour can be related to its DNA bridging ability. Bridging have been widely described as a mechanism for the compaction of the bacterial nucleoid [153], [154], [155] and it can be therefore extrapolated to the mitochondrial nucleoid. In the bacterial nucleoid bridging is caused by dimerization of H-NS molecules that are contacting two DNA segments that are separated in sequence. In the case of Gcf1p, the observed DNA bridging could be due to the dimerization through the coiled coil domain. Alternatively, bridging could be caused by binding of two DNAs via both HMG-box 1 and HMG-box 2. In this second model, dimerization of Gcf1p by a coiled coil at the N-terminal helix should cause DNA kinks induced by HMG-box 2 spaced at a regular distance of approximately 50 bp. A possible third model would be the one in which bridging is caused by the coordinated action of HMG-box 1 and HMG-box 2, and additional bridging events can occur by the formation of coiled coil, The fact that bridging-like structure are also observable in TFAM, suggests that two HMG-boxes suffice for the formation of bridging. Despite bridging events can be observed for TFAM in published papers [90], the relationship between mitochondrial HMG boxes and DNA bridging has never been analysed in detail. It is our hope that our results encourage other researchers to explore the mechanism behind HMG box induced DNA bridging and its implications in mitochondrial DNA metabolism.

Our results do not allow us to discuss any direct relation between Gcf1p structure and recombination events. In parallel to the work shown in this thesis, we have performed preliminary recombination analysis *in vitro* although these experiments have not been included due to a lack of a significant number of replicates. Such experiments, performed using bacterial recombinases RecA and RecB, yielding inconclusive results on the effect of Gcf1p on recombination in lower protein:DNA bp ratios (1:500 to 1:100) whilst higher ratios (above 1:10 proteins per bp) result in an inhibition of the recombination. These results are coherent with preliminary topological assays (performed by collaborator Dr. Belén Martínez from professor Joaquim Roca laboratory) in which similar Gcf1p:DNA bp ratios results in Topoisomerase I not being able to access the DNA. These results suggest that high DNA compaction induced by Gcf1p results in the inhibition of reactions related with the metabolism of DNA, similarly to inhibition of transcription and replication at high TFAM:DNA bp ratios [90]. Nevertheless, our analysis of the protein:DNA complex in solution using SAXS does indicate a different complex is formed when binding recombination intermediates than when binding to linear DNA substrates (see section **R5.2** and **Table R4**). The lack of a crystal structure of a HMG-box protein in complex with a Holliday junction prevent us from doing a precise model that explains our SAXS data, albeit our *ab initio* models reflect a more spherical shape for the Gcf1p:Holliday junction complex than for those formed with linear DNAs (see section **R5.3.3**). A different recognition mechanism for Holliday junction than for linear DNA could be related with the role of Gcf1p in the stabilization of recombination intermediates, as proposed in previous published works [66]. Such a model is not incompatible with our data and would provide a substrate for functional switch in Gcf1p, from structural DNA binding protein to a recombination factor. Nevertheless, our results provide insights on the compaction role of Gcf1p. All our results are compatible with a model in which Gcf1p compacts DNA via a bridging mechanism, in which the two HMG-box domains contact different stretches of DNA and brings them together in a parallel or quasi-parallel arrangement (see sections **R3.6** for the crystal structure, **R5.3** for the validation of such a structure in solution with SAXS and **R6.2** and **R6.3** for electron microscopy evidences and discussion of bridging induced by Gcf1p). Although such a compaction mechanism is not directly related to the invasion strand events that are essential for homologous recombination, our results show that Gcf1p induces a specific arrangement on the DNA. We propose that the DNA arrangement induced by Gcf1p could influence and facilitate the triggering of such events. Our results are compatible with a model in which the architectural role of Gcf1p is strongly affecting the important recombination events that take place in *Candida albicans* mitochondrial DNA, that are directly related to replication [66] and that strongly correlate with the presence of Gcf1p [67]. Moreover, our crystal structure shows an DNA-end binding behaviour similar to that of Abf2p [42], which can be of relevance in some cell events where DNA ends are available, such as recombination after double-strand break.

Together, our results provide a mechanism for DNA recognition and binding by *Candida albicans* mtDNA maintenance protein Gcf1p and suggest an architectural role for this protein that can explain its implications in mtDNA metabolism and *Candida* viability. Gcf1p binds to DNA in an orchestrated manner: it contacts DNA by HMG box1 and it can contact and bend another DNA segment by its HMG-box 2, potentially forming a U-turn. Moreover, Gcf1p multimerizes upon DNA binding via coiled coil interactions mediated by its N-terminal helix. We have obtained experimental evidence that Gcf1p induces DNA bridging events, which poses the question whether the induction of U-turn is the universal mtDNA packaging mechanism or it can coexist with bridging. Further experimental evidence is required to trace a link between Gcf1p structure and DNA recombination. We are sure that our results will be a solid starting point for other colleagues to unveil the details behind mtDNA maintenance amongst species, which potentially can have a significant impact in the way we see the origin and diversification of eukaryotes.

CONCLUSIONS

1- Gcf1p binds to DNA via its two protein domains, HMG-1 and HMG-2 through contacts with the DNA major and minor groove, respectively. In our crystal structure, Gcf1p binds to different DNA molecules and each DNA molecule is, at the same time, contacted by two different proteins. The two DNA molecules are quasi-parallel in our crystal structure and a 90º bend is induced by binding of HMG2 to the minor groove. From such a model, it cannot be affirmed that a U-turn is imposed to DNA by Gcf1p, albeit it could be possible if another DNA substrate is involved. Gcf1p forms an intermolecular coiled coil in our crystal structure that is thermodynamically stable upon energy calculations. In our model, the first 33 residues corresponding to the N-terminal tail of our protein are not traced.

2- Gcf1p is a protein with flexible linkers flanking its DNA binding domains that adopt an extended conformation in solution. Gcf1p adopts a more compacted structure when bound to DNA. Protein-DNA and Protein-Protein interactions of the crystal structure are prevalent in solution.

3- Gcf1p interacts with the crystallized Af2_20 in solution in a sequential, concentration-dependent fashion:

3A- Gcf1p binds to DNA through the HMG-1 domain.

3B- Protein homodimers are formed through stablishment of coiled-coil interaction through the N-terminal portion of the protein.

3C- Protein homodimers further interact with DNA through HMG2, and other proteins are added to the complex forming assemblies of 3 proteins and 2 DNAs and 2 proteins and 3 DNAs.

3D- Upon concentration increase, the tetrameric assembly observed in our crystal structure can be formed, based on the interactions observed in solution.

4- Protein-DNA complex multimerization is hampered when the positive, disordered N-terminal tail of Gcf1p is not present, suggesting a function for this tail in stablishing contacts that stabilize the formation of the coiled coil.

5- Gcf1p binds to DNA showing cooperativity of binding and inducing local distortions, overall DNA compactions and features such as cross-strand binding and DNA wrapping. Gcf1p induces bridging events upon DNA binding as a combination of cooperative binding and cross-strand binding to DNA. At the same time these bridging events justify the observed local distortions as well as the overall DNA compaction.

6- Integration of our results suggests that Gcf1p induces DNA bridging as a result of cross-strand binding through HMG-box 1 and HMG-box 2. Formation coiled-coil mediated dimers on the DNA would result in a DNA binding site of approximately 50 bp. Such a binding site -which is, roughly, twice the one of TFAM and Abf2p- is justified by the formation of Gcf1p homodimers on the DNA through coiled coil formation. Formation of the coiled-coil mediated dimers fixes the HMG-box 2 induced distortions, as well as local cross-strand binding events 50 bp apart. Gcf1p induces DNA bridging as a result of the multiple combinatorial and reversible protein-protein and protein-DNA interactions.

7- Bridging of DNA by Gcf1p as observed in EM, offers a feasible option for mtDNA compaction in *Candida albicans*. It is compatible with our crystal structure and our structural analysis in solution. It is also compatible with the plastic regulation of mtDNA compaction state and with the recombination-driven replication context of *Candida albicans* mitochondrial genome.

REFERENCES

[1] J. C. Pérez, C. A. Kumamoto and A. D. Johnson, "Candida albicans commensalism and pathogenicity are intertwined traits directed by a tightly knit transcriptional regulatory circuit.," *PLOS biology,* pp. 11(3): 1-15 (e1001510), 2013.

[2] M. A. Ghannoum, R. J. Jurevic, P. K. Mukherjee, F. Cui, S. Masoumeh, A. Naqvi and P. M. Gillevet, "Characterization of the oral fungal microbiome (mycobiome) in healthy individuals," *PLoS Pathogenes,* pp. 8;6(1):1-8(e1000713), 2010.

[3] M. M. Barousse, B. J. Van der Pol, D. Fortenberry , D. Orr and P. L. Fidel Jr, "Vaginal yeast colonisation, prevalence of vaginitis," *AIDS patient care and STDs,* pp. 80:48-53, 2004.

[4] J. A. Romo and C. A. Kumamoto, "On commensalism of Candida," *Journal of Fungi,* pp. 6(16): 1-14 (10.3390/jof6010016), 2020.

[5] S. E. Reef, B. A. Lasker, D. S. Butcher, M. M. McNeil, R. Pruitt, H. Keyserling and W. R. Jarvis, "Nonperinatal nosocomial transmission of Candida albicans in a neonatal intensive care unit: Prospective study," *Journal of clinical microbiology,* pp. 36(5): 1255-1259, 1998.

[6] N. Kondori, F. Nowrouzian, M. Ajdari, B. Hesselmar, R. Saalman, A. E. Wold and I. Adlerberth, "Candida species as commensal gut colonizers: a study of 133 longitudinally followed swedish infants.," *Medical mycology,* p. myz091:10.1093/mmy/myz091, 2019.

[7] K. Atarashi, T. Tanoue, M. Ando, N. Kamada, Y. Nagano, S. Narushima, W. Suda, A. Imaoka, H. Setoyama, T. Nagamori, E. Ishikawa, T. Shima, T. Hara, S. Kado, T. Jinnohara, H. Ohno, T. Kondo, K. Toyooka, E. Watanabe, S. Yokoyama, S. Tokoro, H. Mori, Y. Noguchi, H. Morita, I. I. Ivanov, T. Sugiyama, G. Nuñez, J. G. Camp, M. Hattori, Y. Umesaki and K. Honda, "Th17 Cell induction by adhesion of microbes to intestinal epithelial cells," *Cell,* pp. 163(2):367-380, 2016.

[8] L. Markey, L. Shaban, E. R. Green, K. P. Lemon, J. Mecsas and C. A. Kumamoto, "Pre-colonization with the commensal fungus Candida albicans reduces murine susceptibility to Clostridium difficile infection," *Gut microbes,* pp. 9(6): 497-509, 2018.

[9] D. C. Ifrim, J. Quintin, L. Meerstein-Kessel, T. S. Plantinga, L. A. B. Joosten , J. W. M. van der Meer, F. L. van de Veerdonk i . M. G. Netea, «Defective trained immunity in patients with STAT-1-dependent chronic mucocutaneous candidiasis,» *The journal of translational immunology,* pp. 181: 434-440, 2015.

[10] D. W. Denning, M. Kneale, J. D. Sobel and R. Rautemaa-Richardson, "Global burden of recurrent vulvovaginal candidiasis: a systematic review," *THE LANCET Infectious diseases,* pp. 18(11): E339-E347, 2018.

[11] Centers for Disease Control and prevention (CDC), "Centers for Disease Control and prevention (CDC) - Fungal diseases," 31 May 2020. [Online]. Available: https://www.cdc.gov/fungal/diseases/candidiasis/thrush/index.html.

[12] B. J. Kullberg and M. C. Arendrup, “Invasive candidiasis,” *New England journal of medicine,* pp. 374(8): 794-795, 2016.

[13] H. Wisplinghoff, T. Bischoff, S. M. Tallent, H. Seifert, R. P. Wenzel and M. B. Edmond, “Nosocomial bloodstream infections in U.S. hospitals: analysis of 24179 cases from a prospective nationwide surveillance study,” *Clinical infectious diseases,* pp. 39: 309-317, 2004.

[14] S. S. Magill, E. O'Leary, S. J. Janelle, D. L. Thompson, G. Dumyati, J. Nadle, L. E. Wilson, M. A. Kainer, R. Lynfield, S. Greissman, S. M. Ray, Z. Beldavs, C. Gross, W. Bamberg, M. Sievers, C. Concannon, N. Buhr, L. Warnke, M. Maloney, V. Ocampo, J. Brooks, T. Oyewumi, S. Sharmin, K. Richards, J. Rainbow, M. Samper, E. B. Hancock, D. Leaptrot, E. Scalise, F. Badrun , R. Phelps i J. R. Edwards, «Changes in prevalence of health care - associated infections in U.S. hospitals,» *The New England journal of medicine,* pp. 372: 1732-1744, 2018.

[15] J. Morgan , M. I. Meltzer, B. D. Plikaytis, A. N. Sofair, S. Huie-White, S. Wilcox, L. H. Harrison, E. C. Seaberg, R. A. Hajjeh and S. M. Teutsch, “Excess mortality, hospital stay, and cost due to candidemia: A case-control study using data from population-based candidemia surveillance,” *Infection control & hospital epidemiology,* pp. 26(6): 540-547, 2005.

[16] O. Gudlaugsson, S. Gillespie, K. Lee, J. Vande Berg, J. Hu, S. Messer, L. Herwaldt, M. Pfaller i D. Diekema, «Attributable mortality of nosocomial candidemia, revisited,» *Clinical infectious diseases,* pp. 37: 1172-1177, 2003.

[17] «Mucosal damage and neutropenia are recquired for Candida albicans dissemination,» *PLoS pathogens,* pp. 4(2): 1-10 (e35), 2008.

[18] Centers for Disease Control and prevention (CDC), “Drug resistance in Candida,” 2019. [Online]. Available: https://www.cdc.gov/drugresistance/pdf/threats-report/candida-508.pdf.

[19] Centers for Disease Control and prevention, “Candida auris,” 15 May 2020. [Online]. Available: https://www.cdc.gov/fungal/candida-auris/index.html.

[20] G. M. Cooper and R. E. Hausman, “Cell Metabolism,” in *The Cell. A molecular approach*, 2004, pp. 73-103.

[21] P. López-García and D. Moreira, “Open questions on the origin of eukaryotes,” *Trends ecology evolution,* pp. 30(11):697-708, 2015.

[22] G. M. Cooper and R. E. Hausman, “An overview of cells and cell research,” in *The Cell. A molecular approach 4th edition.*, 2004, pp. 3-43.

[23] C. Mereschwosky , “Über natur und ursprung chromatophoren im Pflanzen-reiche,” *Biologisches Centralblatt,* pp. 25:593-604, 1905.

[24] W. Martin i K. V. Kowallik, «Annotated english translation of Mereschkowsky's 1905 paper 'Über Natur und Ursprung der Chromatophoren im Pflanzenreiche',» *European Journal of physiology,* pp. 34: 287-295, 1999.

[25] L. Sagan, "On the origin of mitosing cells," *Journal of theoretical biology,* pp. 14: 225-274, 1966.

[26] J. N. Timmis, M. A. Ayliffe, C. Y. Huang and W. Martin, "Endosymbiotic gene transfer: organelle genomes forge eukaryotic chromosomes," *Nature Reviews,* pp. 5: 131-135, 2004.

[27] M. W. Gray, "Mitochondrial evolution," *Cold Spring Harbor Perspectives in Biology,* pp. 4:a011403 1-16, 2012.

[28] D. G. Whitehouse, B. May and A. L. Moore, "Respiratory chain and ATP synthase," in *Elsevier Reference Collection in Biomedical Sciences*, Elsevier, 2019, pp. DOI: 10.1016/B978-0-12-801238-3.95732-5.

[29] I. Scott and R. J. Youle, "Mitochondrial fission and fusion," *Essays in biochemistry,* pp. 47: 85-98, 2010.

[30] D. C. Logan, "The mitochondrial compartment," *Journal of experimental botanics,* pp. 57(6): 1225-1243, 2006.

[31] C. Kukat, K. M. Davies, C. A. Wurm, H. Spahr, N. A. Bonekamp, I. Kühl, f. Joos, P. Loguercio Polosa, C. B. Park, V. Posse, M. Falkenberg, S. Jakobs, W. Kühlbrandt and N.-G. Larsson, "Cross-strand bindingof TFAMto a singlemtDNA molecule forms the mitochondrial nucleoid," *PNAS,* pp. 112(36) 11288-11293, 2015.

[32] J. M. Berg, L. Stryer and J. L. Tymoczko, "18-Oxidative phosphorylation," in *Biochemistry Sixth edition*, 2008, pp. 502-505.

[33] S. Cogliati, J. A. Enriquez and L. Scorrano, "Mitochondrial cristae: Where beauty meets functionality," *Trends in Biochemical Sciences,* pp. 41(3): 261-273, 2016.

[34] E. Carafoli, "The interplay of mitochondria with calcium: an historical appraisal," *Cell calcium,* pp. 5(1):1-8, 2012.

[35] T. Finkel, S. Menazza, K. M. Holmström, R. J. Parks, J. Liu, J. Sun, J. Liu, X. Pan and E. Murphy, "The ins and outs of mitochondrial calcium," *Circulation Research,* pp. 116(11): 1810-1819, 2015.

[36] M. Lopez-Cruzan, M. Sharma, M. Tiwari, S. Karbach, D. Holstein, C. Martin, J. Lechleiter and B. Herman, "Caspase-2 resides in the mitochondria and mediates apoptosis from the mitochondrial compartment," *Cell Death Discovery,* pp. 2, 16005; doi:10.1038/cddiscovery.2016.5, 2016.

[37] F. Hensen, A. Potter, S. L. van Esveld, A. Tarrés-Solé, A. Chakraborty, M. Solà i J. N. Spelbrink, «Mitochondrial RNA granules are critically dependent on mtDNA replication factors Twinkle and mtSSB,» *Nucleic Acids Research,* pp. 47(7): 3680-3698, 2019.

[38] R. A. Sia, S. Carrol, L. Kalifa, C. Hochmuth and E. A. Sia, “Loss of the Mitochondrial Nucleoid Protein, Abf2p, Destabilizes Repetitive DNA in the Yeast Mitochondrial Genome,” *Genetics,* pp. 181(1): 331-334, 2009.

[39] A. Rubio-Cosials, J. F. Sydow, N. Jiménez-Menéndez, P. Fernández-Millán, J. Montoya, H. T. Jacobs, M. Coll, P. Bernadó and M. Solà, “Human mitochondrial transcription factor A induces a U-turn structure in the Light-Strand Promoter,” *Nature Structure and Molecular Biology,* pp. 18(11): 1281-1290, 2011.

[40] H. B. Ngo, J. T. Kaiser and D. C. Chan, “Tfam, a mitochondrial transcription and packaging factor, imposes a U-turn on mitochondrial DNA,” *Nature Structure and Molecular Biology,* pp. 18(11): 1290-1296, 2011.

[41] A. Cuppari, P. Fernández-Millán, F. Battistini, A. Tarrés-Solé, S. Lyonnais , G. Iruela, Y. Enciso, A. Rubio-Cosials, R. Prohens, M. Pons, C. Alfonso, K. Tóth, G. Rivas, M. Orozco and M. Solà, “DNA specificities modulate the binding of human transcription factor A to mitochondrial DNA control region,” *Nucleic Acids Research,* pp. 47(12): 6519-6537, 2019.

[42] A. Chakraborty, S. Lyonnais, F. Battistini, A. Hospital, G. Medici, R. Prohens, M. Orozco, J. Vilardell and M. Solà, “DNA structure directs positioning of the mitochondrial genome packaging protein Abf2p,” *Nucleic Acids Research,* pp. 45(2): 951-967, 2017.

[43] D. F. Bogenhagen, D. Rousseau and S. Burke, “The layered structure of human mitochondrial DNA nucleoids,” *Journal of biological chemistry,* pp. 283: 3665-375, 2008.

[44] B. Huang, M. Bates and X. Zhuang, “Super-resolution fluorescence microscopy,” *Annual reviews in biochemistry,* pp. 993-1016, 2009.

[45] C. Kukat, C. A. Wurm, H. Spahr, M. Falkenberg, N.-G. Larsson and S. Jakobs, “Super-resolution Microscopy Reveals That Mammalian Mitochondrial Nucleoids Have a Uniform Size and Frequently Contain a Single Copy of mtDNA,” *PNAS,* pp. 108(33):13534-13539, 2011.

[46] P. Jezek, S. Tomás , J. Tauber and V. Pavluch , “Mitochondrial nucleoids: superresolution microscopy analysis,” *Organelles in focus,* pp. 106: 21-25, 2019.

[47] Y. Shi, A. Dierckx, P. H. Wanrooij, S. Wanrooij , N.-G. Larsson, L. M. Wilhelmsson, M. Falkenberg and C. M. Gustafsson, “Mammalian transcription factor A is a core component of the mitochondrial transcription machinery,” *PNAS,* pp. 109(41): 16510-16515, 2012.

[48] S. R. Lee and J. Han , “Mitochondrial nucleoid: shield and switch of the mitochondrial genome,” *Oxidative medicine and cellular longevity,* p. doi:8060949, 2017.

[49] MITOMAP-Foswiki, “MITOMAP,” 2020. [Online]. Available: https://www.mitomap.org/MITOMAP.

[50] C. Piro-Mégy, E. Sarzi, A. Tarrés-Solé, M. Péquignot, F. Hensen, M. Quilès, G. Manes, A. Chakraborty, A. Sénéchal, B. Bocquet, C. Cazevieille, A. Roubertie, A. Müller, M. Charif, D. Goudenège , G. Lenaers, H. Wilhelm, U. Kellner, N. Weisschuh, B. Wissinger , X. Zanlonghi, C. Hamel, J. Spelbrink, M. Sola

and C. Delettre, “Dominant mutations in mtDNA maintenance gene SSBP1 cause optic atrophy and foveopathy,” *Journal of clinical investigation,* pp. 130(1): 143-156, 2020.

[51] D. Bogenhagen, Y. Wang, E. Shen and R. Kobayashi, “Protein components of mitochondrial DNA nucleoids,” *The journal of biological chemistry,* pp. 2(11): 1205-1216, 2003.

[52] J. Kaukonen, J. Juselius, V. Tiranti, A. Kyttälä, M. Zeviani, G. Comi, S. Keränen, L. Peltonen and A. Suomalainen, “Role of adenine nucleotide translocator 1 in mtDNA maintenance,” *Science,* pp. 289(5480):782-785, 2000.

[53] S. R. Lee, N. Kim, Y. Noh, Z. Xu, K. S. Ko, B. D. Rhee and J. Han, “Mitochondrial DNA, mitochondrial dysfunction and cardiac manifestations,” *Frontiers in Bioscience,* pp. 21: 1410-1426, 2016.

[54] T. Olejár, D. Pajuelo-Reguera, L. Alán, A. Dlasková i P. Jezek, «Coupled aggregation of mitochondrial single-strand DNA-binding protein tagged with Eos fluorescent protein visualizes synchronized activity of mitochondrial nucleoids,» *Molecular medicine reports,* pp. 12: 5185-5190, 2015.

[55] D. L. Nelson and M. Cox, “Bioenergetics and metabolism,” in *Lehninger Principles of Biochemistry*, 2008, pp. 505-543.

[56] D. L. Nelson and M. Cox, “Glycolisis, gluconeogenesis and the pentose phosphate pathway.,” in *Lehninger Principles of Biochemistry Sixth Edition*, 2008, pp. 543-587.

[57] D. L. Nelson i M. Cox, «Fatty acid catabolism,» de *Lehininger Principles of Biochemistry Sixth Edition*, 2008, pp. 667-695.

[58] D. L. Nelson i M. Cox, «The citric acid cycle,» de *Lehninger Principles of Biochemistry Sixth Edition*, 2008, pp. 633-667.

[59] D. L. Nelson and M. Cox, “Oxidative phosphorylation and photophosphorylation,” in *Lehninger Principles of Biochemistry Sixth Edition*, 2008, pp. 732-788.

[60] P. Mitchell, “Coupling of phosphorylation to electron and hydrogen transfer by a chemi-osmotic type of mechanism,” *Nature,* pp. 144-148, 1961.

[61] A. Kolondra, K. Labedzka-Dmoch, J. M. Wenda, K. Drzewicka and P. Golik, “The transcriptome of Candida albicans mitochondria and the evolution of organellar transcription units in yeasts.,” *BMC Genomics,* pp. 16(827): 1-22 (doi:10.1186/s12864-015-2078-z), 2015.

[62] D. Williamson, “The curious history of yeast mitochondrial DNA,” *Nature reviews,* pp. 3: 1-7, 2002.

[63] J. W. Willis, W. B. Troutman and W. S. Riggsby, “Circular mitochondrial genome of Candida albicans contains a large inverted duplication,” *Journal of bacteriology,* pp. 164(1): 7-13, 1985.

[64] A. J. Bendich, “Reaching for the ring: the study of mitochondrial genome structure,” *Current genetics,* pp. 24: 279-290, 1993.

[65] A. J. Bendich, “The end of the circle for yeast mitochondrial DNA,” *Molecular Cell Previews,* pp. 39: 831-832, 2010.

[66] J. M. Gerhold, A. Aun, T. Sedman, P. Joers and J. Sedman, “Strand invasion structures in the inverted repeat of Candida albicans mitochondrial DNA reveal a role for homologous recombination in replication,” *Molecular Cell,* pp. 39: 851-861, 2010.

[67] J. M. Gerhold, T. Sedman, K. Visacka, J. Slezakova, L. Tomaska, J. Nosek and J. Sedman, “Replication Intermediates of the Linear Mitochondrial DNA of Candida parapsilosis suggest a common recombination-based mechanism for yeast mitochondria,” *The journal of biological chemistry,* pp. 289(33): 22659-22670, 2014.

[68] M. Bustin, «Regulation of DNA-Dependent Activities by the Functional Motifs of the High-Mobility-Group Chromosomal Proteins,» *Molecular and cellular biology,* pp. 19(8): 5237-5246, 1999.

[69] C. S. Malarkey i M. E. Churchill, «The high mobility group box: the ultimate utility player of a cell,» *Trends in biochemical sciences,* pp. 37(12): 553-562, 2012.

[70] E. Fonfría-Subirós, F. Acosta-Reyes, N. Saperas, J. Pous, J. A. Subirana and J. L. Campos, “Crystal structure of a complex of DNA with one AT-Hook of HMGA1,” *PLoS one,* pp. Issue5, doi:e37120, 2012.

[71] R. Sgarra, S. Zammiti, A. Lo Sardo, E. Maurizio, L. Arnoldo, S. Pegoraro, V. Giancotti and G. Manfioletti, “HMGA molecular network: From transcriptional regulation to chromatin remodeling,” *Biochimica et Biophysica Acta (BBA) Gene regulatory mechanisms,* pp. 37-47, 2010.

[72] N. Zhu and U. Hansen, “Transcriptional regulation by HMGN proteins,” *Biochimica et Biophysica Acta (BBA) Gene Regulatory Mechanisms,* pp. 179 (1-2): 74. doi:10.1016/j.bbagrm.2009.11.006., 2010.

[73] H.-F. Ding, M. Bustin and U. Hansen, “Alleviation of Histone H1-mediated transcriptional repression and chromatin compaction by the acidic activation region in chromosomal protein HMG-14,” *American society for microbiology,* pp. 5843-5855, 1997.

[74] M. P. Crippa, L. Trieschmann, P. J. Alfonso, A. P. Wolffe and M. Bustin, “Deposition of chromosomal protein HMG-17 during replication affects the nucleosomal ladder and transcriptional potential of nascent chromatin,” *The EMBO Journal ,* pp. 12(10): 3855-3864, 1993.

[75] D. Yesudhas, M. Batool, M. A. Anwar, S. Panneerselvam and S. Choi, “Proteins recognizing DNA: Structural uniqueness and versatility of DNA binding domains in stem cell transcription factors,” *Genes,* p. 8(192): doi:10.3390/genes8080192, 2017.

[76] S. Soullier, P. Jay, F. Poulat, J.-M. Vanacker, P. Berta and V. Laudet, “Diversification pattern of the HMG and SOX family members during evolution,” *Journal of molecular evolution,* pp. 48:517-527, 1999.

[77] H. M. Weir, P. J. Kraulis, C. S. Hill, A. R. Raine, E. D. Laue and J. O. Thomas, “Structure of the HMG box motif in the B-domain of HMG1,” *The EMBO Journal,* pp. 12(4): 1311-1319, 1993.

[78] J. E. Masse, B. Wong, Y.-M. Yen, F. H. Allain, R. C. Johnson and J. Feigon, “The S.cerevisiae architectural HMGB protein NHP6A complexed with DNA: DNA and protein conformational changes upon binding,” *Journal of molecular biology,* pp. 323: 263-284, 2002.

[79] Y. Xu, W. Yang, J. Wu and Y. Shi, “Solution structure of the first HMG box domain in Human Upstream Binding Factor,” *Biochemistry,* pp. 41: 5415-5420, 2002.

[80] H. Rong, Y. Li, X. Shi, X. Zhang, Y. Gao, H. Dai, M. Teng, L. Niu, Q. Liu and Q. Hao, “Structure of human upstream binding factor HMG box 5 and site for binding of the cell-cycle regulatory factor TAF1,” *Acta crystallographica section D,* pp. D63: 730-737, 2007.

[81] J. Wang, N. Tochio, A. Takeuchi, J.-i. Uewaki, N. Kobayashi and S.-I. Tate, “Redox-sensitive structural change in the A-domain of HMGB1 and its implication for the binding of cisplatin modified DNA,” *Biochemical and biophysical research communications,* pp. 441: 701-706, 2013.

[82] P. Palasingam, R. Jaunch, C. Keow Leng Ng and P. R. Kolatkar, “The structure of Sox 17 bound to DNA reveals a conserved bending topology but selective protein interaction platforms,” *Journal of molecular biology,* pp. 619-630, 2009.

[83] R. Jauch, C. K. L. Ng, K. Narasimhan and P. R. Kolatkar, “The crystal structure of the Sox4 HMG domain-DNA complex suggests a mechanism for positional interdependence in DNA recognition,” *Biochemical Journal,* pp. 443: 39-47, 2012.

[84] R. Sánchez-Giraldo, F. Acosta-Reyes, C. Malarkey, N. Saperas, M. Churchill and J. Campos, “Teo high-mobility group box domains act together to underwind and kink DNA,” *Acta Crystallographica D,* pp. D71: 1423-1432, 2015.

[85] A. Rubio-Cosials, F. Battistini, A. Gansen, A. Cuppari, P. Bernadó , M. Orozco, J. Langowski, K. Tóth and M. Solà, “Protein flexibility and synergy of HMG domains underlie U-turn bending of DNA by TFAM in solution,” *Biophysical Journal,* pp. 114: 2386-2396, 2018.

[86] T. S. Wong, S. Rajagopalan, S. M. Freund, T. J. Rutherford, A. Andreeva , F. M. Townsley, M. Petrovich and A. R. Fersht, “Biophysical characterization of Human Mitochondrial Transcription Factor A and its binding to tumor supressor p53,” *Nucleic Acids Research,* pp. 37(20):6765-83, 2009.

[87] T. A. Gangelhoff, P. S. Mungalachetty, J. C. Nix and M. E. Churchill , “Structural analysis and DNA binding of the HMG domains of the human mitochondrial transcription factor A,” *Nucleic Acids Research,* pp. 37(10): 3153-3164, 2009.

[88] H. S. Hillen, D. Temiakov and P. Cramer, “Structural basis of mitochondrial transcription,” *Nature structural and molecular biology,* pp. 25:754-765, 2018.

[89] G. Farge, N. Laurens, O. D. Broekmans, S. M. van den Wildenberg, L. C. Dekker, M. Gaspari, C. M. Gustafsson, E. J. Peterman, M. Falkenberg and G. J. Wuite, "Protein sliding and DNA denaturation are essential for DNA organization by human mitochondrial transcription factor A," *Nature communications,* p. DOI: 10.1038/ncomms2001, 2012.

[90] G. Farge, M. Mehmedovic, M. Baclayon, S. M. van den Wildenberg, C. Gustafsson, G. J. Wuite and M. Falkenberg, "In-vitro reconstituted nucleoids can block mitochondrial DNA replication and transcription," *Cell Reports,* pp. 8: 66-74, 2014.

[91] K. Maniura-Weber, S. Goffart, H. L. Garstka, J. Montoya and R. J. Wiesner, "Transient overexpression of mitochondrial transcription factor A (TFAM) is sufficient to stimulate mitochondrial DNA transcription, but not sufficient to increase mtDNA copy number in cultured cells," *Nucleic Acids Research,* pp. 32(20): 6015-6027, 2004.

[92] J. L. Pohjoismäki, S. Wanrooij, A. K. Hyvärinen, Goffart Steffi, I. J. Holt, J. N. Spelbrink and H. T. Jacobs, "Alterations to the expression level of mitochondrial transcription facot A, TFAM, mofidy the mode of mitochondrial DNA replication in cultured human cells," *Nucleic Acids Research,* pp. 34(20): 5815-5828, 2006.

[93] D. J. Dairaghi, G. S. Shadel and D. A. Clayton, "Addition of a 29 residue carboxyl-terminal tail converts a simple HMG Box-containing protein into a transcriptional activator," *Journal of Molecular Biology,* pp. 249(1): 11-28, 1995.

[94] L. Tabernero, N. Verdaguer, M. Coll, I. Fita , G. A. van der Marel, J. H. van Boom, A. Rich and J. Aymamí, "Molecular structure of the A-tract DNA dodecamer d(CGCAAATTTGCG) complexed with the minor groove binding drug netropsin," *Biochemistry,* pp. 32: 8403-8410, 1993.

[95] P. J. Hagerman, "Sequence-directed curvature of DNA," *Annueal reviews in biochemistry,* pp. 59: 755-781, 1990.

[96] D. S. Goodsell and R. E. Dickerson, "Bending and curvature calculations in B-DNA," *Nucleic Acids Research,* pp. 22(24): 5497-5503, 1994.

[97] A. Fiorini, F. Gimenes, Q. Alves de Lima Neto, F. R. Rosado and M. A. Fernandez, "Sequence-directed DNA curvature in replication origin segments," 2011, p. doi:10.13140/2.1.3377.7604.

[98] T. Drsata, N. Spackova, P. Jurecka, M. Zgarbová, J. Sponer and F. Lankas, "Mechanical properties of symmetric and asymmetric DNA A-tracts: implications for looping and nucleosome positioning," *Nucleic acids research,* pp. 42 (11) 7383-7394, 2014.

[99] J. F. Diffley i B. Stillman, «A close relative of the nuclear, chromosomal high-mobility group protein HMG1 in yeast mitochondria,» *Procedures of the National Academy of Sciences USA,* pp. 88: 7864-7868, 1991.

[100] R. W. Friddle, J. E. Klare, S. S. Martin, M. Corzett, R. Balhorn , E. P. Baldwin, R. J. Baskin and A. Noy, "Mechanism of DNA compaction by yeast mitochondrial protein Abf2p," *Biophysical journal,* pp. 86: 1632-1639, 2004.

[101] O. Zelenaya-Troitskaya, S. M. Newman, K. Okamoto, P. S. Perlman and R. A. Butow, "Functions of the High mobility group protein, Abf2p, in mitochondrial DNA segregation, recombination and copy number in Saccharomyces cerevisiae," *Genetics society of America,* pp. 148: 1763-1776, 1998.

[102] J. Bakkaiova, V. Marini, S. Willcox, J. Nosek, J. D. Griffith, L. Krejci and L. Tomaska, "Yeast mitochondrial HMG proteins: DNA-binding properties of the most evolutionarily divergent component of mitochondrial nucleoids," *Bioscience Reports,* pp. 36(1): e00288 1-13, 2016.

[103] D. M. MacAlpine, P. S. Perlman and R. A. Butow , "The high mobility group protein Abf2p influences the level of yeast mitochondrial DNA recombination intermediates in vivo.," *Proceedings National Academy of Sciences USA,* pp. 95: 6739-6743, 1998.

[104] K. Visacka, J. M. Gerhold, J. Petrovicova, S. Kinsky, P. Jõers, J. Nosek, J. Sedman and L. Tomaska, "Novel subfamily of mitochondrial HMG box-containing proteins: Functional analysis of Gcf1p from Candida albicans," *Microbiology,* pp. 155 (4):1226-1240, 2009.

[105] Insightful science, "snapgene.com," [Online]. Available: https://www.snapgene.com/snapgene-viewer/.

[106] Dept. of Protein Evolution MPI-Tuebingen, «Max Planck Institute for Developmental Biology,» 2008-2020. [En línia]. Available: https://toolkit.tuebingen.mpg.de/.

[107] Zimmermann L, "A Completely Reimplemented MPI Bioinformatics Toolkit with a New HHpred Server at its Core.," *Journal of Molecular Biology,* pp. Jul 20. S0022-2836(17)30587-30589., 2018.

[108] B. Mészáros, G. Erdõs and Z. Dosztányi, "IUPred2A: context-dependent prediction of protein disorder as a function of redox state and protein binding," *Nucleic Acids Research,* pp. 46(W1): W329-337, 2018.

[109] S. El-Gebali, J. Mistry, A. Bateman, S. Eddy, A. Luciani, S. Potter, M. Qureshi, L. Richardson, G. Salazar, A. Smart, E. Sonnhammer, L. Hirsh, L. Paladin, D. Piovesan, S. Tosatto and R. Finn, "The Pfam protein families database in 2019," *Nucleic Acids Research,* p. doi: 10.1093/nar/gky995, 2019.

[110] A. Drozdetskiy, C. Cole, J. Procter and G. J. Barton, "JPred4: a protein secondary structure prediction server," *Nucleic Acids Research,* p. 10.1093/nar/gkv332, 2015.

[111] GE-Healthcare Life Sciences, "GST Gene Fusion System Handbook," November 2014. [Online]. Available: https://www.sigmaaldrich.com/content/dam/sigma-aldrich/docs/Sigma-Aldrich/General_Information/1/ge-gst-gene-fusion.pdf.

[112] T. Goulas, A. Cuppari, R. Garcia-Castellanos, S. Snipas, R. Glockshuber, J. L. Arolas and F. X. Gomis-Rüth, "The pCri System: A Vector Collection for Recombinant Protein Expression and Purification," *PLOSOne,* pp. 9(11) e112643 1-10, 2014.

[113] F. van der Ent and J. Löwe, “methods SeMet Recipe,” [Online]. Available: https://www2.mrc-lmb.cam.ac.uk/groups/JYL/methods/SeMet%20recipe.doc.pdf.

[114] F. van der Ent and J. Löwe, “Expression of selenomethionine substituted proteins in non-methionine auxotrophic E.coli,” 11 June 2003. [Online]. Available: https://www2.mrc-lmb.cam.ac.uk/groups/JYL/methods/SeMet%20recipe.doc.pdf.

[115] G. D. Van Duyne, R. F. Standaert, P. A. Karplus, S. L. Schreiber and J. Clardy, “Atomic Structures of the Human Immunophilin FKBP-12 Complexes with FK506 and Rapamycin,” *Journal of Molecular Biology,* pp. 229 (1) 105-124, 1993.

[116] T. Waterboer, P. Sehr, K. M. Michael, S. Franceschi, J. D. Nieland, T. O. Joos, M. F. Templin and M. Pawlita, “Multiplex human papillomavirus serology based on in situ-purified glutathione S-transferase fusion proteins,” *Clinical Chemistry,* pp. 51:10 1845-1853, 2005.

[117] M. A. Herzik, M. Wu and G. C. Lander, “High-resolution structure determination of sub-100 kDa complexes using conventional cryo-EM,” *Nature Communications,* pp. 10: 1-9, 2019.

[118] C. Tang and Z. Yang, “Transmission Electron Microscopy (TEM),” in *Membrane characterization*, Elsevier, 2017, pp. Chapter 8: 145-158.

[119] C. Brack, “DNA Electron Microscopy,” *Critical Reviews in Biochemistry and Molecular Biology,* pp. 113-141, 1981.

[120] S. Lyonnais, R. J. Gorelick, F. Heniche-Boukhalfa, S. Bouaziz, V. Parisi, J.-F. Mouscadet, T. Restle, J. M. Gatell, É. Le Cam and G. Mirambeau, “A protein ballet around the viral genome orchestrated by HIV-1 reverse transcriptase leads to an architectural switch: from nucleocapsid-condensed RNA to Vpr-bridged DNA,” *Virus Research,* pp. 171(2): 287-303, 2013.

[121] J. Dubochet and A. McDowall, "Vitrification of pure water for electron microscopy," *Microscopy,* pp. 124(3): 3-4, 1981.

[122] M. G. Rossmann, “The molecular replacement method,” *Acta crystallographica A,* pp. 46: 73-82, 1990.

[123] Center for Biological Sequence analysis CBS, «TargetP-2.0 Server within CBS prediction servers,» 29 Abril 2019. [En línia]. Available: http://www.cbs.dtu.dk/services/TargetP/.

[124] J. J. Almagro Armenteros, M. Salvatore, O. Winther, O. Emanuelsson, G. von Heijne, A. Elofsson i H. Nielsen, «Detecting sequence signals in tergeting peptides using deep learning,» *Life science alliance,* pp. 2 (5), e201900429, 2019.

[125] Helmoltz Center Munich, 2020. [Online]. Available: https://ihg.gsf.de/ihg/mitoprot.html.

[126] M. G. Claros and P. Vincens, “Computational method to predict mitochondrially imported proteins and their targeting sequences,” *European Journal of Biochemistry,* pp. 779-786 (1996), 1996.

[127] ExPASy , “ProtParam tool within ExPASy-SIB Bioinformatics Resource Portal,” 2020. [Online]. Available: https://web.expasy.org/protparam/.

[128] compbio Dundee, 2020. [Online]. Available: http://www.compbio.dundee.ac.uk/jpred/.

[129] A. Lupas, M. Van Dyke and J. Stock, “Predicting Coiled Coils from Protein Sequences,” *Science,* pp. 252: 1162-1164, 1991.

[130] EMBL-EBI, “Pfam 32.0,” 2020. [Online]. Available: https://pfam.xfam.org/.

[131] S. El-Gebali, J. Mistry, A. Bateman, S. R. Eddy, A. Luciani, S. C. Potter, M. Qureshi, L. J. Richardson, G. A. Salazar, A. Smart, E. L. Sonnhammer, L. Hirsh, L. Paladin, D. Piovesan, S. C. Tosatto and R. D. Finn, “The Pfam protein families database in 2019,” *Nucleic Acids Research,* pp. D1: D427-D432, 2019.

[132] N. Mizushima, T. Noda and Y. Ohsumi, “Apg16p is required for the function of the Apg12p-Apg5p conjugate in the yeast autophagy pathway,” *The EMBO Journal,* pp. 18(14):3888-3896, 1999.

[133] CRG, “TCOFFEE,” 2020. [Online]. Available: http://tcoffee.crg.cat/apps/tcoffee/do:expresso.

[134] C. Notredame, D. G. Higgins and J. Heringa, “T-Coffee: A novel method for fast and accurate multiple sequence alignment,” *Journal of Molecular Biology,* pp. 302: 205-217, 2000.

[135] P. R. Evans and G. Murshudov, “How good are my data and what is the resolution?,” *Acta Crystallographica Section D,* pp. 1204-1214, 2013.

[136] J. Foadi, P. Aller, Y. Alguel, A. Cameron, D. Axford, R. L. Owen , W. Armour, D. G. Waterman, S. Iwata y G. Evans, «Clustering procedures for the optimal selection of data sets from numerous crystals in macromolecular crystallography,» *Acta crystallographica D,* pp. 69(8): 1617-1632, 2013.

[137] I. Tickle, C. Flensburg, P. Keller, W. Paciorek, A. Sharff, C. Vornheim and G. Bricogne, “Global Phasing Ltd.,” 2018. [Online]. Available: http://staraniso.globalphasing.org/cgi-bin/staraniso.cgi. [Accessed 13th March 2020].

[138] D. Liebschner, P. Afonine, M. Baker, G. Bunkóczi, V. Chen, T. Croll, B. Hintze, L. Hung, S. Jain, A. McCoy, N. Moriarty, R. Oeffner, B. Poon, M. Prisant, R. Read, J. Richardson, D. Richardson, M. Sammito, O. Sobolev, D. Stockwell, T. Terwilliger, A. Urzhumtsev, L. Videau, C. Williams and P. Adams, “Macromolecular structure determination using X-rays, neutrons and electrons: recent developments in Phenix.,” *Acta Crystallographica D,* pp. 75: 861-877, 2019.

[139] J. S. Richardson, “II. Basic elements of protein structure, tight turns in protein structure.,” in *The anatomy and taxonomy of protein structure*, 2007, pp. 203-216.

[140] G. B. McGaughey, M. Gagné and A. K. Rappé, “Pi-stacking interactions, alive and well in proteins.,” *Journal of Biological Chemistry,* pp. 25(273): 15458-15463, 1998.

[141] C. S. Malarkey and M. E. Churchill, “The high mobility group box: the ultimate utility player of the cell,” *Trends in biochemical sciences,* pp. 37(12): 553-562, 2012.

[142] T. Drsata, N. Spackova, P. Jurecka, M. Zgarbova, J. Sponer and F. Lankas, “Mechanical properties of symmetric and asymmetric DNA A-tracts: implications for looping and nucleosome positioning.,” *Nucleic Acids Research,* pp. 42(11): 7383-7394, 2014.

[143] D. Gusenko y S. V. Strelkov, «CCFold: rapid and accurate prediction of coiled-coil structures and application to modelling intermediate filaments,» *Bioinformatics,* pp. 34(2): 215-222, 2018.

[144] PDBe, “PDBe-PISA server,” May 2020. [Online]. Available: https://www.ebi.ac.uk/msd-srv/prot_int/cgi-bin/piserver.

[145] N. R. Hajizadeh, D. Franke, C. M. Jeffries i D. I. Svergun, «Consensus Bayesian assessment of protein molecular mass from solution X-ray scattering curves,» *Scientific Reports,* pp. 8:7204 (1-12) | DOI:10.1038/s41598-018-25355-2 , 2018.

[146] A. Cuppari, P. Fernández-Millán, F. Battistini, A. Tarrés-Solé, S. Lyonnais, G. Iruela, E. Ruiz-López, Y. Enciso, A. Rubio-Cosials, R. Prohens, M. Pons, C. Alfonso, K. Tóth, G. Rivas, M. Orozco i M. Solà, «DNA Specificities Modulate the Binding of Human Transcription Factor A to Mitochondrial DNA Control Region,» *Nucleic Acids Research,* pp. 9;47(12):6519-6537, 2019.

[147] P. Bernadó, E. Mylonas, M. V. Petoukhov, M. Blackledge and D. I. Svergun, “Structural characterization of flexible proteins using Small-Angle X-ray Scattering,” *Journal of the American Chemistry Society,* pp. 129: 5656-5664, 2007.

[148] D. Svergun, C. Barberato and M. Koch, “CRYSOL- a program to evaluate X-ray solution scattering of biological macromolecules from atomic coordinates,” *Journal of applied crystallography,* pp. 28. 768-773, 1995.

[149] A. Panjkovic i D. I. Svergun, «Deciphering conformational transitions of proteins by small angle X-ray scattering and normal mode analysis,» *Physical chemistry Chemical physics,* pp. 18:5707-5719, 2016.

[150] P. V. Konarev, V. V. Volkov, A. V. Sokolova, M. H. Koch and D. I. Svergun, “PRIMUS: a Windows PC-based system for small-angle scattering data analysis,” *Journal of Applied Crystallography,* pp. 36: 1277-1282, 2003.

[151] P.-C. Chen and J. S. Hub, “Validating solution ensembles from molecular dynamics simulations by wide-angle X-ray scattering data,” *Biophysical Journal,* pp. 107: 435-447, 2014.

[152] C. J. Knight and J. S. Hub, “WAXSiS: a web server for the calculation of SAXS/WAXS curves based on explicit-solvent molecular dynamics,” *Nucleic Acids Research,* pp. 43: 225-230, 2015.

[153] R. T. Dame, M. S. Luijsterburg, E. Krin, P. N. Bertin, R. Wagner i G. J. Wuite, «DNA Bridging: a property shared among H-NS-like proteins,» *Journal of bacteriology,* pp. DOI: 10.1128/JB.187.5.1845-1848.2005, 2005.

[154] J. M. Chen, H. Ren, J. E. Shaw, Y. J. Wang, M. Li, A. S. Leung, V. Tran, N. M. Berbenetz, D. Kocíncová, C. M. Yip, J.-M. Reyrat and J. Liu, "Lsr2 of Mycobacterium tuberculosis is a DNA-bridging protein," *Nucleic Acids Research,* pp. 36(7): 2123-2135, 2008.

[155] C. J. Dorman and K. A. Kane, "DNA bridging and antibridging: a role for bacterial nucleoid-associated proteins in regulating the expression of laterally acquired genes.," *Federation of European Microbiology Societies,* pp. DOI: 10.1111/j.1574-6976.2008.00155.x, 2008.

[156] A. G. Kikhney i D. I. Svergun, «A practical guide to samll angle X-ray scattering (SAXS) of flexible and intrinsically disordered proteins,» *FEBS Letters,* pp. 589 2570-2577, 2015.

[157] D. I. Svergun and M. H. Koch, "Small-Angle scattering studies of biological macromolecules in solution," *Reports on Progress in Physics,* pp. 66 1735-1782, 2003.

[158] ssrl-BL4-2 website, "Stanford Synchrotron Radiation Lightsource website," 1st February 2017. [Online]. Available: https://www-ssrl.slac.stanford.edu/~saxs/analysis/assessment.htm. [Accessed 10th March 2020].

[159] Cornell Synchrotron, BioSAXS group website, "Cornell High energy Synchrotron Source (CHESS) website," [Online]. Available: https://www.chess.cornell.edu/macchess/biosaxs/whatcan_biosaxs. [Accessed 10th March 2020].

[160] N. S. Murthy, «Recent Developments in Small-Angle X-Ray Scattering,» *Spectroscopy-Special Issues,* pp. 32(11) 18-24, 2017.

[161] A. Alford , V. Kozlovskaya and E. Kharlampieva, "Small Angle Scattering for Pharmaceutical Applications: From Drugs to Drug Delivery Systems," in *Biological Small Angle Scattering: Techniques Strategies and Tips*, Springer Nature Singapore Ltd., 2017, pp. Chapter 15: 239-263.

[162] M. Brennich, P. Pernot and A. Round, " How to analyze an Present SAS Data for Publication," in *Biological Small-Angle X-ray Scattering techniques, strategies and tips.* , Springer Nature Singapore Ltd, 2017, pp. Chapter 4: 47-61.

[163] D. I. Svergun, "Restoring low resolution structure of biological macromolecules from solution scattering using simulated annealing.," *Biophysics Journal,* pp. 2879-2886, 1999.

[164] D. I. Svergun, M. V. Petoukhov and M. H. Koch, "Determination of domain structure of protein from X-ray solution scattering," *Biophysical Journal,* pp. 80: 2946-2953, 2001.

[165] V. V. Volkov and D. I. Svergun, "Uniqueness of ab initio shape determination in small-angle scattering," *Journal of applied crystallography,* pp. 36: 860-864, 2003.

[166] G. Tria, H. D. Mertens, M. Kachala and D. I. Svergun, “Advanced ensemble modelling of flexible macromolecules using X-ray solution scattering,” *International Union of Crystallographers Journal,* pp. 2: 207-217, 2015.

[167] P. Smith, R. Krohn, G. Hermanson, A. Mallia, F. Gartner, M. Provenzano, E. Fujimoto, N. Goeke, B. Olson and D. Klenk, “Measurement of protein using bicinchoninic acid,” *Analytical biochemistry,* pp. 150(1): 76-85, 1985.

[168] D. I. Svergun, “Determination of the redularization parameter in Indirect-transform methods using perceptual criteria,” *Journal of Applied Crystallography,* pp. 25: 495-503, 1992.

[169] D. Franke, M. Petoukhov, P. Konarev, A. Panjkovich, A. Tuukkanen, H. Mertens, A. Kikhney, N. Hajizadeh, J. Franklin, C. Jeffries and D. I. Svergun, “ATSAS 2.8: a comprehensive data analysis suite for small-angle scattering from macromoledular solutions,” *Journal of applied crystallography,* pp. 50: 1212-1225, 2017.

[170] N. Asherie, “Protein crystallization and phase diagrams,” *Methods,* pp. 34(3): 266-272, 2004.

[171] F. Hofmeister, “Zur Lehre von der Wirkung der Salze, translated in (about the science of the effect of salts: Franz Hofmeister's historical papers) Kunz W, Henle J and Ninham BW,” *Current Opinion in colloid and Interface science 2004 (Translation),* pp. 9: 19-37, 1888 .

[172] A. McPherson and J. A. Gavira, “Introduction to protein crystallization,” *Acta Crystallographiva Section F,* pp. 2-20, 2013.

[173] T. Hollis, “Crystallization of protein-DNA complexes,” in *Macromolecular crystallography protocols*, Humana Press, 2007, pp. 225-237.

[174] IQFR-CSIC, “Instituto de Química Física Rocasolano website,” 27 February 2020. [Online]. Available: https://www.xtal.iqfr.csic.es/Cristalografia/index-en.html.

[175] W. H. Bragg and W. L. Bragg, “The reflexion of X-rays by Crystals,” *Proceedings of the Royal Society London,* pp. 88(605): 428-438, 1913.

[176] A. G. Leslie and H. R. Powell, “Processing diffraction data with MOSFLM,” *Evolving methods for Macromolecular Crystallography,* pp. 41-51, 2007.

[177] W. Kabsch, “XDS,” *Acta Cryst D,* pp. 125-132, 2010.

[178] G. Friedel, “Sur les symétries cristallines que peut révéler la diffraction de rayon Röntgen,” *Comptes rendus,* pp. 1533-1536, 1913.

[179] P. R. Evans, “Scaling and assessment of data quality,” *Acta Cryst. D,* pp. 72-82, 2006.

[180] S. French and K. Wilson, “On the treatment of negative intensity observations,” *Acta Crystallographica D,* pp. 517-525, 1978.

[181] K. Diederichs and P. Karplus, “Better models by discarding data?,” *Acta Crystallographica Section D,* pp. 1215-1222, 2013.

[182] A. McCoy, R. Grosse-Kunstleve, P. Adams, M. Winn, L. Storoni and R. Read, “Phaser crystallographic software,” *Journal of Applied Crystallography,* pp. 40: 658-674, 2007.

[183] U. Arndt, R. Crowther and J. Mallett, “A computer-linked cathode-ray tube microdensitometer for X-ray crystallography,” *Journal of Scientific Instruments,* pp. 1(5): 510-516, 1968.

[184] K. Diederichs and P. Karplus, “Improved R-factors for diffraction data analysis in macromolecular crystallography,” *Nature Structural and Molecular Biology,* pp. 4(4): 269-275, 1997.

[185] G. M. Sheldrick, «Experimental phasing with SHELXC/D/E: combining chain tracing with density modification,» *Acta Crystallographica D,* pp. D66, 479-485. , 2010.

[186] A. L. Boyle, “Applications of de novo designed peptides,” in *Peptide Applications in Biomedicine, Biotechnology and Bioengineering*, Elsevier, 2018, pp. 51-86.

[187] P. Skúbak and N. S. Pannu, “Automatic protein structure solution from weak X-ray data.,” *Nature Communications,* p. 4: 2777, 2013.

[188] J. E. Kohn, I. S. Millett, J. Jacob, B. Zagrovic, T. E. Dillon, N. Cingel, R. S. Dothager, S. Seifert, P. Thiyagarajan, T. R. Sosnick, M. Z. Hasan, V. S. Pande, I. Ruczinski, S. Doniach and K. W. Plaxco, “Random-coil behavior and the dimensions of chemically unfolded proteins,” *PNAS,* pp. 101(34) 12491-12496, 2004.

[189] A. Reményi, K. Lins, L. J. Nissen, R. Reinbold, H. Shöler and M. Wilmanns, “Crystal structure of a POU/HMG/DNA ternary complex suggests differential assembly of Oct4 and Sox2 with two enhancers,” *Genes and development,* pp. 17:2048-2059, 2003.

[190] T. Neudegger, J. Verghese, M. Hayer-Hartl, F. U. Hartl and A. Bracher, “Structure of human heat-shock transcription factor1 in complex with DNA,” *Nature Structure Molecular Biology,* pp. 23(2): 140-146, 2016.

[191] D. Esch, J. Vahokoski, M. R. Groves, V. Pogenberg, V. Cojocaru, H. vom Bruch, D. Han, H. C. A. Drexler, M. J. Araúzo-Bravo, C. K. L. Ng, R. Jauch, M. Wilmanns and H. R. Schöler, “A unique Oct4 interface is crucial for reprogramming to pluripotency,” *Nature Cell Biology,* pp. 15(3):295-301, 2013.

[192] F. V. Murphy, R. M. Sweet and M. E. Churchill, “The structure of a chromosomal high mobility group protein-DNA complex reveals sequence-neutral mechanisms important for non-sequence-specific DNA recognition,” *The EMBO Journal,* pp. 18(23): 6610-6618, 1999.

[193] K. Stott, G. S. Tang, K.-B. Lee and J. O. Thomas, “Structure of a complex of tandem HMG boxes and DNA,” *Journal of molecular biology,* pp. 360: 90-104, 2006.

[194] T. Kanki, K. Ohgaki, M. Gaspari, C. M. Gustafsson, A. Fukuoh, N. Sasaki, N. Hamasaki and D. Kang, “Architectural role of mitochondrial transcription factor A in maintenance of Human Mitochondrial DNA,” *American Society of Microbiology,* pp. 9823-9834, 2004.

[196] H. B. Ngo, G. A. Lovely, R. Phillips and D. C. Chan, “Distinct structural features of TFAM drive mitochondrial DNA packaging versus transcriptional activation,” *Nature Communications,* p. doi: 10.1038/ncomms4077, 2014.

[197] T. A. Brown, A. N. Tkachuk, G. Shtengel, B. G. Kopek, D. F. Bogenhagen, H. F. Hess and D. A. Clayton, “Superresolution fluorescence imaging of mitochondrial nucleoid reveals their spatial range, limits, and membrane interaction,” no. 4994-5010, 2011.