Universitat Autònoma de Barcelona

Institut de Biotecnologia i de Biomedicina

Parc Cienfífic de Barcelona

Institut de Recerca Biomèdica de Barcelona

Instituto de Biología Molecular de Barcelona - CSIC

Biochemical and structural characterization of *Mycoplasma genitalium* proteins MG438 and MG200

TESI DOCTORAL

Bárbara Luísa Machado Calisto

Bárbara L. Machado Calisto

Dr. Ignacio Fita Rodriguez

Barcelona, Juliol de 2010

Para a Inês, o José e o André

Pasmo sempre quando acabo qualquer coisa. Pasmo e desolo-me. O meu instintito de perfeição deveria inibir-me de acabar; deveria inibir-me até de dar começo. Mas distraio-me e faço. O que sigo é um produto em mim, não de uma aplicação da vontade, mas de uma cedência dela. Começo porque não tenho força para pensar; acabo porque não tenho alma para suspender. Este livro é a minha cobardia.

Fernando Pessoa
'Livro do desassossego'

ACKNOWLEDGEMENTS

Sempre chegamos ao sítio donde nos esperam.

José Saramago
'A viagem do elefante'

ACKNOWLEDGEMENTS

Na elaboração de uma tese de doutoramento, tal como nas diferentes etapas da vida, estabelecem-se objectivos mais ou menos ambiciosos segundo o que cada um está disposto a arriscar e a sacrificar para consegui-los. Está claro que estes objectivos se definem de acordo com as experiências, o momento e as oportunidades que se apresentam. De acordo com os resultados que se vão obtendo muitas vezes à que mudar de rumo, tomar novas decisões e seguir sempre o caminho que, de boa-fé, pensamos que é o melhor. É sempre bom lembrar que o caminho não é fácil mas é possível.

Llegado este momento me gustaría agradecer antes que nada a Ignasi:

- por haberme aceptado en su grupo con mucho entusiasmo desde el primer momento,
- por haber contribuido de forma inmejorable a mi formación científica permitiendo mi asistencia a diferentes meetings y workshops,
- por haberme enviado al ESRF tantísimas veces,
- por haberse sentado conmigo delante de la Silicon Graphics, olvidando el tic-tac del reloj, mientras nos mirábamos detalladamente cada enlace en nuestras estructuras,
- y principalmente por permitir que llevemos nuestros proyectos de forma bastante independiente y siempre con total confianza asumiendo el riesgo de que las cosas no salgan bien.

Queria também agradecer aos que me inspiraram a iniciar este projecto independentemente dos resultados conseguidos:

- aos professores da UAlg: Paula Ramos, Jorge Martins e Mª. José Castro,
- a Sandra Macedo-Ribeiro por me ter apresentado à cristalização da lisozima e ter assumido a direcção do meu projecto de final de Licenciatura numas condições em que fazer uma miniprep demorava uma semana e nem sempre com bons resultados,
- a Pedro Pereira com quem aprendi rigor e dedicação,
- ao TóZé por me ter feito ver filmes e ler livros em inglês e principalmente por me ter contagiado o entusiasmo de empreender una tese de doutoramento.

Agradecer a los del laboratorio:

- a Nuria Verdaguer por tan generosamente me haber recibido en su casa cuando llegué a Barcelona y por haber seguido acompañándome en el laboratorio,
- al Jordi Q. y Xavi C., mis primeros mentores, que me hicieron la vida mucho más fácil delante de la pollata, del ordenador, de un artículo y de los malos resultados,

- a los Cri2 de siempre: Arnau, Cristina F., Damià, David, Jordi Q., Oriol G., Queralt, Rosa, Sol, Xavi C.; y a los nuevos: Maria, Mercè y Luca siempre incansables en su apoyo, siempre dispuestos a introducirme en su cultura y a salir de FIESTA, pero principalmente por hacer que cada día en el labo sea una fiesta y, en definitiva, por hacer que no quiera marcharme.
- A los Cri's (a los de siempre y a los nuevos): Anna R., Cristina S., Eshter P., Esther F., Laia, Leonor A., Lionel, Maria G., Maylis, Nereida, Pablo, Robert, Roeland, Silvia, Tiago, … que formando un maravilloso conjunto consiguen ser como esos vecinos que son como de la familia y que están para dejarnos los reactivos y las cubetas para los geles, para discutir resultados y experimentos, para compartir los largos viajes al sincrotrón y con quién se puede contar para ir a conciertos, exposiciones, teatros y desde luego para salir de fiesta.

Agradecer a los del Parc:

- a los miembros de la PAC, sobretodo a la ex-PAC Judith que para mí es mucho más que una técnico ejemplar,
- a Jenny a quién le debo muchos de los éxitos de esta tesis por nunca haberme rehusado el hacer aquel ultimo intento de purificar una proteína sino que, incansable, me ayudaba en ello y me incentivaba a más,
- a Vane que nos enseña a todos a luchar,
- a los Macias: Begoña, Lidia y Román, los ases de la PCR y del buen compañerismo.

Agradecer a los del Institut Pasteur:

- al Dr. Pedro Alzari por haberme abierto la puerta su laboratorio y dejado que accediera por completo al ritmo de trabajo del instituto y a todas sus facilities,
- y también a su entrañable equipo por la fantástica acogida.

Agradecer a los de Barcelona:

- a nuestros colaboradores de la UAB: a los Drs. Jaume Piñol y Enrique Querol por haber apostado por el camino de la determinación de la estructura de proteínas de micoplasma y darnos la posibilidad de trabajar con ellos y con su equipo: Alícia Broto, Oscar Q. Pich y Raul Burgos.
- a nuestros colaboradores de la UB, Departamento de Química Orgánica: al Dr. Jordi Robles por habernos prestado tan generosamente sus conocimientos y su laboratorio para la síntesis de un substrato de un enzima de la vía del MEP,
- a nuestros colaboradores de la UB, Departamento de Bioquímica y Biología Molecular: al Dr. Santiago Imperial y particularmente a sus estudiantes Jordi Pérez y Vitor

Giménez por su dedicación a la producción y purificación de diversos enzimas de la vía del MEP que intentámos cristalizar,

- a mis compañeros de los cursos de doctorado, especialmente a Mertxe, a Nereida, a Paloma, a Silvia, a Sebas y a Pau, por haber permitido que nuestra relación vaya mucho más allá de la ciencia pero que la siga teniendo como punto de unión,
- a mis preciosas brujis por nuestras escapadas de baños de sol y de luna, siempre llenas de musicomagia,
- a la familia de la Ronda de la Vía: Alberto, Alex, Angela, Carol, Jordi, Oriol y Sophie por eso, por ser mi familia de acogida y apaciguar mis ganas de bacalao con sus extraordinarias paellas, fideuàs, fondues, cocidos y caldos. También por aceptarme en sus equipos cuando jugamos a trivial y nos tocan preguntas sobre la monarquía española o celebridades de la televisión y por llevarme de excursión, a la playa y infinitas cosas más!
- molt especialment a la Nere, a en Jordi Q., a l'Oriol G. i a en Damià-Arnau per ser tan generosos, perquè m'han donat els seus amics, les seves families i fins i tot els seus temps. Perquè han estat els meus germans, pares i moltes vegadas, fins i tot, els meus fills i nets.

Agradecimentos à minha família:

- ao André por ter elaborado a ilustração da capa deste livro que seguramente ganha muitíssimo mais por ela e ao Oriol por ter pacientemente corrigido o inglês deste manuscrito,
- à minha avó Maria por me ter ensinado a ser uma mulher independente, preserverante e sempre atenta aos outros,
- ao grande Vicente, à Ana e ao André sempre à espera das minhas visitas relâmpago e por terem sido protagonistas de foi uma das mais maravilhosas peripécias desta tese. Não poderei esquecer aquele telefonema desde a cabine da guest-house do ESRF em que ouvi o Vicente pela primeira vez, acabadinho de nascer!
- Aos meus pais por terem aguentado a distância, as ausências e se terem enfrentado a tudo o que isto supõe mantendo sempre esse sentimento de saudade que nos une,
- Ao Oriol por ter a convicção de que a vida só tem sentido quando se é feliz e por aplicá-la em cada uma das suas acções.

SYMBOLS and ACRONYMS

APRd	Acidic and proline rich domain
ATP	Adenosine triphosphate
ATPase	Adenosine triphosphatase
bp	Base pairs
BSGC	Berkeley Structural Genomic Centre
C	Cytosine
CBB	Coomassie Brilliant Blue
CD	Circular dichroism
CIRCE	Controlling Inverted Repeat of Chaperone Expression
CR	Conserved Region
DLS	Dynamic light scattering
DNA	Deoxyribonucleic acid
dNTPs	Deoxyribonucleotides
ds	Double strand
DSF	Differential scanning fluorimetry
E. coli	*Escherichia coli*
EAGR	Enriched in aromatic and glycine residues
EDTA	Ethylenediaminetetraacetic
EM	Electron microscopy
ESRF	European Synchrotron Radiation Facility
cryo-ET	cryo-Electron Tomography
*fl*MG200	full-length MG200
G	Guanine
GFC	Gel filtration chromatography
HMW	High molecular weight
hsd	Host specificity of DNA
HSQC	Heteronuclear Single Quantum Coherence
IEF	Isoelectric Focusing
IPTG	Isopropyl-β-D-thio-galactoside
kb	kilobases
kDa	kiloDaltons

LB broth	Luria-Bertani broth
M. genitalium	*Mycoplasma genitalium*
MAD	Multiple-wavelength Anomalous Dispersion
MTase	Methyltransferase
MIR	Multiple Isomorphous Replacement
MPD	2-methyl-2,4-pentanediol
MR	Molecular Replacement
MS	Mass spectrometry
NMR	Nuclear Magnetic Resonance
OD	Optical density
ORF	Open reading frame
PAGE	Polyacrylamide gel electrophoresis
PCR	Polymerase chain reaction
PMSF	Phenylmethylsulfonyl fluoride
PVDF	Polyvinylidene fluoride
R-M	Restriction-Modification
rmsd	Root Mean Square Deviation
REase	Restriction Endonuclease
RNA	Ribonucleic acid
SAD	Single-wavelength Anomalous Dispersion
SCT-UB	Scientific-Technical Services of the University of Barcelona
SDS	Sodium dodecyl sulphate
SeMet	Selenomethionine
T_m	Temperature of melting
TAE	Tris-acetate EDTA
TEM	Transmission Electron Microscopy
TRD	Target Recognition Domain
Tris	Tris(hydroxymethyl)aminomethane
TX	Triton X-100

ONE and THREE-LETTER CODES for AMINO ACIDS

A	Ala	Alanine
C	Cys	Cysteine
D	Asp	Aspartic acid
E	Glu	Glutamic acid
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
T	Thr	Threonine
V	Val	Valine
W	Trp	Tryptophane
Y	Tyr	Tyrosine

CONTENTS

ABSTRACT

Mycoplasma genitalium genome was completely unravelled in 1995. Since then, big efforts have been done towards the understanding of its proteome structure, function and evolution. *M. genitalium* is considered one of the ideal research models to define the essential functions for a self-replicating cell due to its small genome size. Despite its apparent simplicity, there are still many open questions about its organization, regulation and infectivity. In particular, just a small percentage of its protein structures are known and none from proteins involved in key cellular processes such as cell division, virulence, pathogeneicity or motility.

In the present work two *M. genitalium* proteins, MG438 and MG200, potentially implicated in the microorganism virulence and/or pathogeneicity processes, were mainly studied from a structural point of view.

The *M. genitalium* ORF MG438 encodes for a type I Restriction-Modification (R-M) S subunit. Type I R-M systems are varied and widely spread in prokaryotes, being involved in the protection of the bacterial host against invading DNA. These are hetero-oligomeric complexes composed by two or three subunits, which exert different functions, one of which being the DNA sequence recognition performed by the S subunit. In *M. genitalium*, the MG438 protein appears as an orphan S subunit suggesting it can have other functions rather than its regular activity in the methyltransferase (MTase) and restriction endonuclease (REase) complexes.

The MG438 crystal structure was determined by the Multiple-wavelength Anomalous Dispersion (MAD) method and refined at 2.3 Å resolution. The three-dimensional structure here described was the first to be determined from a type I R-M S subunit, revealing in detail many aspects of the organization of these subunits. The structure consists of two globular domains of about 150 residues each, separated by a pair of 40 residues long antiparallel α-helices, which form a left-handed super-coil. The globular domains correspond to the variable Target Recognition Domains (TRDs) while the coiled-coil structure correspond to the central (CR1) and C-terminal (CR2) Conserved Regions, respectively. Moreover, the MG438 structure presents an overall cyclic

topology with an intra-subunit two-fold axis that superposes the N- and C-half parts of the molecule, each half containing a TRD and a CR. The straight structural resemblance found to exist between TRDs and the small domain of type II N6-adenine DNA *Taq*I MTase, together with the MG438 structural peculiarities, in particular the presence of the intra-subunit quasi-symmetry, allowed the proposal of a model for the recognition of the target DNAs by the S subunits, which is in agreement with previous experimental data. Furthermore, MG438 can form homo-oligomers which are apparent in the crystal, where two MG438 subunits are related by a crystallographic two-fold axis. The contact area between subunits is quite large involving the symmetric interactions of a cluster of exposed hydrophobic residues.

The comparison between the MG438 structure and the almost simultaneously reported structure of an S subunit from the archae *Methanococcus jannaschii*, S.Mja, highlights the preserved structural features despite the low sequence identity (below 20 %). The comparison also reveals important differences in the TRDs and in their disposition with respect to the CRs.

M. genitalium presents a complex cytoskeleton with a differentiated terminal organelle that is involved in cell adherence and motility. The MG200 multi-domain protein, thought to be localized in the terminal organelle by sequence analysis, was found to be directly involved in mycoplasma motility. Production of the full-length MG200 protein revealed that it behaves as a tetramer in solution but also has a high tendency to aggregate and an intrinsic heterogeneity which prevents crystallization and, consequently, the X-ray crystal structure determination. Therefore, a more detailed study of the protein domains was undertaken. The protein consists of four domains:

i) J-like domain, which presents high sequence identity with other J domains;
ii) an Enriched in Aromatic and Glycine Residues box (EAGRb), which is a well conserved domain found in many proteins that locate in the terminal organelle of mycoplasmas from the *pneumoniae* cluster. Moreover, there is no evidence of these protein's presence in other prokaryotic or eukaryotic organisms;
iii) an Acidic and Proline-Rich domain (APRd), which is also conserved in many terminal organelle proteins;
iv) a C-terminal domain with no detectable sequence identity.

The MG200-EAGRb crystal structure was solved to 2.9 Å resolution by the Single-wavelength Anomalous Dispersion (SAD) method on a crystal from a construct of the domain where a methionine residue had been introduced. This structure is the first piece of structural information, at almost atomic resolution, from any terminal organelle protein and revealed that the domain presents an essentially new fold containing an accurate intra-domain symmetry, which relates with a sequence repeat. The EAGRb forms a dimer, contained in the crystal asymmetric unit, which is stabilized by a conserved hydrophobic core that extends throughout the dimer interface. Some of the domain features, such as its plasticity and the presence and organization of the intra- and inter-subunits symmetry axes, which result in the unbalance of interactions, strongly suggest a role for the EAGRb in protein-protein interactions.

Information on the possible quaternary structures for each of the MG200 protein domains, together with preliminary single particle Electron Microscopy (EM) studies, allowed the proposal of a 222 (D2) symmetry for the full-length MG200 protein. The particular properties of its domains, mainly the J domain and the EAGRb, could permit the interaction with other components of the terminal organelle and result in the formation of supra-molecular structures as the ones observed in the terminal organelles of *M. pneumoniae* cells analyzed by cryo-Electron Tomography (cryo-ET).

SUMARI

El genoma del *Mycoplasma genitalium* va ser completament desentranyat l'any 1995. Des d'aleshores, s'han realitzat grans esforços per a entendre la estructura, funció i evolució del seu proteoma. El *M. genitalium* és considerat un model ideal de recerca per a definir quines son les funcions essencials d'una cèl·lula auto-replicant degut a la petita mida del seu genoma. Tot i la seva aparent simplicitat, encara hi ha moltes preguntes obertes referents a la seva organització, regulació e infectivitat. En particular, només un petit percentatge de les estructures de les seves proteïnes és conegut, essent cap d'elles involucrada en processos cel·lulars clau com la divisió cel·lular, virulència, patogenicitat i motilitat.

En el present treball dues proteïnes del *M. genitalium*, MG438 i MG200, potencialment implicades en la virulència del microorganisme i/o processos patògens, han estat objecte d'estudi, principalment des d'un punt de vista estructural.

El ORF MG438 del *M. genitalium* codifica una subunitat S d'un sistema de Restricció i Modificació (R-M) de tipus I. Els sistemes de R-M de tipus I es troben àmpliament disseminats en cèl·lules procariotes, essent involucrats en la protecció de l'hoste bacterial enfront DNA invasor. Aquests sistemes son complexes hetero-oligòmerics composats per dos o tres subunitats, que exerceixen diferents funcions, essent una d'elles el reconeixement de la seqüència de DNA portat a terme per la subunitat S. En el *M. genitalium*, la proteïna MG438 apareix com una subunitat S orfe, suggerint que pot tenir altres funcions apart de la seva activitat regular en els complexes metiltransferasa (MTase) i endonucleasa (REase).

La estructura cristal·lina del MG438 ha estat determinada usant el mètode de dispersió anòmala a múltiples longituds d'ona (MAD) i refinada a una resolució de 2.3 Å. La estructura tridimensional aquí descrita va ésser la primera en ser resolta per a una subunitat S d'un sistema de R-M de tipus I, revelant en detall molts aspectes de la organització d'aquestes subunitats. La estructura consisteix en dos dominis globulars que consten cadascun d'uns 150 residus separats per un parell de llargues hèlices antiparal·leles d'uns 40 residus cadascuna que forman grans bobines levogires.

Els dominis globulars corresponen als dominis de reconeixement de blancs (TRDs) mentre que les estructures tipus bobina corresponen respectivament a les regions conservades central (CR1) i C-terminal (CR2). A més, la estructura de la MG438 presenta una topologia global cíclica amb un eix binari intra-subunitat que superposa les meitats N- i C-terminal de la molècula, cadascuna contenint un TRD i una CR. La ressemblança entre els TRDs i el domini petit de la *Taq*I MTase de tipus II, juntament amb les peculiaritats estructurals de la MG438, en particular la presència de la quasi simetria de la subunitat, va permetre proposar un model per al reconeixement dels DNAs objectiu per les subunitats S, el qual es troba d'acord amb les dades experimentals prèvies. D'afegit, la MG438 pot formar homo-holigòmers els quals són presents al cristall, on dues subunitats MG438 estàn relacionades per un eix binari cristal·logràfic. L'àrea de contacte entre subunitats és força gran i es forma degut a interaccions simètriques entre els cúmuls exposats de residus hidrofòbics.

La comparació entre la estructura del MG438 i la de una subunitat S pertanyent al arqueobacteri *Methanococcus jannaschii*, S.Mja, publicades pràcticament de manera simultània, posa en relleu les característiques estructurals conservades entre elles tot i la baixa identitat entre les seves seqüències (per sota del 20%) i al mateix temps, posa de manifest importants diferències en els TRDs i la seva disposició en relació als CRs.

El *M. genitalium* presenta un citoesquelet complex amb un orgànul terminal diferenciat que està implicat en l'adherència cel·lular i la motilitat. La proteïna MG200, localitzada en l'orgànul terminal, està directament implicada en la motilitat del micoplasma. La producció de la proteïna MG200 entera va revelar que es comporta com un tetràmer en solució amb una gran tendència a agregar-se i una heterogeneïtat intrínseca que impedeix la seva cristal·lització i, en conseqüència, la determinació de la seva estructura mitjançant difracció de raigs X. Per tant, es va emprendre un estudi més detallat dels dominis de la proteïna. La proteïna consta de quatre dominis:

i) un domini DnaJ, la seqüència del qual presenta gran similitud amb les seqüències d'altres dominis J;
ii) una caixa rica en residus aromàtics i glicines (EAGRb), que és un domini ben conservat present a moltes proteïnes localitzades al orgànul terminal dels mycoplasmes del grup *pneumoniae*. No hi ha cap proba de la presència d'aquestes proteïnes en altres organismes procariòtics o eucariòtics;

iii) un domini ric en residus acídics i prolines (APRd), que també es conserva en moltes altres proteïnes de l'orgànul terminal.

iv) un domini C-terminal sense cap identitat de seqüència trobada.

La estructura cristal·lina de la EARGb de la proteïna MG200 va ser resolta a una resolució de 2.9 Å pel mètode de dispersió anòmala de una única longitud d'ona (SAD) aplicat a un cristall d'un fragment mutant on s'hi va introduir un residu de metionina. Aquesta estructura és la primera peça d'informació estructural, a resolució quasi atòmica, d'una proteïna de l'orgànul terminal i va revelar que el domini presenta un nou plec que conté una simetria interna molt precisa que està relacionada amb una repetició de seqüència. La EAGRb forma un dímer, enclòs en la cel·la unitat del cristall, que és estabilitzat per un nucli hidrofòbic que s'estén per tota la interfície del dímer. Algunes de les característiques del domini, tals com la seva plasticitat i la presència i organització del eixos de simetria interns i entre subunitats, resulten en un desequilibri d'interaccions que suggereixen amb força que el EARGb juga un paper important en les interacciones proteïna-proteïna.

La informació sobre possibles estructures quaternàries de cadascun dels dominis de la proteïna MG200, juntament amb els estudis preliminars de microscòpia electrònica (EM), van permetre proposar la existència de una simetria 222 (D2) per la proteïna MG200. Les particulars propietats dels seus dominis, principalment dels domini DnaJ i EAGRb, podrien permetre la interacció amb altres components de l'orgànul terminal i resultar en la formació de les estructures supra-moleculars observades en l'orgànul terminal de cèl·lules de *M. pneumoniae* analitzades per tomografia cryo-electrònica.

SUMARI

CHAPTER I

General introduction

CHAPTER I. GENERAL INTRODUCTION

1. Mycoplasmas general biological properties

Mycoplasmas form a large group of prokaryotic microorganisms with over 200 species widespread in nature as obligatory parasites of humans, mammals, reptiles, fish, arthropods and plants. Mycoplasmas evolved from Gram-positive bacteria by degenerative or reductive evolution, accompanied by significant losses of genetic sequences. They differ phenotypically from Gram-positives by their small size (with a cell diameter of only 300 nm that is too small to be clearly seen in a light microscope), minute genome, total lack of cell wall and for being the smallest self-replicating prokaryote. Such features attracted considerable attention from life scientists into mycoplasmas molecular biology (Table I.1).

Table I.1. Summary of mycoplasmas general properties compared with eubacteria

Property	Mycoplasma	Eubacteria
Cell wall	No	Yes
Genome size (kb)	~500-2200	>1500
Cell size (µm)	~0.3-1.0	>1.0
Sterol and fatty acid requirement for growth	Yes, but there are exceptions	No
Terminal structures	Terminal organelle in few species	Appendages in few species (e.g. pili)
Fried-egg colony morphology	Yes	Only L-forms

The lack of cell wall is used to distinguish these microorganisms from common bacteria and to include them in a separate class named *Mollicutes* (meaning soft skinned). Most human and animal mollicutes are from *Mycoplasma* or *Ureaplasma* genera of the family *Mycoplasmataceae*. The genus *Mycoplasma* contains more than 200 species, of which 16 are considered to be part of the human tract (Razin 2002; Jensen 2006).

Other particular features of most mycoplasma genomes are the low guanine plus cytosine (G+C) content (within the range of 24-33% G+C) that is unevenly distributed along the genome. Mycoplasmas also present an altered genetic code where the UGA codon is used to code for tryptophane instead of the common STOP signal in other bacteria. Mycoplasmas, especially the species with the smallest genomes, have high mutation rates, suggesting that they are in a state of rapid evolution.

Lacking a peptidoglycan layer, mollicute cells are quite sensitive to lysis by osmotic shock, detergents, and alcohols. These cells are covered with a cytoplasmic membrane which contains membrane-anchored proteins, including surface proteins responsible for antigenic variation, lipoproteins capable of activating host-cell cytokines and cholesterol (uniquely among bacteria).

Mycoplasmas have limited biosynthetic capabilities due to its reduced genome, which results in a lack of genes that code for proteins involved in the biosynthesis of amino acids. Thus, the full spectrum of essential amino acids has to be provided by the host or the artificial medium culture for cell survival. Other biosynthetic routes that were lost by mycoplasmas are the co-factor (meaning that all vitamins have to be supplied for survival) and nucleic acids (purine and pyrimidine bases may be provided *in vivo* by the action of the potent mycoplasmal nucleases that degrade DNA and RNA molecules) biosynthetic pathways. Genes involved in fatty acid synthesis are also missing in the mycoplasma genome and although they are capable of synthesizing their own membrane phospholipids and glycolipids there is a huge uptake of exogenous fatty acids into the membrane. Membrane fluidity regulation is then absent by preferential fatty acid biosynthesis but mycoplasmas overcome it by incorporating large quantities of external cholesterol into the membrane which helps to effectively regulate fluidity (Razin 2002). The limitation or complete loss of certain biosynthetic routes makes mycoplasmas exhibit strict host, organ and tissue specificity, being their primary habitat the mucous surface of respiratory, urogenital and digestive tracts, the eyes, mammary glands and joints in humans and animals. To enter in an appropriate host, where they multiply and survive for long periods of time, mycoplasmas evolved fancy molecular mechanisms to deal with the host immune response and their transfer and colonization into new host cells. Mycoplasmas are then difficult to study by classical genetic tools because of its complex nutritional requirements, poor growth yields, and absence of selectable markers but, despite the evident technical problems, significant progress has been made in the identification of the mechanism by which they interact with eukaryotic host cells.

1.1. *Mycoplasma genitalium*

Mycoplasma genitalium, the central microorganism of this study, is predominantly found in the human urogenital tract. It is a causative organism of non-gonococcal and

non-chlamydial urethritis and pelvic inflammatory disease, being one of the 16 known mycoplasma species that infects humans and the mollicutes species with one of the smallest genomes, 580 kb (Fraser, Gocayne et al. 1995).

M. genitalium is a motile mycoplasma that often presents an asymmetrical pear-like shape due to a characteristic membrane protrusion (Figure I.1.), the terminal organelle. This complex structure was first described in 1970 (Biberfeld and Biberfeld 1970) and is now considered to be the scaffold for mycoplasmas adherence to host cells and gliding motility (Balish 2006; Balish and Krause 2006).

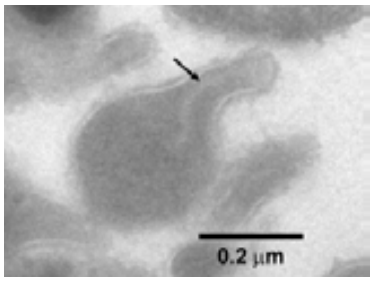


Figure I.1. Electron microscopy (EM) micrograph showing a typical *Mycoplasma genitalium* G37 cell. The arrow indicates the specialized terminal organelle. EM micrograph obtained by Mercè Ratera from Dr. Ignacio Fita laboratory at the Electron Microscopy and *in situ* Molecular Identification Unit of the Scientific-Technical Services of the University of Barcelona (SCT-UB).

1.2. Comparison of the *M. genitalium* and *M. pneumoniae* genomes

M. genitalium is a rather newly discovered microorganism and its genome was one of the firsts to became fully sequenced (Fraser, Gocayne et al. 1995). After the larger *M. pneumoniae* genome (816 kb) was published (Himmelreich, Hilbert et al. 1996) it was shown that these bacterial species are more closely related than anticipated, with a difference of 30 % in genome size, 8 % in DNA G+C content and presenting different tissue specificity. *M. genitalium* genome is considered to be a subset of the *M. pneumoniae* genome but, despite the high similarity level, which allows the correlation of genetic properties between both mycoplasmas, there are genome regions where the similarity is not that big. In these cases, as for the genes encoding for cytoskeleton-like structure proteins, it is worthwhile to do a careful comparison. *M. genitalium* and *M. pneumoniae* protein orthologues have an average identity of 66.1 % at the nucleotide level and of 67.4 % at the amino acidic level. As for the 16S and 23S ribosomal RNAs, the identity is of 98 % (Herrmann and Reiner 1998).

M. pneumoniae was found to have 209 additional open reading frames (ORFs) with respect to *M. genitalium*. Among these, 110 have not been detected and 76

corresponded to homologous genes found as single copies in *M. genitalium* (Himmelreich, Plagens et al. 1997).

1.3. Mycoplasma minimal genome concept

M. genitalium extremely small genome has, consequently, little genomic redundancy compared with the 4.64 Mbp genome of *Escherichia coli*. The first is considered to have a close to the minimal set of genes to sustain bacterial life on Earth, thus being one of the ideal models to determine the essential set of genes needed for cell viability and for synthetic biology studies. In these studies, transposon mutagenesis is used to introduce random insertions into the *M. genitalium* genome and viable cell cultures are further sequenced to identify it. The identification of numerous insertions within a gene suggests that the target gene is non-essential under the defined laboratory growth conditions. The other genes for which no disruptive transposon insertions were found are probably essential for cell growth.

M. genitalium contains 482 protein-coding genes from which 265 are essential (Hutchison, Peterson et al. 1999) and 100 are believed to be non-essential when individually disrupted in the laboratory (Glass, Assad-Garcia et al. 2006). Moreover, 28 % of the essential protein-coding genes encode for proteins of unknown function.

Genome reduction in *M. genitalium* was achieved by the extensive use of operon systems that contributed to a large reduction in the number of regulatory elements needed for gene transcription control. Despite having a reduced genome mycoplasmas still present duplicate, or even multiple, gene copies that seem to be extra to the minimal gene set (Mushegian and Koonin 1996) and that are related with essential physiological needs such as survival in host (Himmelreich, Plagens et al. 1997). Thus, it becomes clear that in order to define a minimal cell, one has to take into account the specific environmental conditions where it must grow.

For a long time now, Dr. Craig Venter's team has been trying to build a minimal cell that contained only essential genes. For this, in 2008, they developed a strategy to produce large DNA molecules that enabled the production of a synthetic *M. genitalium* genome (Gibson, Benders et al. 2008). Though, at this stage, several technical problems had still to be overcome in order to allow the creation of a cell controlled exclusively by

a synthetic genome. Just two years later, the same team was able to accomplish their goal by performing the synthesis, assembly, cloning and transplantation of the 1.08 Mbp *M. mycoides* genome into *M. capricolum* cells (Gibson, Glass et al. 2010). For this study *M. mycoides* was selected instead of *M. genitalium* due to the fact that the latest has an extremely slow growth rate.

1.4. *Mycoplasma genitalium* in the post-genomic era

M. genitalium genome sequencing and its reduced number of protein-coding genes attracted the interest towards this microorganism as an ideal target also for structural genomics initiatives. Later in the year 2000 the Berkeley Structural Genomic Centre (BSGC) was created aiming to determine the complete three-dimensional structural information of the closely related pathogens, *M. genitalium* and *M. pneumoniae*. The initial strategy took them to choose as starting protein targets those predicted to be the most technically 'tractable' and likely to yield new structural and functional information (Kim, Shin et al. 2005). This innovate global strategy led to the development of high-throughput technologies for cloning, protein production, crystallization and three-dimensional structure determination leading to a new structural genomic era. It was seen that most of the processes, from gene to protein structure, can be automated but there are still many steps for which it can be hardly done. In addition, new developments are needed to obtain the structure of a still large number of proteins. By applying this straightforward protocol, without new paths or technologies, structures were determined for only 10 % of the total protein targets. More than two thirds of the solved structures were found to be related with remote homologues of proteins with known structure and function and roughly one half of the proteins with no sequence homologues revealed new folds. At the end of 2005 sixty-nine protein structures have been solved at the BSGC contributing to a 25 % increase in the structural knowledge of *M. genitalium* and *M. pneumoniae* proteomes (Chandonia, Kim et al. 2006).

1.5. *Mycoplasma genitalium* proteome

In a transcendent work from the year 2009 the nearly complete *M. pneumoniae* proteome organization was unravelled (Kuhner, van Noort et al. 2009). The reported results, which were based in proteome wide-screen by tandem affinity purification–mass spectrometry (MS), revealed the existence of 178 soluble protein complexes from which 62 were found to be homo-multimeric while the rest formed hetero-multimers.

This information was further complemented and correlated with experimental data from three-dimensional structural models, single particle EM reconstructions and cellular electron tomograms. In total, almost 90 % of the soluble proteins can form complexes suggesting that this organization can be a general property in bacteria as it is in eukaryotes. Furthermore, it was shown that about one third of the hetero-multimeric complexes can extensively interconnect, suggesting a higher level of proteome organization that includes the possibility of the existence of sequential steps in cellular processes, which could confer multifunctionality. Moreover, such proteome organization is not explained by the genome organization, implying the existence of a much more complex transcriptional regulation than what was previously predicted, resembling the type of organization found in eukaryotes (Guell, van Noort et al. 2009).

2. Mycoplasma pathogenicity

Mycoplasmas can be considered as 'ideal parasites' because they are surface parasites that rarely invade tissues. The infections follow, usually, a chronic course and are only punctually fulminant. The maintenance of mycoplasmas in host cells, their multiplication and survival, show the evolution of their molecular mechanisms to deal with host immune response and to colonize new hosts. The molecular basis of mycoplasma pathogenicity remains largely unknown but tissue damage caused by mycoplasma infections is known to be partially due to the secretion of toxins and harmful metabolites, such as hydrogen peroxide and superoxide radicals. Many components of the host immune system were found to interact with mycoplasmas, inducing macrophage activation and cytokine production. Some mycoplasmal components can act as superantigens (Razin, Yogev et al. 1998).

Mycoplasmas relation with a variety of diseases of unknown etiology has still to be demonstrated as well as its possible oncogenic effects and its role in the induction of apoptosis (Rottem 2003). Human-infecting mycoplasmas are known to enter the cell and persist intracellularly (Jensen, Blom et al. 1994), which seem to explain their capacity to survive after antibiotic treatment (Taylor-Robinson 1996).

2.1. Adhesion to host cells

Mycoplasmas must adhere to host cells for colonization and subsequent infection. The lack of cell wall and cell wall associated components, indicates that mycoplasmas

adhesion mechanism is rather different from those existent in other bacteria and that this process must be related to membrane-associated components, the adhesins. It was already demonstrated that the loss of adhesion by mycoplasma cells implies the lack of infectivity, while the reestablishment of a cytadherent phenotype results in a regain of infectivity (Razin, Yogev et al. 1998). Some of the most intensively studied mycoplasma adhesins are P1, P30 and MgPa from *M. pneumoniae* and *M. genitalium*, which belong to a family of conserved cytadhesins in mycoplasma species that elicit strong immunological responses. Adhesins are used to colonize widely different hosts by attaching to cell membrane receptors as diverse as sialo-oligosaccharides and sulphated glycolipids (Razin and Jacobs 1992), being essential for successful surface parasitism. The adhesins are the most important players of the cytadherence process although there is also a number of accessory membrane proteins involved in it. These proteins are related with components of the cytoskeleton facilitating the movement and concentration of the adhesins at the specialized terminal organelle (Rottem 2003) by which mycoplasma cells first attaches to host cells.

Several of these cytadherence-associated proteins have been identified and characterized and some of them were found to have proline-rich regions and sequence repeats characteristic of eukaryotic cytoskeletal proteins (Krause 1996). Serious efforts are being made to localize these proteins and characterize the interactions that drive to the cytoskeleton network formation given that mycoplasmas pathogenicity may depend on the proper formation of the specialized terminal organelle, which is probably essential for cell infection. However, the events in pathogenesis subsequent to cytadherence are still poorly understood (see Rotem and Yogev, 2000).

The gliding motility mechanism can be also involved in mycoplasmas pathogenicity because of its importance in the spreading of the microorganism from the initial infection site and its association with cell adherence. However, it is another of the mycoplasmas mechanisms that has not been extensively studied (Jordan, Chang et al. 2007).

2.2. Strategies to evade and modulate the host immune system

Successful bacterial pathogens, as mycoplasmas, evolved complex mechanisms to evade host immune cell response such as mimicry of host antigens, survival within host

phagocytes and generation of antigenic variation. As described above, a common survival strategy on pathogens is to avoid the host immune system by varying the antigenic composition of the surface components at high frequency. Mycoplasmas are capable of rapidly change its surface antigenic repertoire by molecular switching events and by spontaneously generate distinct lipoprotein populations. Given this, they can keep a dynamic surface that allows them to survive and quickly adapt to changing environments, avoiding detection by the host's immune system. Another unexpected observation is that the number of mycoplasmal genes involved in diversifying the antigenic nature of the cell surface is very high given the limited genetic information of these microorganisms (Citti and Rosengarten 1997).

Mycoplasma cell membranes are extremely rich in lipoproteins that can act as adhesins and can also be responsible for antigenic variation by acquiring specificities to various host cell receptors or adhesion factors. Like this, they increase the possibilities of the cell population to successfully bind to displayed receptors in the host cell surface, giving the microorganism more chances to adapt to different environments during cell infection and colonization (Herbelin, Ruuth et al. 1994; Kostyal, Butler et al. 1994).

A new genetic approach to search for pathogenicity-associated genes is based on transposon *Tn4001* mutagenesis, which was applied to study pathogenicity factors in *Spiroplasma citri*. In this case, one of the *Tn4001* mutants was non-pathogenic and did not multiply and pathogenicity was only recovered when the mutant was complemented with the wild-type gene (Jacob, Nouzieres et al. 1997). A disadvantage of this approach is the non-random insertion of the transposon in specific hot spots of the genome, but altered methods are being developed for the appropriate target disruption of genes based on homologous recombination (Dhandayuthapani, Rasmussen et al. 1999).

2.3. Virulence

Several potent mycoplasmal virulence factors have been described, like the urease from ureaplasmas which damages the tissues adjacent to living cells by the action of the ammonia resultant from the urea hydrolysis (Ligon and Kenny 1991). Another well known ureaplasma virulence factor is phosholipase A_2 that can release large amounts of arachidonic acid, which inhibits prostaglandin synthesis, an essential component for pregnancy implantation and maintenance (Lamont 1990). Despite the various

mycoplasmal virulence factors that have already been studied there is still no clear relationship between them and pathogeneicity, contributing to the poor understanding of mycoplasmal pathogenic mechanisms.

Mycoplasmas can loose virulence due to mutations in adhesion proteins related with mycoplasmas reduced speeds, suggesting that motility is associated to virulence. Furthermore, adherence is considered the major mycoplasmal virulence factor. The relationship between motility and virulence is still unclear because mutants deficient in gliding are often inhibited in attachment, making the connection between gliding and virulence hard to establish (Hasselbring, Jordan et al. 2005).

2.3.1. Restriction-Modification systems in mycoplasmas

Restriction-Modification (R-M) systems are varied and widespread in prokaryotes and arose early in their evolution. They are multi-component systems with the ability to degrade foreign DNA (restrict) and methylate (modify) unmethylated or hemimethylated DNA, protecting the bacterial hosts from invading genetic material, particularly bacteriophage DNA. The host genome is specifically methylated and is not affected by the restriction endonuclease activity. Otherwise, invading DNA, which usually lacks the specific methylation pattern, is cleaved by the restriction endonuclease (REase). However, in a few cases, the incoming DNA sequences are methylated before recognition by the REase, resulting in a productive phage infection or successful DNA uptake and maintenance (Bickle 1987). In addition, REase expression must be delayed in respect to DNA replication, until the newly synthesized host DNA is appropriately methylated, to avoid undesired self-DNA restriction. In many R-M systems this is accomplished at the level of transcriptional regulation, mediated by the associated R-M controller protein, or C-protein (Ives, Nathan et al. 1992; Tao and Blumenthal 1992).

R-M enzymes are divided in three main categories - types I, II, and III - and their most significant features are summarized in Table I.2.

Table I.2. Main characteristics of the three main R-M systems (adapted from Razin 2002).

	Type I	Type II	Type III
Structural Features	Hetero-oligomeric complex Multifunctional	Independent polypeptides for MTase* and REase‡ activities	Hetero-oligomeric complex Multifunctional
Co-factors	Mg^{2+}, ATP, S-adenosyl methionine	Mg^{2+}	Mg^{2+}, ATP, S-adenosyl methionine
Recognition Site	Bipartite, asymmetric, separated by a non-specific spacer 6-8 bp long	Palindromic	Asymmetric
DNA cleavage	Occurs at sites far away from binding sites. ATP hydrolysis.	Occurs within the recognition sequence	Occurs at fixed distances from the recognition sequence
Enzyme Activities	MTase and REase are mutually exclusive	MTase and REase are independent	MTase and REase are simultaneous

* MTase, methyltransferase; ‡ REase, restriction endonuclease.

The most complex R-M systems are type I enzymes because DNA recognition, methylation and restriction reside in three different polypeptides. The DNA sequence recognition subunit (S) and the methyltransferase subunit (M) form the active methyltransferase complex (MTase), while a restriction subunit (R) is needed to complex with both S and M subunits to form an active REase. Restriction occurs at random sites after translocation of the DNA by adenosine triphosphate (ATP) consumption, as far as 7 kbp apart from the target DNA sequence recognized by the S subunit, upon encountering an obstacle. This type of systems can more easily evolve newer DNA specificities because DNA sequence recognition is independent of other enzymatic activities. Type I R-M systems are further divided into five subtypes (IA, IB, IC, ID and IE) according to the cross-hybridization of genes, antibody cross-reactivity and sequence conservation (Chin, Valinluck et al. 2004).

Type II R-M enzymes were identified during the early 1970s (Kelly and Smith 1970; Roy and Smith 1973) and are the most commercially available. They constitute the simplest systems, consisting of a site-specific REase and a cognate MTase, with each of the activities occurring on independent and single polypeptides. The MTase recognizes the target DNA sequence, binds to it and converts adenosine or cytosine bases on the unmethylated or hemimethylated DNA recognized sequence to 6-methyladenosine (m^6A) or 5-methylcytosine (m^5C), respectively. The REase usually cut within the recognition sequence that is relatively short, 4 to 8 bp, to which it binds specifically.

In type III R-M enzymes specific DNA sequence recognition and methylation are undertaken by the same protein. MTase activity can be carried out independently or as part of an MTase-REase complex, while restriction depends on the formation of this complex to cleave DNA usually at a determined distance from the recognition sequence (25-30 bp). In this case, DNA recognition sequence consists of two inverted repeats of the same DNA sequence (Meisel, Bickle et al. 1992), which produce a loop of DNA, bringing the bound enzymes together for DNA cleavage (Janscak, Sandmeier et al. 2001).

Although a number of R-M systems have been identified in several mycoplasma species, very few of them were subjected to functional studies. *M. arthritidis* strain 1581 is not readily transformable as happens with the H6061 strain. However, it can be transformed with DNA that has been modified *in vitro* using the *Alu*I MTase, which methylates cytosine in the sequence AGCT to m^5C. Genomic DNA of strain 1581 is resistant to cleavage by the *Alu*I restriction enzyme, indicating that the genome is methylated at the AGCT sequence. Therefore, *M. arthritidis* is predicted to have an R-M system that is an isochizomer of *Alu*I that serves as a barrier to transformation (Voelker and Dybvig 1996). Interestingly, some virulent *M. arthritidis* strains have the MAV1 virus, which have genes homologous to MTases and REases, integrated in their genomes (Voelker and Dybvig 1999).

A high complexity type I R-M system was identified in *M. pulmonis* strain KD735 where two *hsdS* genes are flanking *hsdM* and *hsdR* genes. One of the *hsdS* genes, *hsdM* and *hsdR* are organized as a single operon while the second *hsdS* gene is encoded on the other DNA strand and oriented in the opposite direction to that of the other three genes. The production of chimeric *hsdS* genes is then possible via site-specific inversions, resulting in altered restriction specificity, unique feature among the known R-M systems (Sitaraman and Dybvig 1997).

Type I R-M systems persistence in mycoplasmas small genome can be explained by the change in restriction specificity that can result in double-stranded break at random sites in the mycoplasma chromosome, which could promote recombination and lead to evolutionary diversification. Also the unspecific cleavage of foreign DNA by these

enzymes could facilitate its recombination and integration into the mycoplasma genome (Dybvig 1993; Kusano, Sakagami et al. 1997).

3. Cell Division

3.1. Division mode, DNA replication, and chromosome segregation

Mycoplasmas cell division machinery does not resemble much that of cell-walled bacteria but they can divide by binary fission like most of the prokaryotes although other cell division modes cannot be discarded for the more than two hundred known mycoplasma species.

In mycoplasmas, the DNA replication process precedes cell division and must occur almost simultaneously to the duplication of other biomolecules and cytoplasmatic division. The chromosomes are then faithfully segregated into daughter cells, as seen in the case of *M. capricolum* (Seto and Miyata 1999), implying the existence of special mechanisms to promote it as occurs in eubacteria, being the mycoplasma genome attachment to the terminal organelle a possibility for chromosome segregation (Seto, Layh-Schmitt et al. 2001). However, chromosome segregation and cell partition into daughter cells are even less understood in mycoplasmas than in eubacteria. Limited information on mycoplasmas chromosome segregation concerns *M. genitalium* topoisomerase IV that is capable of relaxing super-helical DNA and converting knotted DNA to encircled DNA *in vitro*. This protein is thought to be evolutionarily conserved due to its role in cell division, (Bailey, Younkins et al. 1996).

The last stage of the reproduction cycle consists in the formation of a constricted region followed by the final division of the cell. In conventional bacteria cell septation is thought to be mediated by cell wall components while in cell wall-less bacteria the components that promote this mechanism are unknown, although some mycoplasmal genes were found to be homologous to walled bacteria cell division genes (Fraser, Gocayne et al. 1995; Himmelreich, Hilbert et al. 1996). One of these gene products encodes for the FtsZ protein that forms tubulin-like elements by hydrolysing GTP. In addition, FtsZ localizes at the cell septation site forming a constriction ring between the dividing cells, initiating cell division.

3.2. Duplication of the terminal organelle

The number and location of terminal organelles found in mycoplasma cells has been correlated with chromosome replication, suggesting that terminal organelle duplication precedes cell division (Seto, Layh-Schmitt et al. 2001). Therefore, it is thought to have a role on the initiation of the cell division because bifurcation of the organelle seems to occur in a pre-replicative state of the cell cycle and it might work as a mitotic-like apparatus in mycoplasma. Adhesion and motility of the terminal organelle, which is enriched in adhesion proteins, is thought to be important for cell division once it was seen to replicate, migrate to the opposite sides of cells that are about to divide and then stay at the head of the daughter cells while they separate. Moreover, the terminal organelles of the daughter cells are capable of pulling them apart by means of their gliding motility, which is considered the force that completes cell division.

There are two hypothetical mechanisms by which the terminal organelle could duplicate. The first involves a semi-conservative process in which the bar-like elements of the terminal organelle electron-dense core (see section I.5.2.1) separate and serve as templates for the assembly of duplicate organelles (Hegermann, Herrmann et al. 2002). The second mechanism consists in a *de novo* assembly based on the fact that coreless mutants can reconstitute a fully functional terminal organelle when transformed with the recombinant wild-type proteins that were missing (Krause and Balish 2004).

4. Components of the mycoplasmas cytoskeleton

Non-spherical cell morphology is found for a variety of mycoplasmas suggesting the existence of a cytoskeleton-like structure that supports the membrane and for which numerous experimental evidences have been accumulated over the past 30 years. However, despite the clear presence of a characteristic cytoskeleton, examination of the *M. pneumoniae* genome revealed no genes analogous to the ones that constitute eukaryotic cytoskeletons, suggesting the existence of a novel bacterial cytoskeleton composed of proteins different from those found in eukaryotes. It seems that this cytoskeleton-like structure can function as a substitute for the missing cell wall and that it could provide the necessary scaffold for maintaining and stabilizing the asymmetric shape of mycoplasma cells, allowing the correct formation of the specialized terminal organelle (Feldner, Gobel et al. 1982).

By analogy with eukaryotic cells, where a detergent insoluble fraction of the cells was found to contain a cytoskeleton-like structure, *M. pneumoniae* cells were treated with 0.5-2 % (v/v) Triton X-100 (TX) in order to determine mycoplasmas cytoskeleton or cytoskeleton-associated elements and as a starting point to identify proteins that are potentially relevant for cytadherence (Regula, Boguth et al. 2001). From this experiment an undefined, insoluble and structured protein complex enriched in about 100 proteins remained, called the TX insoluble fraction. These proteins were further identified by bidimensional gel electrophoresis coupled to MS and by immunoblotting. The most abundant proteins found in the TX insoluble fraction of *M. pneumoniae* cells were the house keeping genes DnaK (a heat-shock protein), elongation factor Tu, FtsH and the redox enzyme thioredoxin. All these proteins have the capacity to form multimers and/or to transiently interact with aggregates or TX insoluble proteins, which may contribute to their partial insolubility. The identity of the cytoskeleton filaments is particularly interesting because, of the known bacterial cytoskeletal proteins, only FtsZ (MPN317) has been identified in the *M. pneumoniae* TX insoluble fraction and it is known that in eukaryotes FtsZ form a linear homo-polymer that accounts for its insolubility in TX. Elongation factor Tu (MNP665) has been shown to form filaments *in vitro* under appropriate conditions (Beck, Arscott et al. 1978) becoming a potential candidate to be part of the cytoskeleton-like structure. Furthermore, two genes of unknown function, *yabB* and *yabC*, were identified in *M. pneumoniae* and *M. genitalium* genomes, which belong to the cluster of genes involved in cell division in other organisms. In mycoplasmas, these genes appear in the *ftsZ* associated operon and were identified in the TX insoluble fraction of *M. pneumoniae*, suggesting a possible presence in the mycoplasma cytoskeleton.

Proteins involved in folding, such as DnaK, GroEL and trigger factor, were also identified in the *M. pneumoniae* TX insoluble fraction probably because of their association with other proteins and given that they are highly abundant in mycoplasmas. Thus, their presence in this fraction may not be related to their presence in the mycoplasmas cytoskeleton, although DnaK (MPN434, also known as Hsp70), which is a protein chaperone structurally homologous to actin (Flaherty, McKay et al. 1991), was shown to be associated with the P1 adhesin by chemical cross-linking (Layh-Schmitt, Podtelejnikov et al. 2000).

Mycoplasmas terminal organelle particular cytoskeletal structures and constituent proteins will be extensively discussed in later sections.

5. Gliding motility

Mycoplasma cells were shown to move in the 70's (Bredt 1973) and, to date, there are 13 mycoplasma species (*M. hominis*, *M. mobile*, *M. pulmonis*, *M. penetrans*, *M. agassizii*, *M. testudineum*, *M. pneumoniae*, *M. testudinis*, *M. amphoriforme*, *M. gallisepticum*, *M. imitans*, *M. genitalium*, and *M. pirum*) that have been reported to glide. These cells were found to glide on solid surfaces, such as glass, plastic and animal cells, and present no flagella, pili or other conventional motor proteins. From these species five genomes have already been sequenced and no genes related to other bacterial motility systems, neither genes homologous to eukaryotic motor proteins were identified (Dandekar, Huynen et al. 2000; McBride 2001). Therefore, these observations imply that mycoplasmas motility mechanism has to be distinct from any other known.

Mycoplasmas terminal organelle was demonstrated to be the cells molecular motor by the analysis of a *M. pneumoniae* mutant in ORF MPN311. This mutant revealed that 'run-away' terminal organelles could exist that stretch the cell when it gets stuck, continue gliding and stretch the cell until it breaks, leaving it motionless while the leading end keeps on moving forward for more than 30 minutes (Hasselbring and Krause 2007).

Molecular and genomic studies suggest that the machineries responsible for cell movement differentiate mycoplasma species in two main groups: the fast and the slow gliders. Proteins involved in the gliding motility system of *M. mobile* (main representative of the fast gliders) were found to share no amino acidic similarity with the motility system proteic components of *M. pneumoniae* (main representative of the slow gliders, or with any ORFs of its genome (Jaffe, Stange-Thomann et al. 2004). Moreover, no homologues of the *M. mobile* terminal organelle ultrastructures were found in *M. pneumoniae* (Uenoyama and Miyata 2005). Also *M. gallisepticum* have a cell morphology distinct from that of *M. pneumoniae* despite having some ORFs homologous to proteins that compose its cytoskeleton (Nakane and Miyata 2009). *M. amphoriforme* presents a unique combination of features with an overall morphology that resembles that of *M. gallisepticum* and an internal ultrastructure more similar to its

relative *M. pneumoniae* (Hatchel, Balish et al. 2006). *M. insons* was recently characterized as having a distinct cellular organization when compared to other motile mycoplasma species, lacking a differentiated terminal organelle and the associated cytoskeletal extension while conserving a polar distribution of adhesins (Relich, Friedberg et al. 2009). Altogether, these facts demonstrate a non-predicted high degree of ultrastructural complexity among mycoplasma species with substantial differences in cell and terminal organelle dimensions, despite the large phylogenetic relatedness between species (Hatchel and Balish 2008). Cell polarity and varied gliding machinery among mycoplasma species suggest that motility evolved more than once and in multiple independent steps in this group of bacteria.

5.1. Gliding motility in the fast gliders

5.1.1. Head-like ultrastructure and architecture

M. mobile is the known fastest gliding mycoplasma (gliding on glass at an average speed of 2.0 to 4.5 μm/s). The analysis of *M. mobile* EM micrographs reveals the presence of a unique ultrastructure at the cell head, designated the *jellyfish* structure, shaped as an oval bell filled with a hexagonal lattice of 12 nm periodicity. Connected to this structure, at the level of the cell neck, there are dozens of *tentacles* covered with particles of 20 nm in diameter at intervals of 30 nm (Figure I.2A). This structure was suggested to be related with the cell gliding machinery due to its co-localization with the gliding motility proteins and by the analysis of mutants lacking the essential gliding proteins, which presented a disordered *jellyfish* structure (Nakane and Miyata 2007).

5.1.2. Proteins of the head-like structure

Four major proteins, which are clustered at the cell neck and are responsible for the correct positioning of the other gliding proteins, have been also directly related with *M. mobile* gliding motility. These proteins, which will be only briefly discussed in here, are coded in tandem in the genome. None of them has significant sequence similarities with proteins from other bacteria, except for orthologues from closely related species. After their identification and from the analysis of gliding mutants and recombinant proteins a role was assigned to each of them (Figure I.2A):

i) Gli123 (123 kDa) was given the role of *mount*, providing the sites for the proper localization of the other motility proteins into the cell neck (Uenoyama and Miyata 2005);
ii) Gli 349 (349 kDa) was determined to be the *leg* (Uenoyama, Kusumoto et al. 2004);
iii) Gli521 (521 kDa) is considered the *gear* (Seto, Uenoyama et al. 2005);
iv) P42 (42 kDa), a nucleoside triphosphatase enzyme, should be the system *motor* given the ATPase activity showed by the recombinant protein.

5.1.3. Centipede model for gliding motility

An hypothetical model for *M. mobile* gliding motility, called the centipede model (or power stroke model), was proposed by Miyata (Miyata 2008). In this model the *leg* proteins, present in high number in the cell neck and that are probably connected to the cytoskeleton, are thought to propel cell movement in a centipede-like manner. Gliding motility would be droved by the alternately binding and unbinding of the *legs* to a solid surface and powered by the force generated by ATP hydrolysis (Jaffe, Miyata et al. 2004; Uenoyama and Miyata 2005). Each gliding unit is in a given mechanical state that is sequentially changing following the transition steps: pull (stroke), movement, release, return, initial binding and tight binding (Figure I.2B).

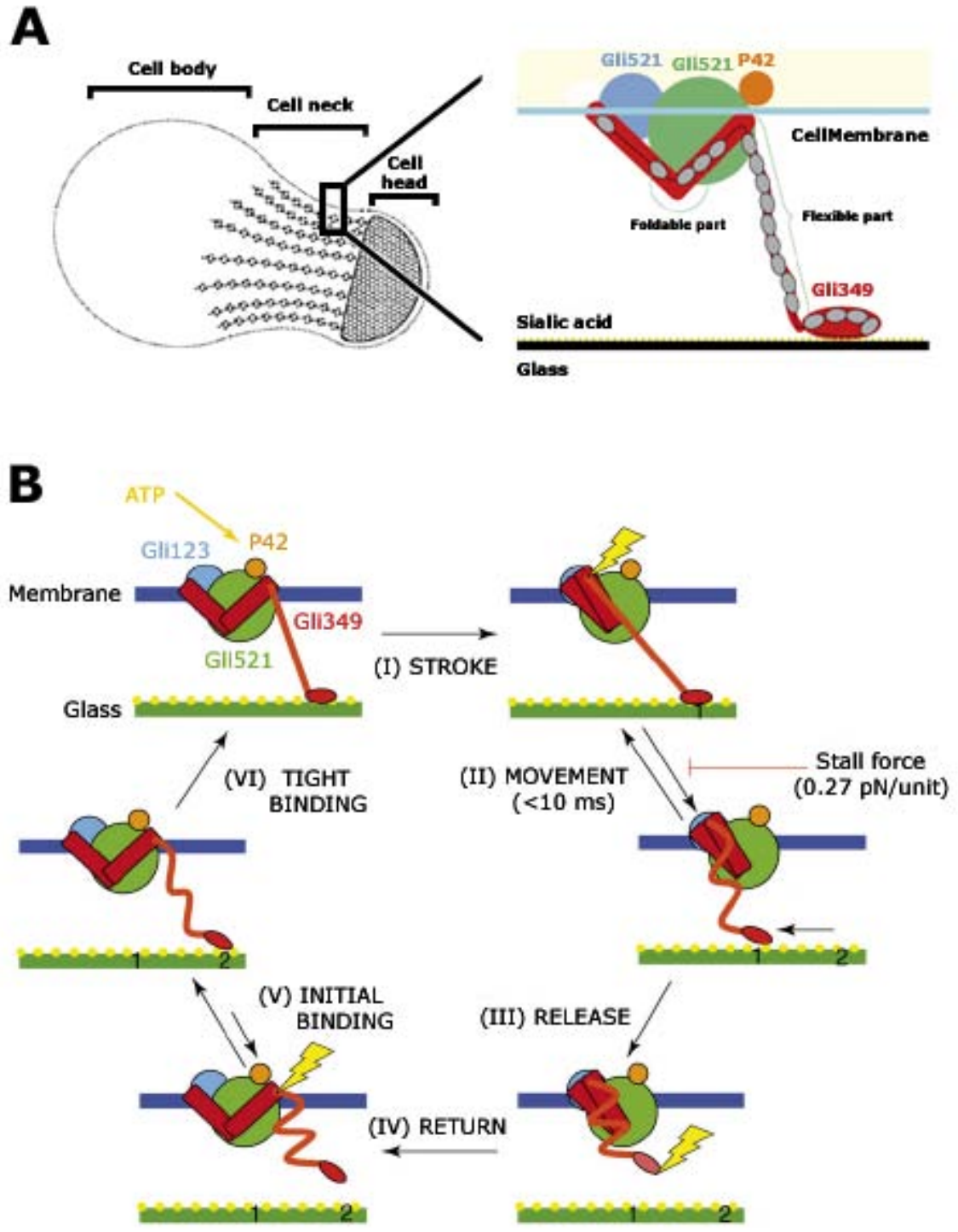


Figure I.2. Centipede model for *Mycoplasma mobile* cell architecture and gliding. (A) Schematic illustration of three subcellular regions of *M. mobile*. The neck region is specialized on gliding containing the gliding units, composed by four distinct proteins that localize near to the cell membrane and are aligned along this subcellular region. The right panel shows a single gliding unit including a scheme of Gli123 (*mount*), Gli349 (*leg*), Gli521 (*gear*) and P42 (*motor*) proteins. (B) Hypothetical centipede model for *M. mobile* gliding motility. The cycle of transition of a gliding unit occurs in six steps numbered I-VI. The 'tension-sensitive' action of the gliding machinery is marked by the yellow lightning bolts. The figure was modified from Nakane et al. 2007, and Miyata 2008.

5.2. Gliding motility in the slow gliders

5.2.1. Terminal organelle ultrastructure and architecture

Some of the proteins found in *M. pneumoniae* TX insoluble fraction are thought to be involved in cytadherence. Furthermore, the ultrastructures observed in EM micrographs of *M. pneumoniae* intact cells were detected in *M. pneumoniae* TX insoluble fractions suggesting that these structures persist to the detergent treatment.

Cryo-electron tomography (cryo-ET) is a suitable technique to produce three-dimensional images of intact pleiomorphic cells no thicker than half a micron in a close to native state (due to mycoplasma cells size) and it does not rely on the averaging of structures, in contrast to other reconstruction methods. In cryo-ET, suspensions of intact cells are spread into thin films across EM grids, plunge-frozen in liquid ethane and imaged iteratively while being tilted (Li and Jensen 2009). Mycoplasmas are particularly difficult samples for EM studies because, due to the lack of cell wall, their cell body can be easily deformed and their cytoplasmic membrane can disrupt during sample preparation. Despite this, recent cryo-ET studies of mycoplasma cells contributed to important advances in the determination of the terminal organelle ultrastructure (Henderson and Jensen 2006; Seybert, Herrmann et al. 2006).

A cytoskeletal network was observed in several cryo-ET studies of mycoplasma cells, especially at the terminal organelle. This structure is a conformationally flexible multi-subunit dynamic motor surrounded by a translucent area and a membrane, which presents organized surface proteins. The terminal organelle, thought to be stabilized by this intracellular cytoskeleton-like structure, consists of three parts (Figure I.3):

i) a terminal button, located at the distal end of the terminal organelle, that is wider than the rest of the core and that connects it with the cell membrane;
ii) a ~300 nm long electron-dense core, placed longitudinally at the centre of the terminal organelle, composed by two segmented paired plates (the rods) which are formed by quasi-periodical macromolecular complexes;
iii) a wheel complex, located at the proximal end of the terminal organelle, close to the cell body.

M. genitalium terminal organelle terminal button and electron-dense core structures resemble the *bell* and *tentacle* structures of the terminal organelle *jellyfish* structure of *M. mobile* (Nakane and Miyata 2007).

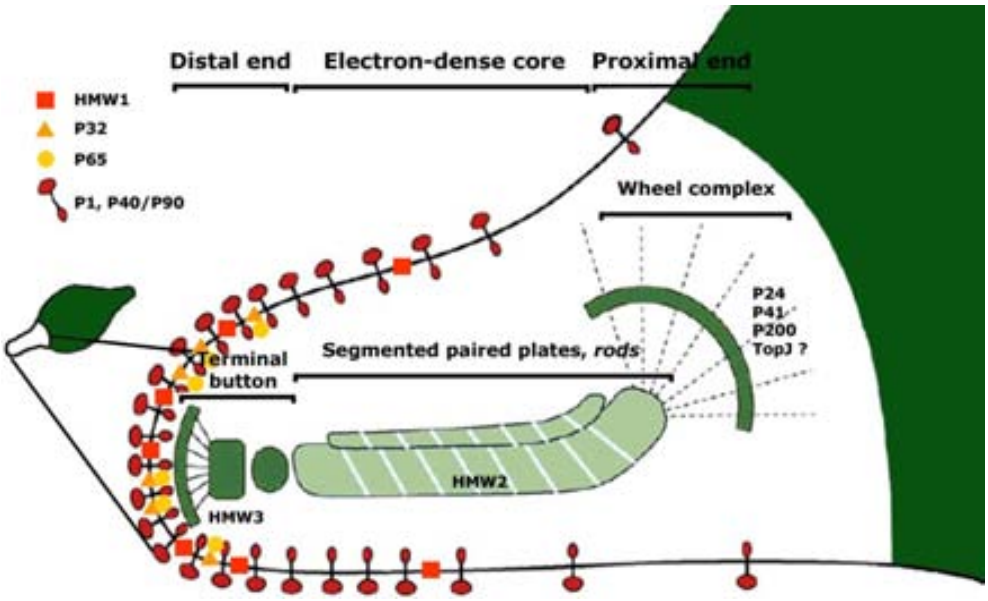


Figure I.3. ***Mycoplasma pnemoniae*** **terminal organelle architecture.** Schematic illustration of the terminal organelle with an outline of the entire cell on the left. The cytoplasm is coloured in green. The terminal organelle intracellular cytoskeleton-like structure consists of three parts: the distal end, the electron-dense core and the proximal end. P1, P40/P90 surface protein complexes are indicated as well as the proteins localized in each of the terminal organelle parts: HMW1, P32, P65, HMW3, HMW2, P24, P41, P200 and TopJ. The figure was modified from Miyata 2008.

The terminal organelle surface protein complexes are well organized on the organelle extracellular surface and densely clustered near its front end. These large complexes form packed rows of 5.5 nm thickness that extend from the terminal button to the end of the electron translucent area (Henderson and Jensen 2006; Kuhner, van Noort et al. 2009). Their extracellular domain has a length of about 16 nm, while the intracellular domain extends ~24 nm into the cytoplasm.

The electron-dense core is constituted by two morphologically different segmented-paired plates, the rods, which are highly flexible. The rods have different thicknesses and lengths, are bent ~150º proximal to their midpoint and are separated by a gap of ~7 nm. The outer rod thickness varies in a range of 13 to 31 nm while the inner rod thickness is of ~8 nm. The outer rod is also longer than the inner one and can be contacting the terminal button. The electron-dense core is surrounded by an intriguing

translucent area deprived of large macromolecular complexes although there is no visible barrier that to justifies it.

5.2.2. Terminal organelle proteins

To date, 11 distinct proteins have been localized into the terminal organelle. They are generally involved in cell adhesion and motility processes and are also responsible for the terminal organelle proper functioning (Table I.3). Furthermore, all of them are coded in three unlinked loci in the genome, together with other proteins of unknown function. Adhesins form the protein clusters found across the whole surface of the terminal organelle while the others are aligned from the distal to the proximal end (Figure I.3).

Among the known terminal organelle proteins are the *M. pneumoniae* adhesins P1, P30, P40/P90 (expressed as a single peptide from ORF MPN6 gene downstream of the P1 adhesin gene and processed into two mature proteins) and the high molecular weight (HMW) proteins 1-3. Lack of any one of these results in non-adherent cells that lack the cytoskeleton-like structure or have it malformed (Krause 1996). Furthermore, chemical cross-linking studies suggested that P1 may physically interact with P30, P40/P90, HMW1, HMW3 and P65 proteins, suggesting that they are forming a large working complex or are in close proximity in the cell.

Table I.3. Terminal organelle structural and cytadherence-associated proteins.

MG proteins		MPN proteins		Sequence Identity (%)	Subcellular Localization†	Structural features				
ORF	Name	MPN no.*	Name			TM‡	CC	PR	APRd	EAGRb
191	P140	141	P1	45.4	S	5	-	+	-	-
192	P110	142	P40/P90	50.2	S	4	-	-	-	-
200	MG200	119	TopJ	34.9	-	-	-	+	+	1
217	MG217	309	P65	41.6	T	-	+	+	+	-
218	MG218	310	HMW2	57.0	R	-	+	+	-	-
218.1	MG218.1	311	P41	53.7	W	-	-	-	-	-
219	MG219	312	P24	15.9	W	-	-	-	-	-
312	MG312	447	HMW1	32.9	R	-	+	+	+	1
317	MG317	452	HMW3	33.5	t/r	-	-	+	+	-
318	P32	453	P30	43.2	T	1	-	+	-	-
386	MG386	567	P200	29.7	W	-	-	-	+	6

MPN number according to the *M. pneumoniae* genome re-annotation (Dandekar, Huynen et al. 2000). † Subcellular localization in *M. pneumoniae*: *S*, surface of the terminal organelle; *T*, terminal button; *R*, rod; *W*, wheel complex. ‡ TM, predicted transmembrane segment; CC, predicted coiled-coil structure; PR, peptide repeat (4-10 amino acids in most cases); APRd, acidic and proline-rich domain; EAGRb, enriched in aromatic and glycine residues box. Adapted from Regula et al. 2001 and Burgos et al. 2009.

P1 adhesin is a 170 kDa protein with several transmembrane segments, presumably associated with P40/P90 adhesins. It is the major terminal organelle adhesin, responsible for mycoplasmas binding to solid surfaces (probably through sialyllactose or sulphated glycoplipids) and gliding motility (Seto, Kenri et al. 2005).

HMW1, HMW3, P65 and P200 accumulate in the *M. pneumoniae* TX insoluble fraction and present unusual amino acidic sequence compositions with Acidic and Proline-Rich domains (APRd). P65 N-terminal region was predicted to form an extended coiled-coil structure (Proft, Hilbert et al. 1995; Proft, Hilbert et al. 1996).

HMW2 is a 215 kDa protein composed of repetitions of seven amino acids that have hydrophobic residues in positions 1 and 4 and hydrophilic residues in positions 3, 5, 6 and 7. This periodicity usually gives rise to regular α-helical structures that lead to the formation of α-helical coiled-coil structures. These are characteristic of cytoskeletal proteins filamentous domains, suggesting that HMW2 can be the major terminal organelle structural element (Krause, Proft et al. 1997).

HMW1, a 112 kDa protein, is composed of an APRd, an Enriched in Aromatic and Glycine Residues box (EAGRb) and two predicted coiled-coil motifs and is known to associate with the terminal organelle cell surface (Proft, Hilbert et al. 1996).

Disruption of *M. pneumoniae* MPN119 gene, by insertion of the *Tn*4001 transposon, resulted in a cytadherence defective mutant from which the TopJ corresponding protein was missing (Cloward and Krause 2009). TopJ have a predicted well known J domain, an EAGRb and an APRd.

Localization of some of the terminal organelle proteins was performed by fluorescence microscopy. HMW1 and HMW2 proteins localize in the flexible electron-dense core region and are required in early stages of the organelle formation (Razin 2002). In turn, P65 and P30 co-localize and form the terminal button and the boundary of the electron-dense core structure, at the distal end of the terminal organelle, which is surrounded by HMW3 (Seto and Miyata 2003; Kenri, Seto et al. 2004). The extracellular C-terminal region of P30 has multiple tandem proline-rich repeats. P41, P24 and P200 are localized in the wheel complex (Kenri, Seto et al. 2004), at the proximal end of the terminal

organelle, and might have a role in connecting the organelle with the cell body. TopJ is also hypothetically localized in the wheel complex because it shares several domains with proteins that localize in this region of the cell but there is still no experimental evidence that confirms it.

Since it was not easy to differentiate between adhesins and terminal organelle structural proteins the term cytadherence-associated proteins was adopted as a generic term to mention them (Krause 1996).

5.2.3. Terminal organelle assembly

The study of non-cytadherent mutants was crucial for the identification of the terminal organelle components. Their phenotypes included altered cell morphology, improper protein localization and instability of specific proteins. Subcellular localization of cytadherence-associated proteins in wild-type and cytadherence-deficient mutant cells, by immuno-EM revealed the order by which the terminal organelle proteins assemble during the terminal organelle formation. The results point out that HMW1 may be the first protein to be translocated, followed by HMW3, P1, P30, P40/P90 and finally P65, which is thought to be the last component to assemble into the terminal organelle. HMW3, P1, P30 and P40/P90 proteins seem capable to assemble independently of each others (Seto, Layh-Schmitt et al. 2001).

5.2.4. Inchworm model for gliding motility

The motile mycoplasma species that are closely related to *M. pneumoniae* cells are slimmer and more flexible than those of *M.* mobile and glide sporadically with an average speed of 0.3-0.5 μm/s (~10x slower than *M. mobile* cells).

The inchworm model for mycoplasma gliding motility is based on the fact that mycoplasmas are capable of using their terminal organelle to attach to and move invariably in its direction through solid surfaces. This model proposes that the terminal organelle terminal button and wheel complex are bound to the cell membrane, or any membrane associated-molecule, and that the electron-dense core, which has many flexible parts and a bend, could change in length and curve causing repeated dissociation, displacement and association of membrane-surface molecules with the solid surface, propelling the cell body (Henderson and Jensen 2006; Seybert, Herrmann

et al. 2006). Furthermore, the conformational changes that can undo the terminal organelle filamental proteins are thought to produce the forces that drive propulsion and cell division (Wolgemuth, Igoshin et al. 2003). These forces could drive to cell motion when coupled to ATP hydrolysis and the dynamic movements of the active elastic filaments could generate the mechano-chemical deformations needed for propulsion (Figure I.4).

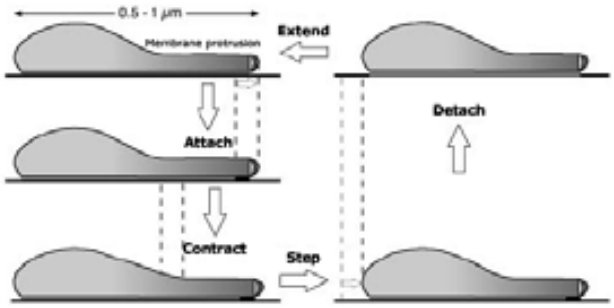


Figure I.4. Inchworm model for *Mycoplasma pneumoniae* gliding motility. The filamental terminal organelle proteins at the front of the cell attach to the solid surface and contract pulling the cell body forward. Then, the membrane protrusion detaches and extends completing the mechano-chemical cycle (adapted from Wolgemuth, Igoshin et al. 2003).

OBJECTIVES

OBJECTIVES

Mycoplasma genitalium was first identified in the 80's and since then many efforts have been done in order to completely understand this microorganism, which is considered to have the minimal set of genes to sustain bacterial life on Earth. Despite all the investigations, data on mycoplasmas virulence, pathogeneigity and motility mechanisms is still limited and relationships between these processes are hard to establish.

The present thesis focuses on the biochemical and structural characterisation of *M. genitalium* MG438 and MG200 proteins. MG438 is an orphan type I R-M S subunit and it is already known that the existence of such orphan subunits correlates well with virulence. Likewise, mycoplasma motility has also been strongly associated with virulence although the main elements participating in this process are still largely unknown in terms of their structural and functional properties. *M. genitalium* MG200 protein likely localizes in the terminal organelle, which contains all the motility machinery.

The strong indications that MG438 and MG200 proteins can be related with *M. genitalium* virulence and/or pathogeneicity mechanisms, together with the fact that these proteins represent important targets for structure determination as models to the study of protein homologues, led to the proposal of the following objectives:

Chapter II. MG438, an orphan type I R-M S subunit

i) Determine the crystal structure of the MG438 protein;
ii) Establish the bases for DNA recognition in type I R-M S subunits;
iii) Bring some light on other possible functions for orphan type I R-M S subunits.

Chapter III. MG200, a terminal organelle motility protein

i) Characterise biochemical and biophysically the MG200 protein;
ii) Determine the MG200 protein crystal structure;
iii) Characterise biochemical and biophysically the MG200 soluble domains;
iv) Determine the MG200 EAGR box crystal structure;

v) Unravel the role of MG200 on *M. genitalium* gliding motility and on the formation of the terminal organelle ultrastructures.

CHAPTER II

MG438, an orphan type I R-M S subunit

CHAPTER II. MG438, an orphan type I R-M S subunit

1. Introduction

Restriction-modification systems were described earlier on section I.2.3.1. Briefly, they are hetero-oligomeric complexes that catalyse both the restriction and specific modification of DNA. In the present work special attention will be given to type I R-M systems since this chapter interest protein constitutes a type I R-M S subunit.

Complete type I R-M systems comprise three different subunits: HsdR (R), HsdM (M), and HsdS (S), coded by three closely linked genes that are named *hsd*, as acronym for host specificity of DNA (Murray 2000). The three subunits can assemble into complexes in order to present MTase and/or REase activities. The largest complex, which assemble with stoichiometry $R_2M_2S_1$, have MTase and REase activities, while the smaller M_2S_1 complex shows only MTase activity.

All type I R-M systems allocated into a family are very closely related illustrating only a significant diversification within the specificity gene. The S subunit is responsible for the DNA sequence recognition on both MTase and REase complexes. In general, both complexes recognize specifically an asymmetric nucleotide target composed of two components of 3-4 and 4-5 bp that are separated by a 6-8 bp long unspecific spacer (Glover and Colson 1969).

Functional type I R-M systems were identified in a large variety of bacteria such as *Bacillus subtilis* (Xu, Willert et al. 1995), *Citrobacter freundii* (Daniel, Fuller-Pace et al. 1988), *Klebesiella pneumoniae* (Valinluck, Lee et al. 1995; Lee, Rutebuka et al. 1997), *Lactococcus lactis* (Schouler, Clier et al. 1998), *Mycoplasma pumonis* (Dybvig and Yu 1994), *Pasteurella haemolytica* (Highlander and Garza 1996) and *Staphylococcus aureus* (Sjostrom, Lofdahl et al. 1978) in addition to the firstly characterized R-M systems that were from *Echerichia coli* and *Salmonella enterica* (Bullas, Colson et al. 1980).

An overview of several genome databases reveals that most of the bacterial genomes encode for the three subunits of the type I R-M systems. However, incomplete *hsd* gene

systems were found in some bacteria as *Helicobacter pylori* (strains 26695 and J99), *Mycobacterium tuberculosis* (strain H37Rv) and in several mycoplasma species, including *M. genitalium*, which contained only two or one solitary genes, mainly *hsdS*. The intriguing presence of these type I R-M orphan S subunits among bacteria with reduced genomes can suggest that they could have a multifunctional role in addition to their regular activity in the MTase and REase complexes. Some genome-wide DNA microarray analysis reinforced this idea when comparison between *Helicobacter pylori* (Bjorkholm, Guruge et al. 2002) and *Francisella tularensis* (Bjorkholm, Guruge et al. 2002), virulent and attenuated strains, showed a correlation between the absence of *hsdS* genes with the loss of virulence, even when the organisms already lacked the related *hsdM* and *hsdR* genes (Alm and Trust 1999).

The ubiquitous maintenance of R-M systems in bacteria is usually explained by their protective role against external DNA. However, the loss of some R-M systems was found to lead to cell death, most likely through restriction enzyme attack on unmethylated recognition sites in newly replicated chromosomes. This process, which belong to a phenomenon known as 'post-segregational host killing' or 'genetic addiction' (Kobayashi 2001), results in the stable maintenance of these R-M systems in a population of variable cells. This finding led to the hypothesis that some R-M systems may behave as 'selfish' genetic elements, as do transposons or viral genomes (Kusano, Naito et al. 1995; Naito, Kusano et al. 1995).

S subunits primary sequences revealed that these proteins contain a central (CR1) and a C-terminal (CR2) Conserved Regions, which main role can be the interaction with the M subunits (Dryden, Sturrock et al. 1995). Furthermore, it contains two highly variable regions that alternate with the CRs, which are known as Target Recognition Domains (TRDs). These define the specificity for the DNA target sequences, each TRD recognizing one half of the bipartite DNA site independently (Glover and Colson 1969).

All these interesting but still little understood data on orphan type I R-M S subunits prompt to new researches on the subject. In here, structural studies were carried out on the protein encoded by the *M. genitalium* ORF MG438. This ORF is currently annotated as an orphan S subunit based on amino acidic sequence similarity (COG0732) and on domain architecture, which is in accordance with what is defined in Pfam (Finn, Mistry

et al. 2010) for type I R-M S subunits (motif 01420). Neither M nor R subunits have been identified in the *M. genitalium* genome.

2. Results and Discussion

2.1. MG438 purification and crystallization

The MG438 protein was purified to approximately 95 % through a Ni^{2+}-affinity column in a one step protocol. Small protein crystals were readily obtained in 0.1 M tri-sodium citrate pH 5.5 with 2.0 M ammonium sulphate (Figure II.1A-*a*). Crystal growth was easily optimized and large and thin needles were obtained but, apparently, were not unique crystals given that smaller needles were observed to grow from the largest ones (Figure II.1A-*b*). These crystals did not diffracted when tested in the in-house rotating anode X-ray generator Rigaku 007 (Automated Crystallography Platform of the Barcelona Scientific Park) even when frozen in presence of different cryoprotectants. Some crystals were then extracted from the crystallization drops, cleaned in fresh crystallization solution and analysed by SDS-PAGE. The silver-stained denaturant polyacrylamide gel showed a discrete protein band with an apparent molecular weight identical to the one expected for MG438, confirming that the obtained crystals were indeed from the interest protein (Figure II.1B). Diffraction data were collected to 7 Å resolution when using synchrotron radiation. The crystals unit cell parameters were a=b=113.8 Å, c=420.2 Å, α=β=90°, γ=120° and upon indexing they were shown to belong to the $P6_x$ space group.

Meanwhile, crystals with a different morphology grew in a distinct condition (Figure II.1A-*c*) and further optimization yielded even bigger crystals (Figure II.1A-*d*). However, they were also more fragile and very sensible to manipulation. Despite this, crystals grown in high concentrations of NaCl diffracted to 4.5 Å resolution at the European Synchrotron Radiation Facility (ESRF) ID13 beam line and were shown to belong to the trigonal space group $P3_121$. Further attempts were done in order to improve crystal growth and quality by testing the effect of adding small organic compounds to the crystallization drops (Hampton Research additive screen) or microseeding. Larger crystals, with a more defined morphology, were obtained in around the same conditions but now adding 0.01 M ethylenediaminetetraacetic (EDTA) disodium salt dihydrate to the crystallization drop (Figure II.1A-*e*). Native protein crystals grown in these very fine-tuned conditions proved to diffract further than the former ones and atomic resolution, nearly 3 Å, was achieved.

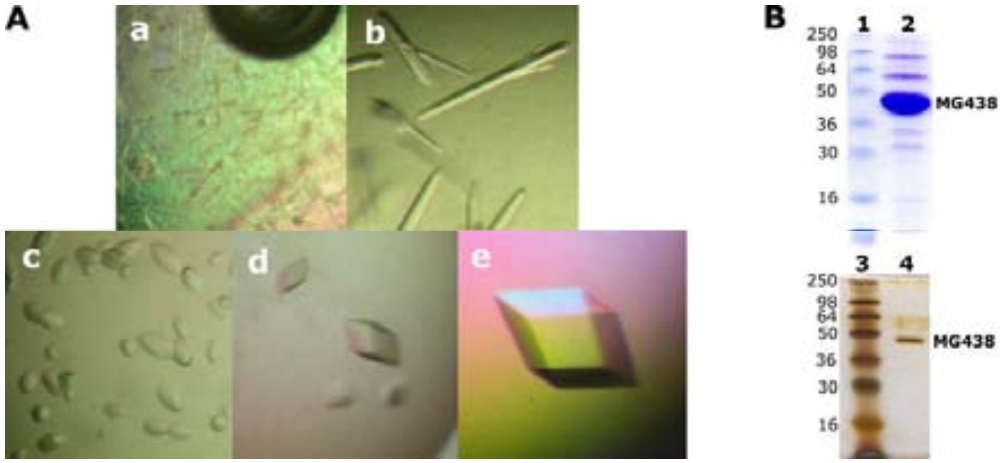


Figure II.1. MG438 crystals and their analysis by SDS-PAGE. A) MG438 crystals, from nano-crystals to single diffracting-quality crystals: (a) small nano-crystals grown in a high-throughput crystallization plate in presence of ammonium sulfate at pH 5.5 were optimized to (b) large and thin needle crystals that diffracted to 7 Å resolution when exposed to synchrotron radiation. Crystals with a different morphology were obtained at pH 5.0 using NaCl as precipitant (c and d) and optimized to quality-diffracting crystals by adding EDTA disodium salt dihydrate to the previous crystallization condition (e). C) Coomassie brilliant blue (CBB)-stained denaturing gel showing a major protein band corresponding to the pure MG438 protein solution (upper gel) and silver-stained denaturing gel showing a MG438 protein band coming from the dissolution of crystals (lower gel), proving that these are from the interest protein. The molecular weights (MW) of protein standards and the positions of the MG438 protein bands are indicated.

2.2. MG438 heavy-atom derivative crystals and initial phasing

At this stage, a new set of experiments were designed to produce MG438 heavy-atom derivative crystals with the purpose of phasing, since no three-dimensional structure of a MG438 homologue was available that would allow solving the protein structure by the molecular replacement (MR) method.

In general, native protein crystals were very sensitive to manipulation and sometimes dissolved or disintegrated after opening the crystallization drop. The protein crystals that maintained their appearance after manipulation were transferred as quickly as possible to the heavy-atom harvesting solutions. Some crystals presented severe crashes when monitored under the light microscope immediately or a few minutes after immersion in the heavy-atom harvesting solution, as were the cases for crystal soaking in thiomersal or $GdCl_3$ solutions for which no diffraction datasets are available.

Analysis of the heavy-atom derivative crystals diffraction data revealed that some did not diffract at all, as were the case for the $HgCl_2$-derivatized crystals, crystals soaked for several days in K_2PtCl_4, or crystals soaked in concentrations of $KAu(CN)_2$ higher than

0.1 mM. In general, derivative crystals diffracted less than the native ones and presented higher mosaicity. The main problem in working with these crystals was in the collection of complete datasets with high multiplicity for the anomalous data, maybe because they became more fragile after soaking.

The presence of ordered heavy-atoms in the crystals was analysed by the inspection of various statistical parameters calculated for data after raw data processing (Table II.1).

Table II.1. Summary of the heavy-atom derivative crystals data collection.

Heavy-atom	Chemical Formula	Soaking Conditions†	Resolution (Å)	Completeness (%)	R_{merge} (%)	Derivative
Mercury*	$Hg(CH_3COO)_2$	5 / 72 h	3.7	91.9 (89.2)	12.8 (56.8)	X
	$HgCl_2$	10 / 24 h	ND			
	$HgCl_2$	2.5 / 72 h	ND			
	$HgCl_2$	10 / 6 h	ND			
Platinum*	K_2PtCl_4	20 / 72 h	ND			
	K_2PtCl_4	15 / 72 h	ND			
	K_2PtCl_4	20 / 24 h	ND			
	$K_2Pt(CN)_4$	5 / 72 h	4.0	91.1 (93.2)	6.3 (2.5)	X
Gold*	$KAu(CN)_2$	0.25 / 24 h	ND			
	$KAu(CN)_2$	0.1 / 24 h	3.4	85.0 (83.4)	8.2 (41.4)	X
Osmium‡	$K_2Os_4Cl_6$	10 / 24 h	3.3	100 (100)	10.9 (96.6)	✓
Tantalum‡	$Ta_6Br_{14}.8H_2O$	10 / 24 h	5.5	100 (100)	7.0 (31.7)	X
Bromide‡	KBr	1 M / 10’	3.6	100 (100)	15.8 (77.3)	X
Iodine‡	NaI	0.5 M / 1’	ND			

† Heavy-atom soaking concentration (mM) / soaking time, * Diffraction data collected at the ESRF beam line ID13, ‡ Diffraction data collected at the ESRF beam line BM16, ND Crystals that did not diffracted.

A single MG438 crystal derivative was found that contained osmium. A clear heavy-atom site appears in the anomalous difference Patterson map that was calculated using the FFT Patterson tool from the CCP4 program suite (Collaborative Computational Project 1994, Figure II.2A). Osmium substructure was determined by SHELXD using the anomalous signal from osmium in a Single-wavelength Anomalous Dispersion (SAD) experiment. Initial phases, calculated excluding data beyond 4.0 Å, produced an electron density map difficult to interpret due both to the low connectivity and mainly to the low resolution, though possible helical structures can be already identified (Figure II.2B).

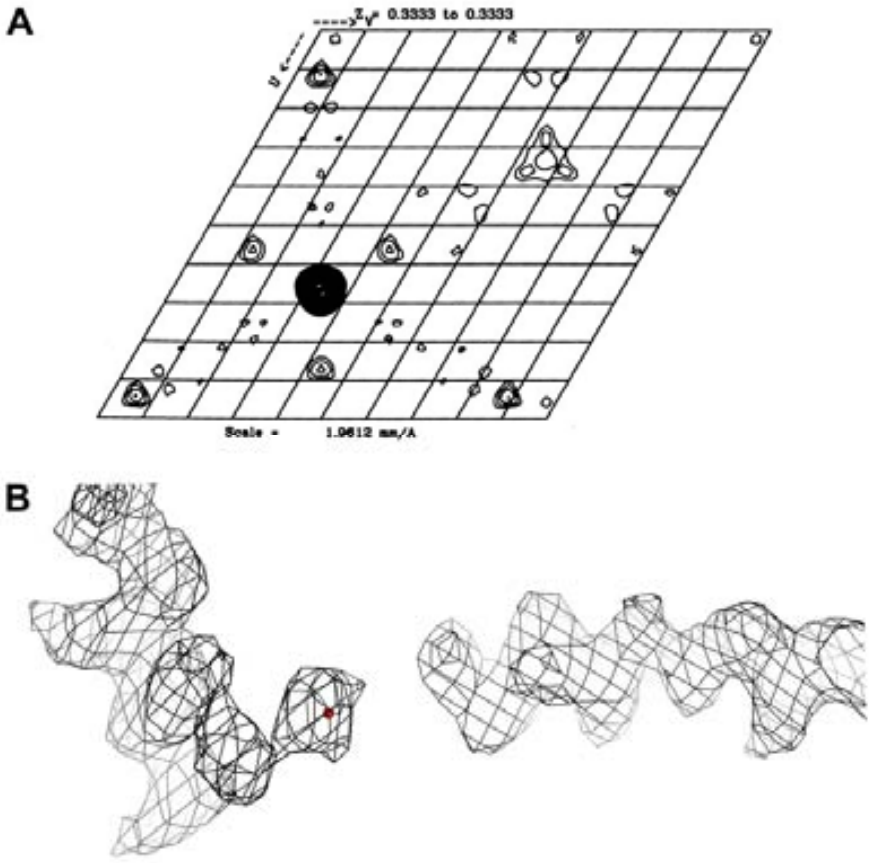


Figure II.2. Anomalous difference Patterson and electron density map calculated from an Os-MG438 derivative crystal. A) Anomalous difference Patterson map of the peak wavelength dataset of an Os-MG438 crystal: the map is a Harker section in space group $P3_121$ showing a clear 6σ peak height corresponding to the self-vector. B) Sections of the 4 Å resolution experimental electron density map obtained from osmium phasing contoured at 1σ, with 2Fo-Fc computed by using COOT (Emsley and Cowtan 2004). The osmium atom position located is indicated by a red sphere.

2.3. SeMet-labelled MG438 structure determination

Preparation of diffraction-quality heavy-atom derivatives of native MG438 crystals was a difficult task because crystals became more fragile after manipulation and soaking. Thus, another strategy was followed in order to obtain the phases for solving the protein structure that consisted in producing selenomethionine (SeMet)-labelled MG438. A highly pure protein fraction was obtained using the same protocol used to purify the native protein, the only difference being in the salt concentration in which the proteins were more stable. The native protein is stable in 0.7 M NaCl while 1.2 M NaCl is needed to stabilize the SeMet-labelled MG438 protein.

The SeMet-labelled MG438 crystals, which were obtained at about the same conditions determined for the native protein, allowed the X-ray crystal structure determination of

MG438 by using Multiple-wavelength Anomalous Dispersion (MAD) data (Table II.2). Crystals contained one molecule per asymmetric unit with a high solvent content (64 %), parameter that could contribute to the moderate atomic resolution achieved when exposing them to potent synchrotron radiation.

The final model, refined at 2.3 Å resolution, contains residues 1-374, two chloride atoms and 129 water molecules. The goodness of fit of the model to the experimental structure factors is reflected in an R_{factor} of 19.7 (R_{free} of 23.1). The stereochemistry of the model is good with 97 % of residues in the most favoured region, 2 % in the allowed region and only 1 % in the disallowed region of the Ramachandran plot. Details of refinement statistics and geometry parameters are given in Table II.2.

Table II.2. Data collection, phasing and refinement statistics.

	Se (1)[a]			Se (2)[a]	Os[b]
Data collection [c,d]					
Space group		$P3_121$		$P3_121$	$P3_121$
Unit cell parameters a=b, c (Å)		76.4, 174.4		76.6, 174.8	76.5, 174.9
Data set	Peak	Inflection	Remote		Peak
Wavelength (Å)	0.97935	0.97950	0.97565	0.97855	1.13984
Resolution (Å)	35-2.5 (2.64-2.50)	35-2.5 (2.64-2.50)	35-2.5 (2.64-2.50)	20-2.3 (2.36-2.30)	40-3.30 (3.42-3.30)
R_{merge}	6.6 (18.4)	6.9 (23.5)	7.1 (37.6)	7.9 (42.5)	7.8 (30.8)
$< I/\sigma(I) >$	6.6 (3.9)	6.4 (3.1)	6.5 (2.0)	5.7 (1.7)	7.6 (4.0)
Completeness (%)	99.6 (100)	99.6 (100)	99.7 (100)	99.6 (98.8)	99.5 (99.4)
Redundancy	7.2 (7.3)	6.9 (7.1)	7.2 (7.3)	7.0 (6.2)	9.2 (7.6)
Refinement					
Resolution (Å)				20-2.3	
No. of reflections (work/free)				25 685 / 1842	
R_{factor} **(%)**				19.7 (24.6)	
R_{free} **(%)**				23.1 (29.4)	
No. atoms					
Protein				3033	
Ions				2	
Water				129	
Averaged $B_{factors}$ (Å^2)					
Protein				40.8	
Ions				47.1	
Water				40.9	
Rms deviations					
Bond lengths (Å)				0.02	
Bond angles (º)				1.5	

[a] Dataset (1) was used for the initial phasing while dataset (2) was used for refinement. Dataset (2) was also a SeMet derivative because only about 3 Å resolution could be obtained from the best diffracting native crystals. [b] Osmium derivative dataset that was not used to solve the protein structure. [c] Values in parentheses are for the highest resolution shell. [d] The overall figures of merit of phasing from SOLVE and RESOLVE were 0.31 and 0.67, respectively for dataset (1). For the osmium dataset the overall phasing figure of merit from SHELXD was 0.65 to a cut-off resolution of 4 Å.

Despite the quality of the electron density maps that allowed the confident tracing of most residues' side chains, no density was visible for the protruding loop Asp310-Pro313 side chains, for the 23 N-end residues extension originated from the plasmid and the histidine tag used and for 9 C-end residues of ORF MG438. The residues that have not been included in the model are presumed to be disordered in the crystal structure.

2.4. Overall structure description

MG438 consists of a single polypeptide chain of 383 amino acids. It folds into three distinct structural regions (Figure II.3A-E): an N- and a central-globular domains and a pair of 40 residues long antiparallel α-helices. The long α-helices constitute the two conserved regions that separate the two globular domains.

The N-terminal domain, which correspond to the variable target recognition domain one (TRD1) as previously described by sequence analysis comparison (Bickle and Kruger 1993), is constituted by the N-terminal residues Met1 to Asn142 plus the C-terminal residues Pro371 to Thr374. Similarly, the central globular domain also well corresponds to TRD2, including residues Ser183 to Ser330. Finally, the two antiparallel α-helices are constituted by residues Lys143 to Phe182 and Phe331 to Leu369 that correspond to the central (CR1) and C-terminal (CR2) conserved regions, respectively (Figure II.3A). This organization, in which the CRs separate the two TRDs, supports the two domain model for S subunits proposed by Argos (Argos 1983) and later extended in other studies by Murray and others (Fuller-Pace and Murray 1986; Gann, Campbell et al. 1987).

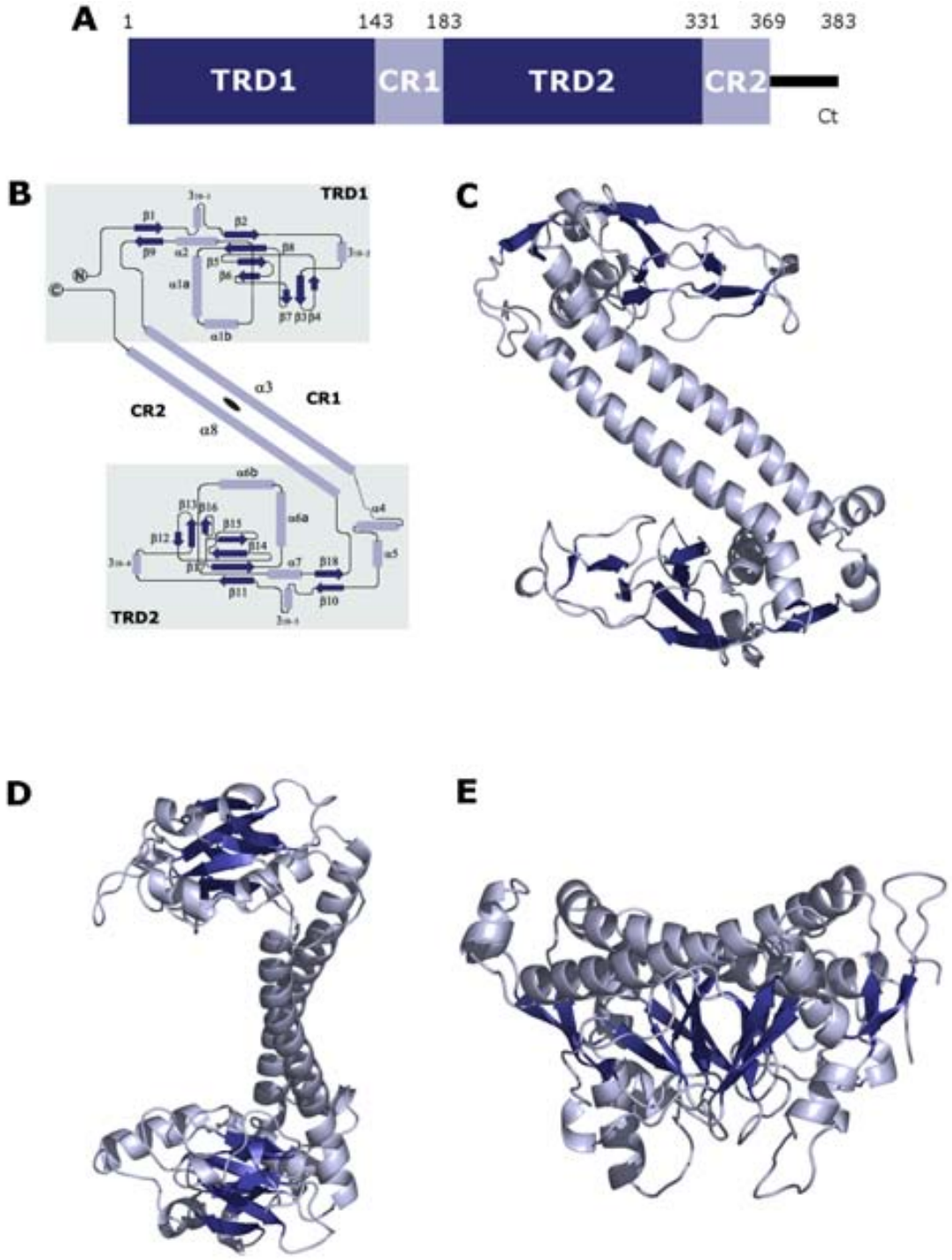


Figure II.3. MG438 overall structure representation. A) Schematic representation of the domain assignment of an S subunit. B) Scheme of the MG438 topology with helices depicted in light blue cylinders and β-strands as dark blue arrows. The cyclical organization of one subunit with an internal quasi-two fold axis, indicated by the oblong dot, is apparent in this representation. The N- and central-globular domains are referred to, for consistency with sequence analysis, as the target recognition domains TRD1 and TRD2, respectively. The long α-helices separating the TRDs correspond to the conserved regions CR1 and CR2. It is worth noting the overall S shape of the representation of the S subunit topology from a type I R-M system. Ribbon representations of the MG438 structure as view (C) down the intra-subunit two-fold axis, (D) after rotating 90º to the right and (E) after rotating the initial view 90º forwards.

Structure of the globular TRDs

The two globular TRDs, of ~150 amino acids each, have very similar fold according to the DSSP algorithm (Kabsch and Sander 1983) and as expected given the 24 % sequence identity (Figure II.3B and II.4).

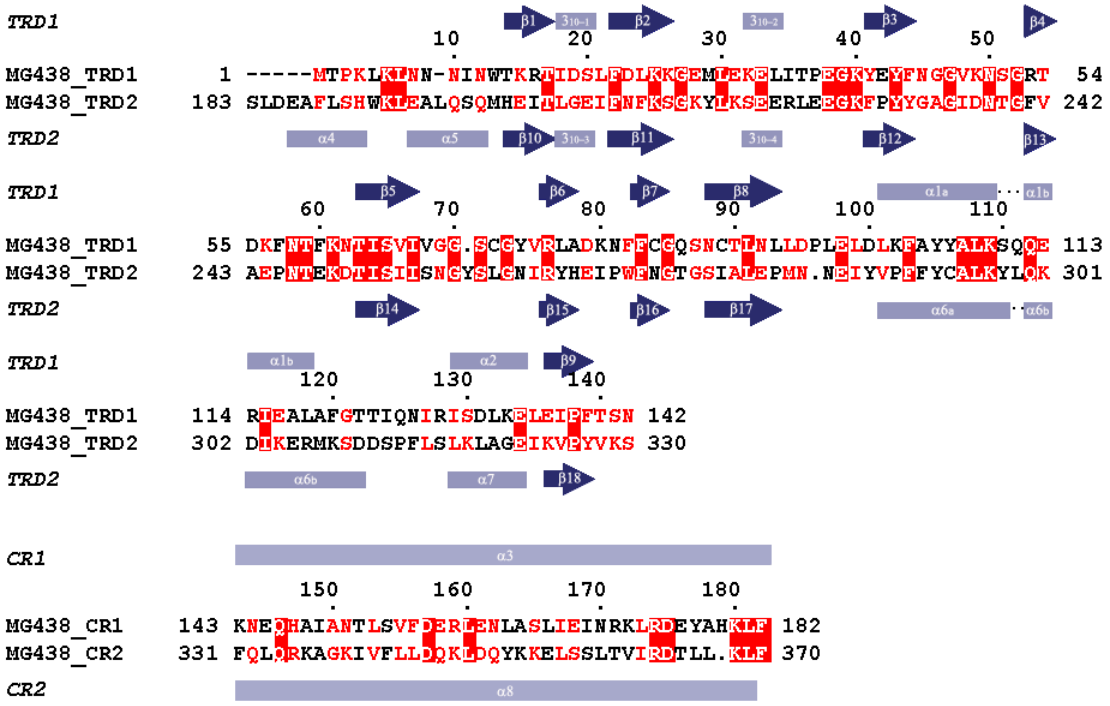


Figure II.4. Structure-based sequence alignment of MG438 TRD1/TRD2 and CR1/CR2. Identical residues have a red background colour, and residues presenting high and moderate homology are depicted in red.

Despite the high structural similarity between TRD1 and TRD2 some small differences were encountered between them. In TRD1 the two α-helices α4 and α5 found at the beginning of TRD2 are missing. The equivalent locations of these helices in TRD1 are occupied by the N- and C-terminal residues of the MG438 structure. The remaining secondary structure elements, initiated and ended by two antiparallel β-strands (β1-β9 in TRD1 or β10-β18 in TRD2), have equivalents in both TRDs. Each TRD consists of two antiparallel β-sheets with four and three short β-strands, respectively (β2-β8-β5-β6 and β4-β3-β7 in TRD1 and β11-β17-β14-β15 and β13-β12-β16 in TRD2) and of three consecutive α-helices (α1a-α1b-α2 in TRD1 or α6a-α6b-α7 in TRD2) (Figure II.3B). The first α-helix has a strong hydrophobic character while the other two are amphiphilic. In each domain there are also two short 3_{10}-helices (residues Ile18-Leu21 and Lys32-Leu34 in TRD1 or Leu 206-Glu208 and Ser220-Glu222 in TRD2). The four-

stranded β-sheet is the central motive of the TRDs interacting on one side with the hydrophobic helix and on the opposite side forming a small β-sandwich with the three-stranded β-sheet. On the other side, this three-stranded β-sheet is exposed to the solvent at the top of the domain and surrounded by two prominent loops (centred on residues Gly39 and Thr123 in TRD1 or Ser220 and Asp310 in TRD2). The hydrophobic α-helix is oriented perpendicularly to the CRs and in direct contact with them defining a hydrophobic core at the junction between each TRD and the two CRs (Figure II.5A-B). These hydrophobic cores include the apolar residues Leu18, Phe22, Leu24, Leu64, Val66, Val68, Val75, Cys89, Leu91, Leu108, Ile115, Leu18, Ile127, Ile129, Ile132, Ile135 in TRD1, and Leu206, Phe212, Ile252, Ile254, Ile264, Ile278, Leu280, Val289, Leu296, Ile303, Leu307, Leu315, Leu317, Ile323 in TRD2. The hydrophobic cores from both TRDs are rather enlarged and delimited by the amphyphilic helices.

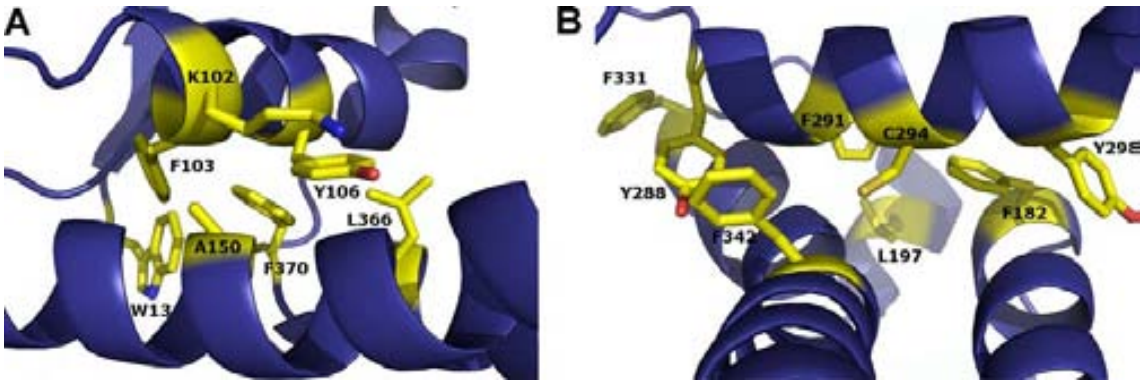


Figure II.5. Representation of the interactions between TRDs and CRs. A) Hydrophobic interactions formed at the boundary of TRD1-CRs (B) and TRD2-CRs.

Structure of the two CRs

The two CRs, of ~40 residues each, have conserved primary sequences among type I R-M S subunits, unlike the sequence variable TRDs and often exist in the middle and at the C-end of their amino acidic sequences. The CRs have been suggested to provide a structural motif for protein-protein interactions with the other type I R-M subunits based on the evidence that desired specific activity can be fully recovered after domain swapping (Kneale 1994).

In MG438 the two CRs correspond to the two longest antiparallel helices, α3 and α8, which occupy the structure central region. These helices are organized following very well the principles of coiled-coils, having an approximate heptad periodicity of the

hydrophobic residues in each helix and forming a left-handed super-coil (Crick 1953). As a result of this organization the two helices are held together mainly by hydrophobic interactions, which are complemented by five salt bridges ($O^{\delta 1}$Asp157-NH_1Arg362-$O^{\delta 1}$Asp363, NH_2Arg159-$O^{\varepsilon 2}$Asp354 and $O^{\delta 1}$Asp175-NH_1Arg174-$O^{\delta 1}$Asp345) (Figure II.6). Interactions between S subunits CRs were already taken as the base to perform reconstruction of active S subunits from truncated S polypeptides (Mernagh, Reynolds et al. 1997; Smith, Read et al. 2001).

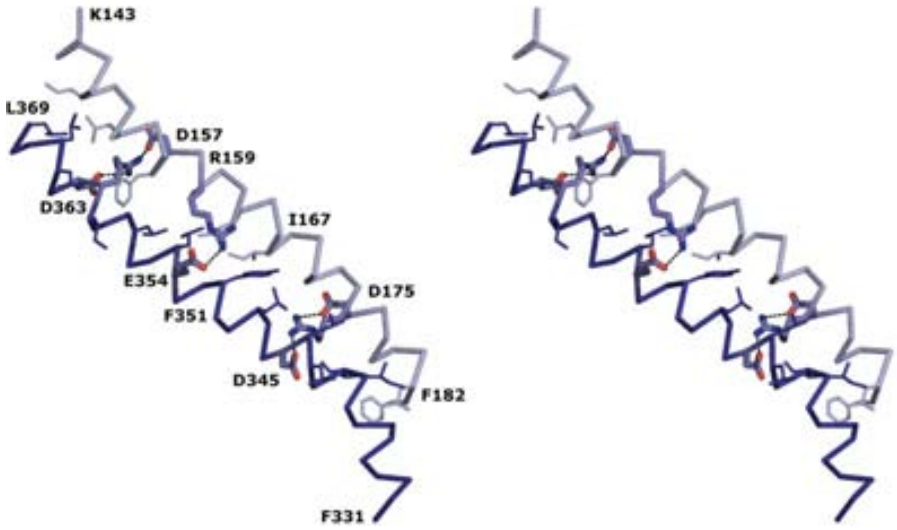


Figure II.6. Stereo view of the CRs coiled-coil structure. The two CRs, located in the middle and at the C-end of MG438 primary sequence, form almost perfect coiled-coil structures with a large number of hydrophobic residues between the two helices (represented as sticks). Interactions between the helices are completed with the presence of five salt bridges, which are also indicated. Residues involved in the salt bridges are depicted in the chicken-box representation and are coloured according to their atoms type.

Structural homologues of this coiled-coil structure have been frequently observed in other protein structures, such as the coiled-coil region of the DNA repair system RAD50 ATPase (Hopfner, Craig et al. 2002), the cytosolic helical bundle structure of chemotaxis receptor protein in signal transduction (Kim, Yokota et al. 1999) and the prefoldin/β-tubulin binding postchaperonin cofactor in protein-protein interaction (Siegert, Leroux et al. 2000).

The intra-subunit two-fold axis

Superposition of the two TRDs with each other gives root mean square distance (rmsd) of 1.64 Å for 130 structurally equivalent C^{α} atoms. Moreover, the whole subunit can be superposed with itself, by a pure rotational symmetry of exactly 180º, resulting in rmsd

of 1.55 Å for 345 equivalent C^{α} atoms. Such a good superposition reveals both the accuracy of the intra-subunit binary symmetry and the anticipated circular organization of the subunit (Figure II.7) (Kneale 1994).

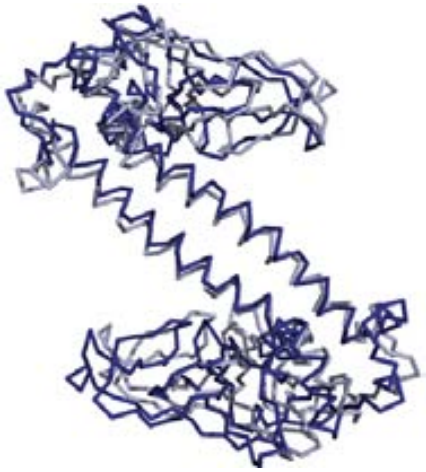


Figure II.7. Superposition of the whole MG438 protein with itself reveals an accurate intra-subunit two-fold rotation axis.

The electrostatic potential of the protein is highly polarised, as can be clearly appreciated in the electrostatic surface charge representation along the intra-subunit two-fold axis (Figure II.8), allowing the definition of two molecular faces named as **S** and **Z** because of their overall shapes. The **S** face is flat and shows two conspicuous patches of exposed hydrophobic and positively charged residues. In turn, the **Z** face is rugged, presenting the CRs at a lower level in relation to the TRDs surfaces and will also be called the 'DNA binding' face for reasons described later in section II.2.9.

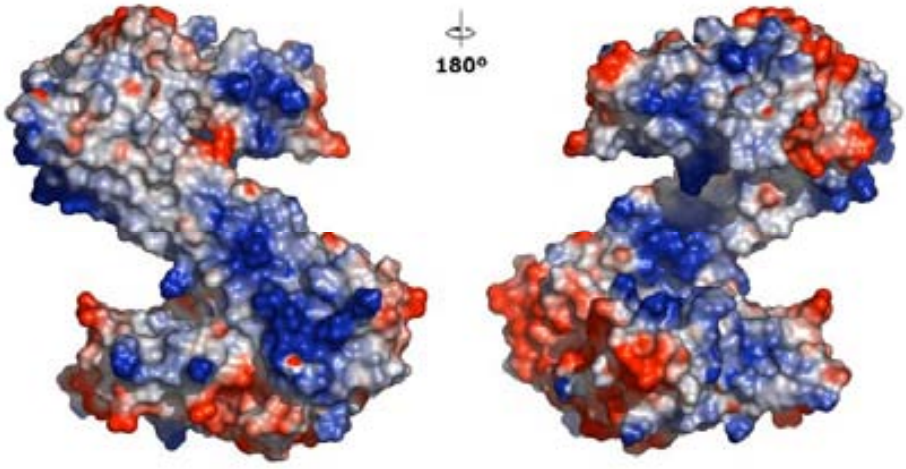


Figure II.8. Views along the intra-subunit two-fold axis of the S (left panel) and Z (right panel) faces of MG438. The molecular surface is coloured according to the local electrostatic potential as calculated with the program PYMOL (DeLano 2002). The **S** face shows two conspicuous patches of exposed hydrophobic (white) and positively charged (blue) residues. In turn, the **Z** face, or DNA binding face, present a mostly polar character where positive and negative (red) charges alternate.

2.5. Structural comparison of MG438 with its close homologue S.Mja

MG438 was assigned to the type I R-M S subunit family of proteins by sequence homology. MG438 shares close overall structural homology with the almost simultaneous solved structure of an S subunit from *Methanococcus jannaschii* (S.Mja; Kim, Degiovanni et al. 2005). The S.Mja structure, solved to 2.4 Å resolution at BSGC, superposes with rmsd of 2.47 Å for 218 (57 %) structurally equivalent C^{α} atoms (Figure II.9A). Superposing only the TRDs, the corresponding values are 1.91 Å for 113 (79 %) equivalent C^{α} atoms between TRD1s and 1.76 Å for 117 (81 %) equivalent C^{α} atoms between TRD2s, respectively (Figure II.9B and C). When restricting the comparison to the CRs, the equivalent C^{α} atoms rise to 76 (95 %) with rmsd of 1.66 Å (Figure II.9D).

However, differences are apparent between the MG438 and S.Mja structures. Major differences are observed both in the different disposition of the TRDs with respect to the CRs, with a relative rotation of 21.5º (Figure II.9E) and in differences between the TRDs both at the level of the secondary structure and in the conformation of some loops (Figure II.9B and C).

The orientations of the TRDs with respect to the CRs seem to be very precise and rigidly defined on both structural homologues, as reflected by the accuracy of the intra-subunit two-fold symmetry axis that is not affected, even by the different crystal environments around each TRD. In the MG438 structure, which have shorter TRDs, a β-hairpin is missing that is found at the N-terminal end of both S.Mja TRDs. Besides, S.Mja TRDs structure has four and five strands in the small and central β-sheets while in MG438 there are only three and four, respectively.

The largest differences between S subunit loops from *M. genitalium* and *M. jannaschii* are respectively for Lys64-Leu78 with respect to Lys32-Lys49 from TRD1 and for Cys290-Lys345 with respect to Ile 236-Pro245 for TRD2.

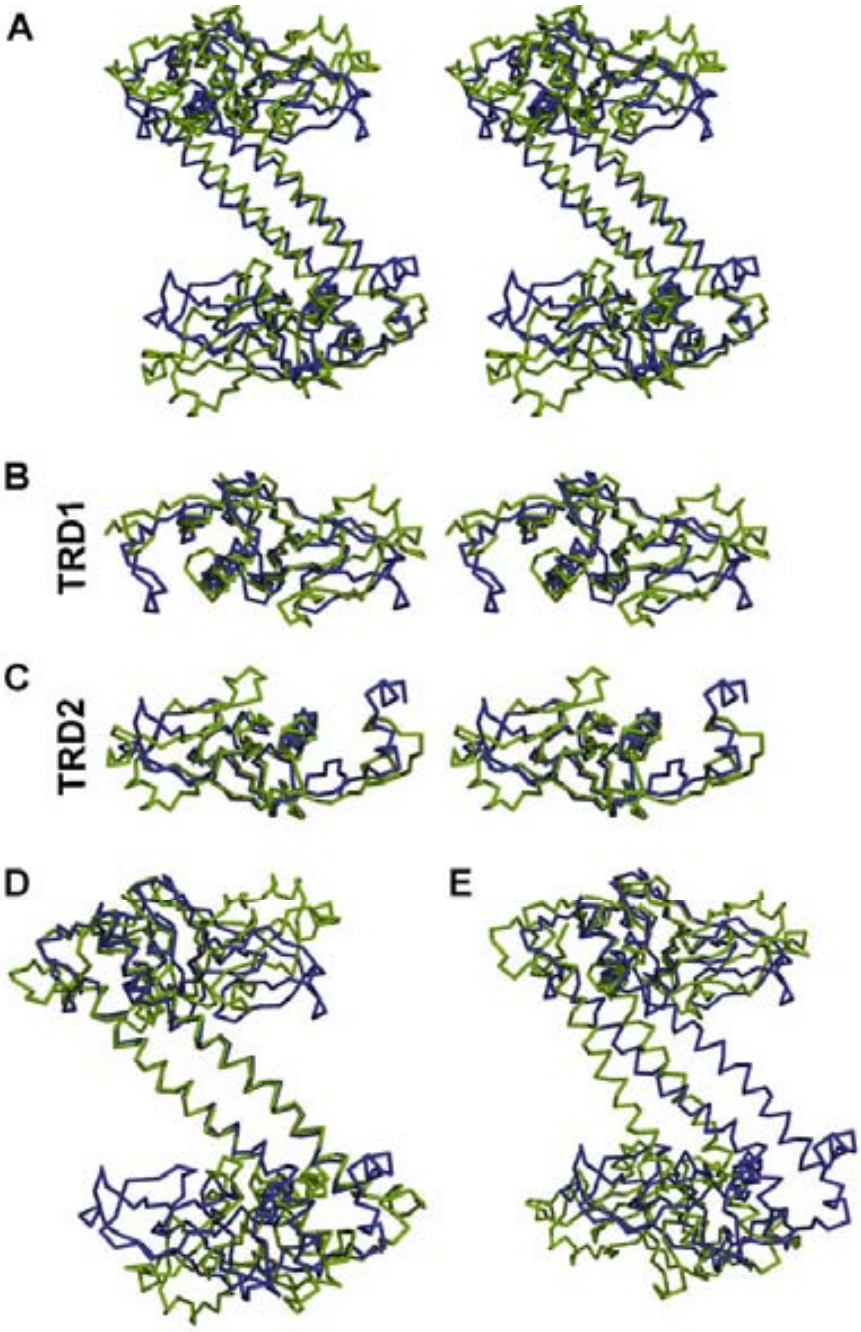


Figure II.9. Structural comparison between the *M. genitalium* MG438 protein and its close homologue from *M. jannaschii*, S.Mja. A) Stereo view of the overall superposition of the MG438 (blue) and S.Mja (green) structures. B) Superposition of TRD1s from both structures with rmsd of 1.91 Å. C) Superposition of TRD2s from both structures with rmsd of 1.76 Å. D) Overall superposition of the MG438 and S.Mja structures when using the transformation obtained from the superposition of the CRs. E) Same as (D) but now using the transformation obtained from the superposition of TRD1s. The different orientation of the TRDs with respect to the TRD1s (~21.5º) is evident in (D) and (E).

2.6. Primary sequence comparison of MG438 with other S subunits

A structure-based alignment of MG438 with S.Mja served as restraint to perform careful primary sequence alignments with representative S subunits from type I R-M systems families IA, IC, ID, and IE (Figure II.10). The IB family has been excluded from the alignment due to its low sequence similarity (Murray, Gough et al. 1982).

In general, the alignments show very weak consensus sequence with only a few particularly relevant features (numbering refers in all cases to the MG438 amino acidic sequence, Figure II.10):

i) residues corresponding to the MG438 β-strand β5 from TRD1, and also to its equivalent β14 in TRD2, present a pronounced hydrophobic character, which would be explained by the inaccessibility of these strands to solvent in the MG438 structure;

ii) a similar explanation can be given for the hydrophobic character of residues corresponding to the hydrophobic helix α2a in TRD1 and α6a in TRD2;

iii) the two CRs match well with sequences of about 40 contiguous residues, showing an approximate heptad periodicity for the hydrophobic amino acids;

iv) Pro138 in TRD1, or its equivalent Pro326 in TRD2, located at the end of their respective TRDs, are fully conserved residues always preceded by a hydrophobic one, mostly isoleucine. The main-chain oxygen of this conserved proline is inside the hydrophobic core without making any hydrogen bond, which likely requires the reduced conformational flexibility of proline residues for stability;

v) Gly215 in TRD2 is also a fully conserved residue. The equivalent location in TRD1 (Gly 27) is not always, but only often, a glycine. Due to the proximity of Gly215 atoms to the main-chain of neighbour residues, with shortest distances of 2.90 Å to Gly274 and 3.16 Å to Asn273, steric hindrances are expected for any amino acids different from glycine in this position.

Figure II.10. Multiple-sequence alignment of the S subunit homologues from MG438: *Methanococus jannaschii* DSM 2661 S.Mja, *Escherichia coli* S.EcoKI (prototype for family IA), *Escherichia coli* S.EcoR124I (prototype for family IC), *Salmonella enterica* StySBLI (protoptype for family ID), *Klebsiella penumoniae* KpnBI (prototype for family IE). The secondary structure elements, assigned according to the MG438 structure, are also depicted with light blue rectangles for helices and dark blue arrows for β-strands. The structural alignment between MG438 and S.Mja was performed with the program SHP (Stuart, Levine et al. 1979) and used as a constraint during the sequence alignment performed with ClustalW (Larkin, Blackshields et al. 2007). Despite the low sequence identity the conservancy of the polar or neutral character becomes clear when using a similarity colour code: blue for the positively charged residues K and R; red for the negatively charged D and E; light orange for the hydrophobic residues A, C, F, I, L, M, V; dark orange for the aromatic residues H, W, Y; green for the polars N, Q, S,T; green background for proline and red background for glycine residues.

Altogether, S subunits structural and sequence comparisons strongly support the idea of a common origin for the type I R-M systems, with *hsdS* genes coding for molecules having an accurate intra-subunit two-fold symmetry that could have arisen by either gene duplication or duplication followed by gene fusion (Sharp, Kelleher et al. 1992). The low consensus between TRDs appears due mainly to the variable requirements imposed by the diversity of target DNAs, but the alternative roles of S subunits could also contribute to this weak homology.

2.7. MG438 crystal packing

The MG438 protein found in the crystal is very close to a symmetry related MG438 subunit. The two subunits are related by a crystallographic two-fold axis and present a contact surface of 2340 $Å^2$ (Figure II.11A-B) from which 1570 $Å^2$ are due to hydrophobic interactions. These are established between symmetry related apolar side-chains from residues Leu5, Leu7, Leu97 (TRD1), Val155, Phe156 (CR1), Leu358, Ile361, Leu365 and Leu369 (CR2). In a single MG438 subunit this cluster of apolar side-chains would be exposed to the solvent at the **S** face of the structure (see more on Figure II.8 left panel).

The inter-subunits interface is complemented by the presence of four strong hydrogen bonds, according to their short bond distance, between Ser141-Glu226* and Asn151-Ser357* side chains. The absence of solvent, associated to the low $B_{factors}$ of atoms from residues located in this interface, suggest the existence of a strong and rigid interaction between the two subunits. It can be also appreciated when analysing the quality of the experimental electron density map of this region (Figure II.11C).

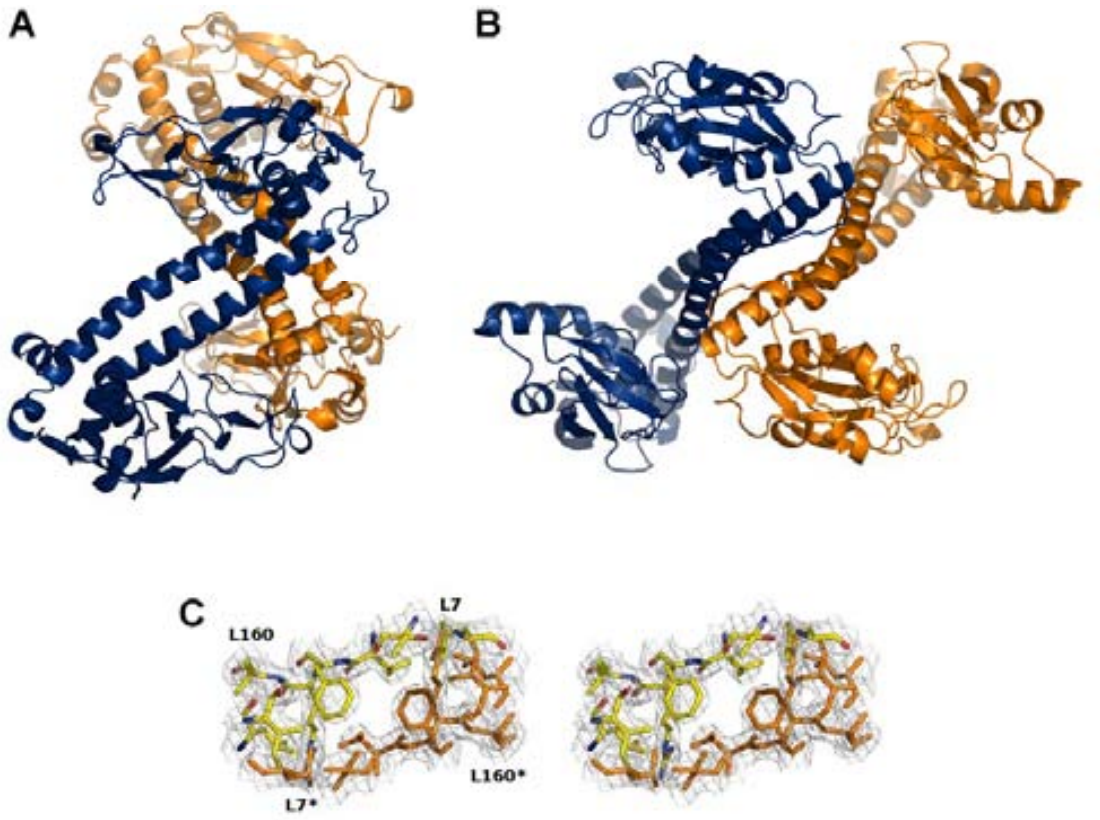


Figure II.11. Dimeric organization of MG438 in the crystal and detail of the hydrophobic inter-subunits interface. A) View down the reference intra-subunit two-fold axis. Subunits are depicted in cartoon representation with the reference subunit in blue. B) View along the crystallographic two-fold axis of the MG438 dimer, presenting a large contact area (about 90º apart from the orientation in (A)). C) Stereogram, in the vicinity of the crystallographic two-fold axis, of the 2Fo-Fc electron density map (chicken box representation) corresponding to the cluster of hydrophobic residues that conform most of the dimer interface. The view orientation is the same as in (B) and residues from the reference subunit are coloured according to their atom types.

The inter-subunit crystal two-fold axis is close to perpendicular (96º) to the intra-subunit binary axis of one of the subunits. However, these inter- and intra-subunit two-fold axes do not intersect ruling out the possibility of point group-like molecular organization combining these symmetries. Instead, when the intra-subunit symmetry is considered, the apolar cluster of contacting residues is only in part mirrored by solvent exposed residues still retaining a hydrophobic character, such as Leu166, Ile169, Leu173, Leu343 and Leu344 from the CRs. This second hydrophobic patch at the **S** face of MG438, though smaller than the first, must also have a tendency to participate in intermolecular interactions. This possibility is supported by the S.Mja structure where the two subunits found in the asymmetric unit, related by a local two-fold symmetry, present a contact surface corresponding to the location of the second hydrophobic patch

in MG438 (Figure II.12). The possibility of oligomers containing simultaneously the two types of dimers seem to be excluded by steric clashes.

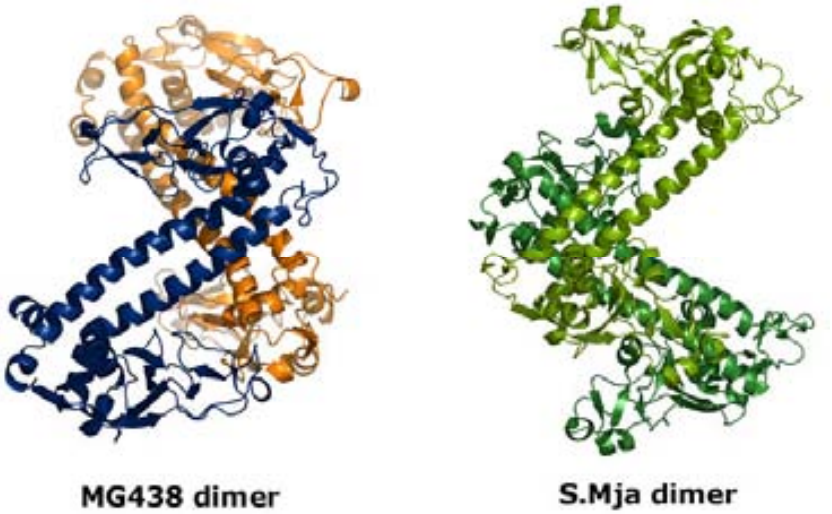


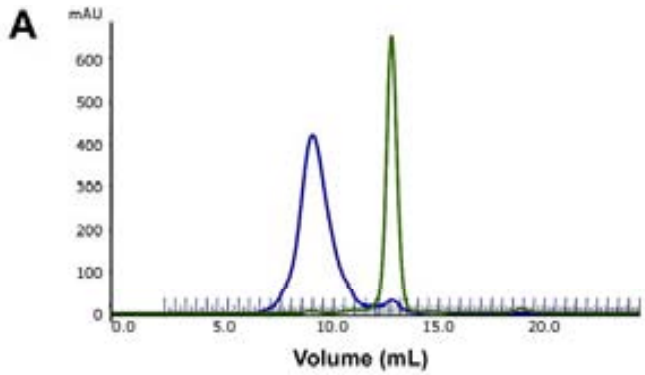
Figure II.12. Comparison between the MG438 crystallographic dimer and the non-crystallographic dimer found in the S.Mja crystal structure.

Two solvent-exposed hydrophobic regions that exist at the **S** face of the MG438 structure were already described. In the rest of the structure the presence of this kind of surfaces is limited to residues Phe57, Phe60, Phe83 in TRD1 and Phe229, Phe241, Ile269 in TRD2. In this way, the presence of large hydrophobic surfaces seem to be confined to the **S** face of the subunit, in contrast to what is observed in the S.Mja structure, likely reflecting the solitary character of the *M. genitalium* type I R-M S subunit. These hydrophobic surfaces can also be related to the poor solubility found for isolated S subunits (Patel, Taylor et al. 1992; Dryden, Cooper et al. 1997), which could increase when burying these surfaces either by oligomerization, as in the case of MG438 and S.Mja dimeric structures, or by interactions with type I R-M M subunits in the MTase complex (read more on section II.2.9).

2.8. MG438 possible oligomeric structures

In order to study the oligomeric state of MG438 in free solution several biochemical and biophysical techniques were applied. Gel filtration chromatography (GFC) performed at physiological pH (~7.5) show the presence of large, discrete oligomeric structures that could reach a molecular weight of approximately 600 kDa (corresponding to almost 12 MG438 subunits). This experiment performed at the crystallization pH (5.0) revealed the exclusive presence of MG438 dimers, which likely correspond to the ones observed in the crystal (Figure II.13A). To cross-validate these

results a MG438 protein solution was measured in a dynamic light scattering (DLS) instrument at both pHs. Again, it is apparent that MG438 forms large oligomers at pH 7.5 while at pH 5.0 exists as a dimer (Figure II.13B).



B

DLS parameters

Sample	RH (nm)	PDI (%)	MW (kDa)	Intensity (%)	Mass (%)
pH 7.5	8.1	27.7	630	19.1	94.1
pH 5.0	4.3	18.3	102	98	100

Figure II.13. Analysis of MG438 size in solution by GFC and DLS. A) Fractionation of MG438 by GFC. The protein eluted at pH 7.5 (blue) in a volume of 9.4 mL (indicating a relative molecular weight of ~600 kDa) and at pH 5.0 (green) in a volume of 26.2 mL (indicating a relative molecular weight of ~102 kDa). Recombinant MG438 has a theoretical molecular weight of 94 kDa. B) Summary of DLS measurements of MG438 in solution at pHs 7.5 and 5.0. Data were acquired at 20 ºC on 7.5 mg/mL protein samples. Both samples give homogeneous peaks (unaggregated) with polydispersity indexes of 27.7 and 18.3 for samples at pH 7.5 and 5.0, respectively.

2.9. Modelling of a type I MTase ternary complex

The search for structural homologues of the whole MG438 with the DALI server (Holm and Sander 1993) retrieved no solutions suggesting that there is no reported protein structure with an identical topology. However, by limiting the search to the TRDs, these were found to be similar to the small domain (DNA binding domain) of the type II N^6-adenine DNA methyltransferase from *Thermus aquaticus* (*Taq*I MTase) for which there are a number of crystal structures available, including a DNA complex (Schluckebier, Kozak et al. 1997; Goedecke, Pignot et al. 2001).

Type I R-M M subunits consist of large enzymes which combine with the S subunits to specifically bind to and methylate DNA. A ternary methylation complex consisting of two M subunits, one S subunit and a DNA target recognition sequence was modelled in

order to determine if specific association is possible and to identify probable interacting regions.

The superposition of the *Taq*I MTase (PDB accession code 1G38) small domain with the TRDs disposes the short DNA target sequences in a way that immediately suggested an interaction model between S subunits and DNA (Figures II.15 and II.16) which present three basic features:

i) the DNA contacts only the **Z** face (the DNA binding face) of both TRDs at the location defined in MG438 by the exposed side of the three-stranded β-sheet and the surrounding loops (Figure II.3.);

ii) the portion of the DNA that bridges the separation between the TRDs, defined by the length and relative orientation of the CRs (about 22-28 Å in MG438) is not in contact with the S subunit;

iii) the two-fold symmetry relating the two DNA chains can be matched by the intra-subunit symmetry of the S subunit.

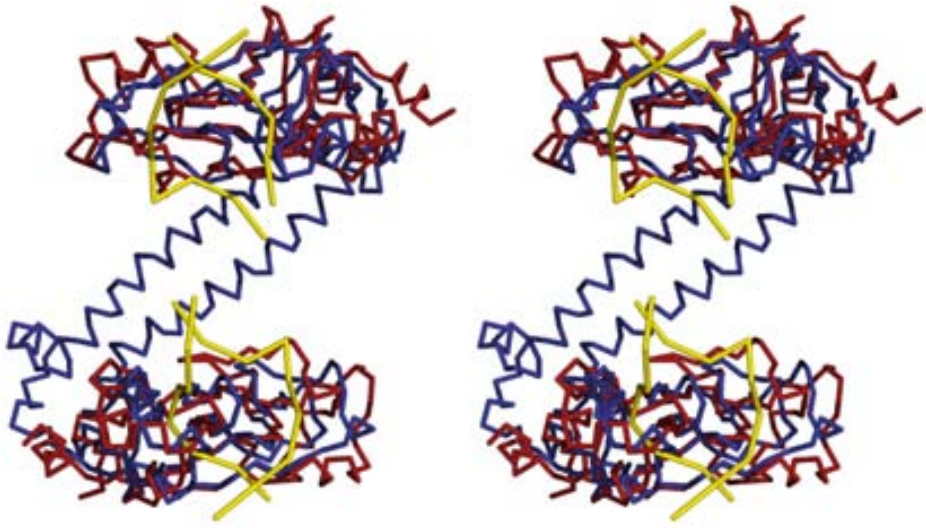


Figure II.14. Stereo view of the superposition of the *Taq*I MTase (PDB accession code 1G38, red) small domain onto the two TRDs of MG438 (blue). The corresponding DNA fragments present in the DNA-*Taq*I MTase complex structure are also shown depicted as ribbons and coloured yellow.

Taking these features into account the MTase complex was modelled as follows:

i) a regular 20 bp *B*-DNA model was docked onto MG438 such that the intra-subunit two-fold axis was coincident with a DNA two-fold axis and the

overlapping with the short DNA oligomers, coming from the superposition of the *Taq*I MTase complex, maximized (Figures II.16A-C);

ii) on the *B*-DNA oligomer the adenine bases to be extruded were assumed to be on the fifth base pair on each side of the intra-subunit two-fold axis. This disposition was chosen as 8 bp is the most common interval between the adenine bases in recognition sequences (type I R-M families IA and IC). The changes in the location of the adenine bases according to the possible spacer sequences, from 6 to 8 bp, would require only minor readjustments in the model;

iii) the C-terminal region of the *Taq*I MTase complex, comprising about 245 residues, was added by using now the DNA as superposition target (Figure II.15D-F).

Type I R-M systems M subunits share approximately 40 % sequence identity with the C-terminal region of type II MTases that have, with very few exceptions (Kim, Degiovanni et al. 2005), an extension of about 200 residues that is mainly hydrophobic for the first 50. This extension ends close to the CRs in the model obtained for the MTase complex which would easily allow the direct interaction between M subunits and CRs. Therefore, this is in agreement with a large amount of experimental data available (Kneale 1994; Dryden, Sturrock et al. 1995) and with the fact that type I R-M M subunits are also very well conserved between families (Smith, Read et al. 2001).

Several MG438 residues can be proposed, from the model analysis, to interact with the DNA. These include residues from the TRDs small β-sheets and residues from loops Met29-Lys32 and Thr122-Asn126 from TRD1 and Ser214-Tyr217 and Asp311-Phe314 from TRD2. Besides, this is in accordance with mutagenesis experiments performed on EcoKI (O'Neill, Dryden et al. 1998; O'Neill, Powell et al. 2001).

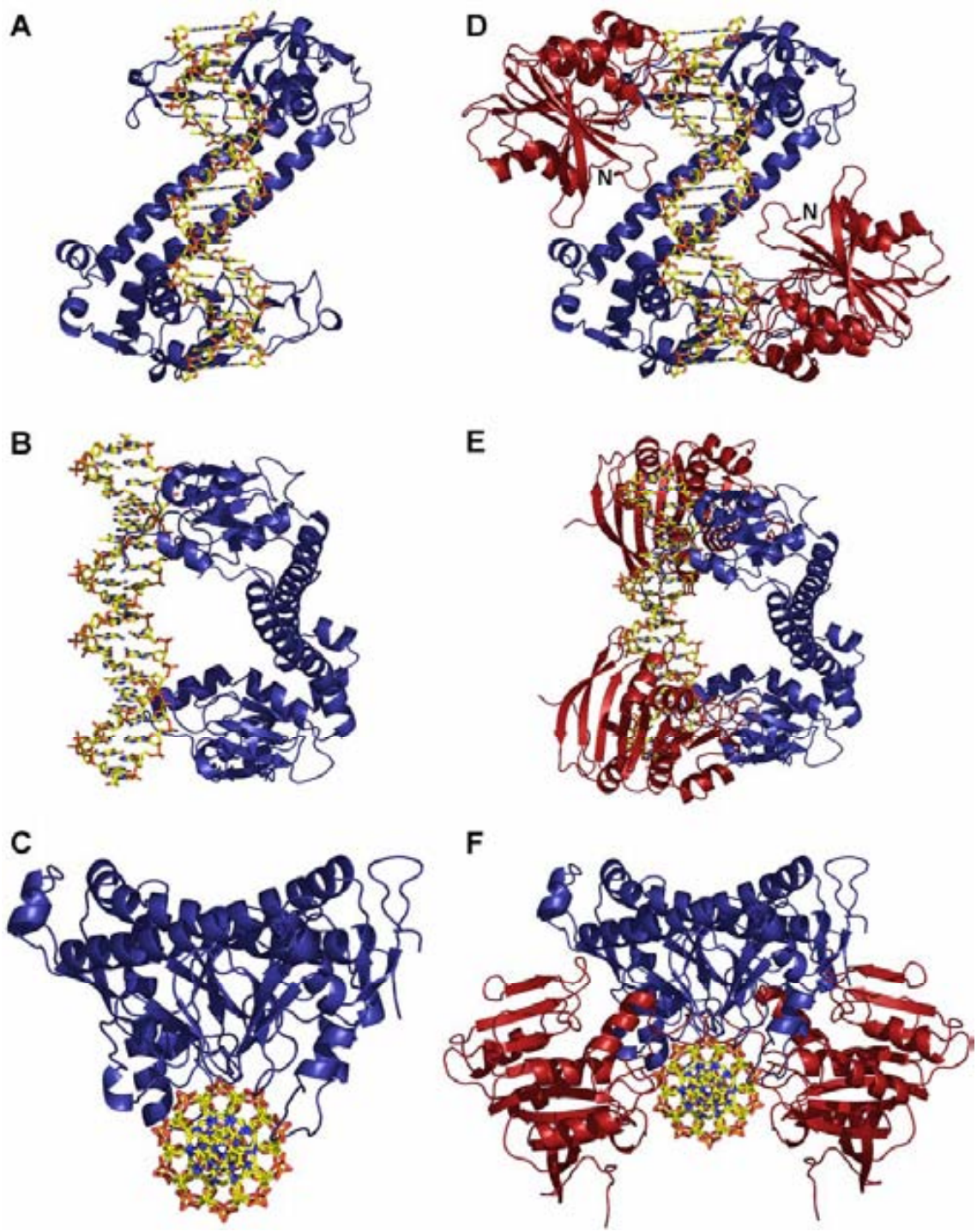


Figure II.15. Modelling of the MTase ternary complex. A standard *B*-DNA oligomer of 20 bp was firstly docked onto MG438 and after the C-terminal regions of two *Taq*I MTases were added as described in the text above. Views of the MG438-DNA complex: (A) the **Z** face or DNA binding face, (B) after rotating 90º to the left and (C) after rotating the initial view 90º forwards. (D-F) Views corresponding to (A-C) including now the M subunits. DNA is displayed with all atoms while the MG438 and the MTases are displayed as blue and red cartoons, respectively. Type I R-M M subunits have an extra N-terminal extension of about 200 amino acids with respect to the corresponding protein domains from type II MTases. This N-terminal extension (starting in the location indicated with an **N** in (D) would have, in the docking obtained, a suitable disposition to interact with the CRs.

The MTase complex model also reveals that only weakly specific interactions could be anticipated between the S subunit and the DNA with the closest contacts involving mainly the DNA phosphodiester backbone. This would explain the negative results obtained when performing MG438-DNA binding assays. These data also agree with previous experimental work suggesting that sequence-specific DNA recognition by the S subunit is only fulfilled when interacting with the M subunits (Mernagh, Reynolds et al. 1997).

The MTase complex model obtained from starting with the MG438 structure is surprisingly distinct from the one proposed for the S.Mja structure. The differences are at the level of the DNA conformation and in the disposition of the M subunits in the complex. In the MTase complex modelled starting from the S.Mja structure important distortions in the centre of the DNA molecule are assumed where no contacts with the S subunit are expected. The stability of these distorsions is dubious, except if they could be explained by interactions with the M subunits.

In MG438 the different orientation of the TRDs with respect to the CRs allowed the docking of an essentially straight *B*-DNA oligomer, while in S.Mja an important bending is required. Furthermore, the position of the M subunits away from the CRs in the MTase complex modelled starting from the S.Mja structure appears difficult to match with the already referred experimental data. A conclusive answer will require the structure determination of an MTase complex.

Despite having been studied for more than 50 years, high resolution structural information for type I R-M systems is very limited. These systems are composed by three independent subunits: S, M and R, which combine to perform DNA methylation, DNA restriction and DNA-dependent ATPase activity. To date, only a few type I R-M systems have been functionally characterized and most of them were from Enterobacterial species. The studied systems generally recognize non-palindromic sequences composed of a bipartite target site, each half-site including an adenine on opposite strands and generally ~10 bp apart that are the methylation substrates. The first half-site is of 3-4 bp and the second of 4-5 bp in length, separated by a non-specific spacer of 6-8 bp. S subunits are responsible for the recognition of both half-sites as well as for spacer length determination and for interaction with the other peptides of the

enzyme complex. From the MG438 structure it becomes evident that the two half-sites of the DNA target sequence are recognized by the highly variable TRDs while the amino acidic segments that govern the length of the spacer variable region of the DNA recognition sequence constitute the large coiled-coil structures, whose sequences are very well conserved between type I R-M S subunits that are members of a particular family. Combining structural data from MG438 and type II MTases in the MTase ternary complex model reveals that type I MTases have two-fold rotational symmetry in which each inversely oriented TRD make contact with one M subunit. The M subunits can arrange on either side of the S subunit, encircle the DNA and gain access to the methylation targets via base flipping, similarly to the mechanism followed by type II MTases (Roberts and Xeng 1998; Kennaway, Obarska-Kosinska et al. 2009).

2.10. S subunit structures as models for the design of novel DNA specificities

The conserved tertiary structure within the TRDs and the regular structure of the CRs, observed both in MG438 and S.Mja structures, imply that it may be feasible to derive the amino acid recognition code used by TRDs to recognize new DNA sequences and define the spacer length by changes in the length of the CRs, to enhance the potential for diversification of target sequences towards the generation of novel potent tools for use in Molecular Biology. Indeed, rational engineered type IIB REases have already been produced that presented previously undescribed specificities (Jurenaite-Urbanaviciene, Serksnaite et al. 2007). In that work the authors noticed, by comparative analysis, the close similarity between the C-terminal regions of type IIB REases, responsible for the specificity and some type I R-M S subunits and also that these enzymes recognize similar bipartite DNA targets. Therefore, they were able to model the CRs of three different type IIB REases using the S.Mja crystal structure as template. After careful prediction of the TRDs of the different enzymes, new polypeptides were produced that were hybrids between two enzymes, generated by domain swapping. The newly engineered type IIB REase hybrid acquired new DNA specificity and demonstrated its potential to be used as a conventional type II REase. The crystal structures determination of type I R-M S subunits together with computational protein structure modelling were crucial to design this experimental procedures. The very successful results that were achieved suggest that more structural studies on R-M enzymes will exponentially increase the possibilities to successfully engineer restriction enzymes with new specificities.

2.11. Other type I MTase models

In silico models for the most studied type I R-M S subunit, S.EcoR124, and also for the EcoR124 MTase complex were performed based on the MG438 and S.Mja crystal structures (Obarska, Blundell et al. 2006). These structural models took to the prediction of a mechanism for DNA binding where the regions adjacent to the M subunits undertake a conformational shift involving one domain of the M subunit, which would account for dramatically diminish the diameter of the protein confirming previous results on important conformational changes suffered by EcoR124 MTase upon specific DNA substrate binding (Taylor, Davis et al. 1994).

More recently, a 3D density map generated by EM of EcoKI, a type I R-M system, and single particle EM analysis of the central core of the restriction complex, M.EcoKI, bound to the T7 phage anti-restriction protein ocr (a DNA mimic) was published (Kennaway, Obarska-Kosinska et al. 2009). The EM model determined for the M.EcoKI in complex with ocr was combined with the S.Mja atomic model and with the DNA-*Taq*I MTase complex X-ray crystal structure. The new EM structure and computational model of M.EcoKI puts together numerous experimental results allowing the rationalization of most of the published data on the subject. A mechanistic explanation for type I MTases is suggested but what becomes clear is the need for a more detailed analysis of the interfaces formed between S-M and M-M subunits and between these subunits and the target DNA sequence.

2.12. MG438 possible functional roles

The crystal structure determination of the orphan S subunit from the type I R-M system from *M. genitalium*, MG438, is a hallmark on the way to the structural and functional understanding of R-M systems, having already promoted several computational modelling studies on MTases and other experimental work directed to the construction of restriction enzymes with new DNA specificities.

Despite all the important findings, the MG438 crystal structure retrieved very few clues about the functional role of this orphan protein in *M. genitalium*. Surprisingly, MG438 transposon mutants (mg438$^-$ mutants) revealed that *mg438* is not an essential gene given that mycoplasma cells were viable and behaved apparently like wild-type cells. Also the protein profile of the mg438$^-$ mutant reveals no additional changes to the absence of the

MG438 protein band in comparison with wild-type cells (Pich, O. Q., Piñol, J.; personal communication). Immuno-EM analysis indicates that MG438 protein localizes in the cell cytoplasm, in a region just below and near to the membrane (Pich, O. Q.; unpublished results). Another intriguing finding is the fact that MG438 forms oligomers with discrete molecular weight at nearly physiological pH which can be a clue to its potential multifunctionality, but this remains to be proved by further experimental work.

3. Experimental Procedures

In this section some experimental details will be given that concern the work performed specifically with the MG438 protein. General methods will be described on Chapter IV.

3.1. Cloning

The *mg438* gene was cloned between the *Nde*I and *Bam*HI restriction sites of the expression vector pET19-b (Novagen, Madison, WI, USA) for heterologous expression in *E. coli* in collaboration with the laboratory of Doctors Enrique Querol and Jaume Piñol (Institut de Biotecnologia i de Biomedicina, UAB) by Oscar Q. Pich. The *mg438* gene contained two TGA that had to be mutated to TGG for recombinant expression of MG438 in *E. coli* (Calisto, Pich et al. 2005). The final pET19b-MG438 construct carries an N-terminal His_9-tag sequence followed by an Enterokinase cleavage site.

3.2. Purification of the recombinant MG438 protein

Four liters of a 90 min induced *E. coli* BL21(DE3) pLysS culture containing the pET19b-MG438 construct were harvested by centrifugation at 5000 *g* for 15 min at 4ºC and resuspended in 50 mL binding buffer (0.02 M Tris-HCl pH 7.9, 5 mM imidazole, 0.5 M NaCl, 1 mM PMSF and 20 μg/ml DNase I). Cells were lysed by sonication (5x30 seconds, 30 W) and the lysate centrifuged at 39000 *g* for 20 min. The recombinant protein present in the cleared lysate was purified according to the pET instruction manual (Novagen, Madison, WI, USA). The purified protein was dialyzed against 0.02 M Tris-HCl pH 7.4, 0.5 M NaCl. The total protein yield was 20 mg of recombinant protein per liter of culture. Protein aliquots were stored at –80 C.

3.3. MG438 crystallization

MG438 protein solution concentrated to 10 mg/mL was subjected to high-throughput sparse-matrix crystallization screenings at 20 ºC. Nano-sized crystals were obtained in the Wizard™ I screen (Emerald BioSystems, WA, USA) condition 0.1 M tri-sodium citrate pH 5.5 containing 2.0 M ammonium sulphate. Crystal growth was then optimized by changing both the crystallization buffer pH and the precipitant concentration. Large and thin needles grew within 7 days to a maximum size of 0.05 × 0.05 × 0.8 mm. These crystals were diffracted in the in-house rotating anode X-ray generator Rigaku 007 (Automated Crystallography Platform of the Barcelona Scientific

Park) in presence of different cryoprotectants such as Paratone-N, 2-methyl-2,4-pentanediol (MPD), glucose or glycerol. The best diffracting crystals were measured in the ESRF ID13 beam line and diffracted up to 7 Å resolution.

MG438 also crystallized in the Index screen (Hampton Research, CA, USA) condition 0.1 M Bis-Tris pH 5.0, 3.0 M NaCl and this was also optimized in order to obtain better crystals, improved by means of changing the crystallization buffer composition and pH to 0.1 M sodium acetate at pH 5.0, lowering the precipitant concentration and changing the drop volume and the protein:reservoir solution ratio. A screening of additives was the last step of crystal optimization and the addition of EDTA disodium salt dihydrate to the previous crystallization condition yielded crystals that diffracted to ~3 Å resolution in the ESRF BM16 beam line.

3.4. Preparation of MG438 heavy-atom derivative crystals

Three of the heavy-atom compounds used to derivatize the native protein crystals were selected within the historically successful so-called 'magic seven' (Boggon and Shapiro 2000): $HgCl_2$, K_2PtCl_4 and $KAu(CN)_2$. To increase the possibilities of getting a heavy-atom crystal derivative other compounds were selected by using the predictions of the Heavy-Atom Database System server (Sugahara, Asada et al. 2005), avoiding the conventional procedure for derivatives preparation that is usually a time-consuming 'trial-and-error' process. The database search suggested the mercury compounds ehtylmercurythiosalicylate (known as thiomersal), $HgCl_2$ and $Hg(CH_3COO)_2$ and $K_2Pt(CN)_4$, $K_2Os_4Cl_6$ and $GdCl_3$ as heavy-atom compounds with potential to derivatize the native protein crystals based on the protein amino acidic sequence and crystallization conditions. Other classes of heavy-atoms were also tested to have broader heavy-atom derivatives available for phasing: the multi-metal cluster complex $Ta_6Br_{14}.8H_2O$ (Thygesen, Weinstein et al. 1996) and the halide salts KBr and NaI (Dauter, Dauter et al. 2000).

In general, native protein crystals were soaked in heavy-atom harvesting solutions which contained the same components as the reservoir solution 1.2× concentrated. After incubating the crystals in 0.1-20 mM heavy-atom harvesting solutions for periods ranging from 4 to 72 h, they were transferred to a heavy-atom-free harvesting solution containing 20 % (v/v) glycerol, because back-soaking and cryoprotection were intended.

The conditions where crystals were not apparently severely damaged after soaking were selected to further improve soaking conditions by changing incubation times and heavy-atom compounds concentration.

Crystals soaked in the different heavy-atom harvesting solutions were diffracted at the ESRF ID13 fixed energy (0.975 Å) and BM16 bending magnet stations, which allowed diffracting the derivative crystals on the absorption edge of the derivatising heavy-atom compound. The collected data sets were integrated with MOSFLM (Leslie 1992) and scaled with SCALA (Collaborative Computational Project 1994).

3.5. SeMet-labelled MG438 crystallization

SeMet-labelled MG438 protein was produced from the pET19b-MG438 plasmid in the non-auxotroph *E.coli* BL21(DE3) pLysS cells by adding SeMet to the culture medium (Doublie, S., 1997) and purified following the same protocol used to obtain the native protein.

Crystals of the SeMet-MG438 protein were grown at 20 ºC using the hanging-drop vapour-diffusion method. The protein concentration was 7.5 mg/mL in 0.02 M Tris-HCl pH 7.4, 1.2 M NaCl. Derivative crystals grew within a week to their maximum size, of about 0.2 × 0.2 × 0.15 mm, after mixing protein, reservoir solution (0.1 M tri-sodium citrate pH 5.0, 2.2 M NaCl) and additive (0.1 M EDTA disodium salt dihydrate) in a 1:1:0.25 ratio. The crystals belonged to the trigonal space group $P3_121$ with unit cell parameters a=b=76.6 Å, c=174.8 Å and had an estimated solvent content of 64 %, consistent with one protein molecule per asymmetric unit.

3.6. Data collection and processing

X-ray diffraction intensities were measured at 100 K from crystals flash-frozen in mother-liquor containing 20 % (v/v) glycerol. Complete datasets were collected using synchrotron radiation at the European Synchrotron Radiation Facility (ESRF, Grenoble, France) ID23eh1 beam line.

Fluorescence data at the Se K absorption edge (Figure II.16) allowed the determination of the scattering factors *f'* and *f''* values and their associated optimal wavelength for anomalous data collection (Table II.3).

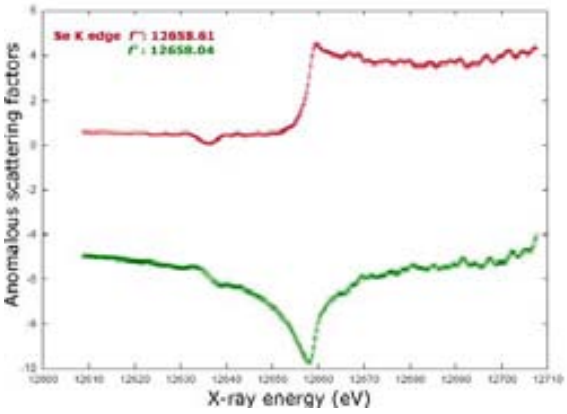


Figure II.16. Fluorescence spectrum of SeMet-labelled MG438 crystals. The values of real (*f'*) and imaginary (*f''*) scattering components as a function of incident photon energy are also shown. Figure produced using CHOOCH (Evans and Pettifer 2001).

Table II.3. Anomalous scattering factors.

Dataset	X-ray energy (eV)	Wavelength (Å)	*f'*	*f''*
Peak	12658.04	0.97935	-8.28	4.52
Inflection	12658.61	0.97950	-9.74	3.06
High energy remote	12708.00	0.97565	-4.09	4.32

Datasets were integrated with MOSFLM (Leslie 1992) and scaled with SCALA (Collaborative Computational Project 1994).

For the initial phasing, three-wavelength MAD data from a single SeMet crystal were collected, with complete anomalous information to 2.5 Å resolution. For refinement high resolution data to 2.3 Å was used. However, these refinement data were also from a SeMet crystal as only about 3 Å resolution could be obtained from the best diffracting native crystals.

3.7. Structure determination and refinement

Coordinates for six Se atoms, over a total of six possible sites, were found using both SHELXD (Schneider, T. R.; 2002) and SOLVE (Terwilliger, T. C.; 1999) softwares with MAD data using the scattering factors given on Table II.3. The resulting phases, with an overall figure of 0.31, were input into RESOLVE using a cut-off resolution of 2.6 Å (Terwilliger, T. C.; 2000) and density modification. In particular, solvent flattening was applied yielding a map with excellent connectivity.

Semi-automatic model building and structural refinement using data to 2.3 Å resolution was carried out with the software package ARP/wARP 6.0 (Perrakis, A.; 1999) and a preliminary model was successfully calculated which was 66 % complete. Several

rounds of model refinement with REFMAC5 (Murshudov, G. N.; 1997) and manual model rebuilding/refitting with TURBO/FRODO (Roussel and Cambillau 1989) resulted in a final model with good stereochemistry.

Structure factors and coordinates were deposited in the RCBS Protein Data Bank under accession code number 1YDX.

CHAPTER III

MG200, a terminal organelle motility protein

CHAPTER III. MG200, a terminal organelle motility protein

1. Introduction

As was already remarked, *M. genitalium* hides a complex cytoskeleton that shapes and polarizes the cell behind its apparent simplicity. Mycoplasma cells show a differentiated terminal organelle that is involved in key biological processes, such as locomotion by gliding across solid surfaces (Hasselbring and Krause 2007; Burgos, Pich et al. 2008). However, little is known about the mechanism of this locomotive activity in part because surface adhesion is a prerequisite for gliding motility and because many proteins involved in it are also directly or indirectly related with cell adhesion (Hasselbring, Page et al. 2006).

Some *M. genitalium* gliding deficient mutants have been transposon-generated. After isolation and characterization of these mutants *mg200* and *mg386* genes were determined to be involved in gliding motility but not in cytadherence (Pich, Burgos et al. 2006). In addition, terminal organelle ultrastructure remains unaffected. Further complementation of the deficient mutants with their respective *mg200* and *mg386* wild-type copies resulted in restoration of the gliding phenotype. Orthologues of these proteins were also found in other mycoplasmas from the *pneumoniae* cluster, such as *M. pneumoniae* and *M. gallisepticum*, suggesting the existence of specific motility machinery.

Mycoplasmas present a high number of multi-domain proteins resembling more what happens in eukaryotes where 60-80 % of the proteins have multiple domains, while in prokaryotes these constitute only ~40 % of the proteome (Teichmann, Park et al. 1998; Han, Batey et al. 2007).

M. genitalium MG200 protein presents a particularly complex domains organization being constituted by a J domain at its N-terminal, a central region that accommodates an EAGRb followed by an APRd and a final C-terminal domain.

J domains are present in the DnaJ family of proteins and are involved in the modulation of the ATPase activity of the molecular chaperone DnaK or Hsp70 (Hennessy,

Cheetham et al. 2000; Walsh, Bursac et al. 2004). Molecular chaperones have also been described to participate in motility mechanisms in other microorganism. In *E. coli*, DnaK, DnaJ and GrpE chaperones are required for flagellum synthesis (Shi, Zhou et al. 1992). In the non-flagellar microorganism *Myxococcus xanthus* a DnaK protein homologue is needed for motility and, in this case, the gene is not regulated by heat-shock (Yang, Geng et al. 1998).

In addition to *mg200*, *M. genitalium* possesses two extra paralog genes, *mg002* and *mg019*, coding for proteins with a J domain at their N-terminal regions. The majority of the DnaJ proteins are 300 to 400 amino acids long. This characteristic is shared by MG002 and MG019 while MG200 consists of a 601 amino acids long polypeptide. From the *M. genitalium* J domain-containing proteins only MG019 has a J domain linked to a glycine repeat (G repeat) region and a zinc-finger domain, as described for type I J proteins, in addition to a regulatory sequence CIRCE (Controlling Inverted Repeat of Chaperone Expression) within its promoter region. Consequently, MG019 is believed to be the true homologue of the DnaJ chaperone in *M. genitalium*. Surprisingly, from the *M. genitalium* proteins belonging to the DnaJ protein family only MG200 was found to be up-regulated after heat-shock (Musatovova, Dhandayuthapani et al. 2006), but the significance of such up-regulation remains unclear.

MG200 was tentatively classified into the type III J proteins group, which lack both the G repeat and the zinc-finger domain and represent a very diverse and functionally distinct group of proteins. Type III J proteins are currently thought not to have a role as molecular chaperones (Walsh, Bursac et al. 2004).

After the J domain the MG200 protein accommodates a small globular domain remarkably enriched in aromatic and glycine residues, the EAGRb. Such a domain is exclusive of mycoplasmas appearing in proteins related with gliding motility and/or with the terminal organelle. To date, a total of 32 EAGR boxes were identified in *M. genitalium* MG200, MG312 and MG386 proteins and in its protein homologues from *M. pneumoniae* (TopJ, HMW1 and P200, respectively), and *M. gallisepticum* (MGA1228, MGA0205 and MGA0306, respectively). In addition, deletion of a small region including the MG312-EAGRb revealed the specific role of this small domain in

gliding motility (Burgos, Pich et al. 2007) although there is still no data on the specific function of this motif.

In the MG200 amino acidic sequence an APRd follows the EAGR box. This domain is a common feature of several mycoplasma terminal organelle proteins (MG217, MG312, MG317, and MG386; see Table I.3). Proline-rich domains are also present in a high number of eukaryotic proteins and are presumably involved in protein-protein interactions (Kay, Williamson et al. 2000). In the context of mycoplasma motility, it is particularly interesting that some proteins containing proline-rich regions have been found to have a role in the regulation of actin polymerization, activity that is closely related to cellular motility (Holt and Koffer 2001).

The MG200 protein C-terminal domain, which is ~308 amino acids long, has only very weak similarity with its orthologues from *M. pneumoniae* and *M. gallisepticum*, TopJ and MGA0205 proteins, respectively. So far no protein homologue was encountered for this domain.

2. Results and Discussion

2.1. Full-length MG200 protein production and characterisation

M. genitalium MG200 protein was cloned in the laboratory of Doctors Enrique Querol and Jaume Piñol (Institut de Biotecnologia i de Biomedicina, UAB) by Oscar Q. Pich. The full-length MG200 (*fl*MG200) protein expressed in the most standard conditions was totally insoluble but after several optimization assays soluble protein was observed, when the expression was undertaken at 16 ºC. To assure that the major protein band, observed in the cell extract analysed by SDS-PAGE, corresponded to the interest protein a western blot was performed with an anti-His-tag antibody. Surprisingly, several protein bands were found to react with the antibody that can probably correspond to MG200 degradation products (Figure III.1A-B).

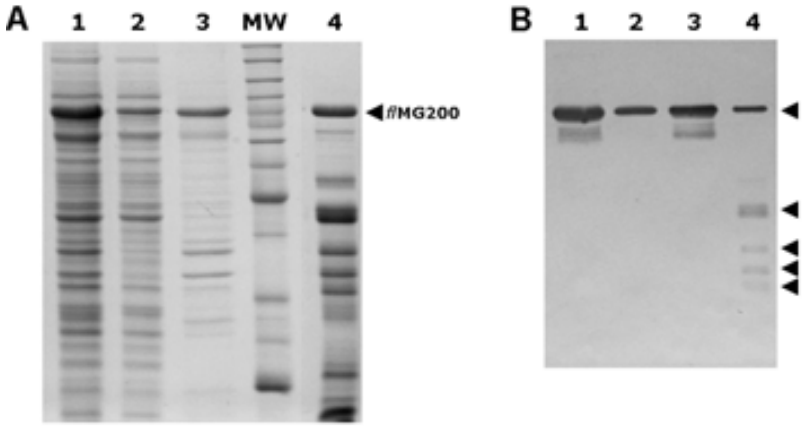


Figure III.1. SDS-PAGE and western blot analysis of IPTG-induced expression of pET19b-*fl*MG200 in *E. coli* BL21(DE3). A) CBB-stained 10 % (v/v) SDS-PAGE and (B) western blot (revealed with an anti-His-tag antibody) analysis of the total protein extract and of its soluble and insoluble fractions (lanes 1-3 in both panels, respectively). Lane 4, in both panels, corresponds to the analysis of *fl*MG200 purified through Ni^{2+}-affinity chromatography. *fl*MG200 is indicated as well as *fl*MG200 degradation products, which are marked by black arrowheads. Results obtained in the laboratory of Doctors Enrique Querol and Jaume Piñol (Institut de Biotecnologia i Biomedicina, UAB) by Oscar Q. Pich.

The electrophoretic mobility of the *fl*MG200 protein is that of a 90 kDa protein (Figure III.1) when its theoretical molecular weight is of 71 kDa. This abnormal electrophoretic mobility is frequently observed in proline-rich proteins that generally run in a denaturing gel with a molecular weight higher than the expected.

To further investigate the identity of the major expression band and the nature of the degradation bands and also given the relatively high amount of protein produced, the protein peak recovered from Ni^{2+}-affinity chromatography was loaded into a Superdex 200 10/300 GFC column (GE Healthcare Life Sciences, Uppsala, Sweden). The elution profile is quite heterogeneous and several peaks can be appreciated that were not completely separated during the chromatography (Figure III.2A). In addition, SDS-PAGE analysis of fractions collected during sample elution revealed that indeed no highly pure protein fraction could be recovered and also that degradation products are quite stable and appear as discrete bands in polyacrylamide denaturing gels (Figure III.2B). This sample aggregated and precipitated when taken to concentrations higher than 1 mg/mL.

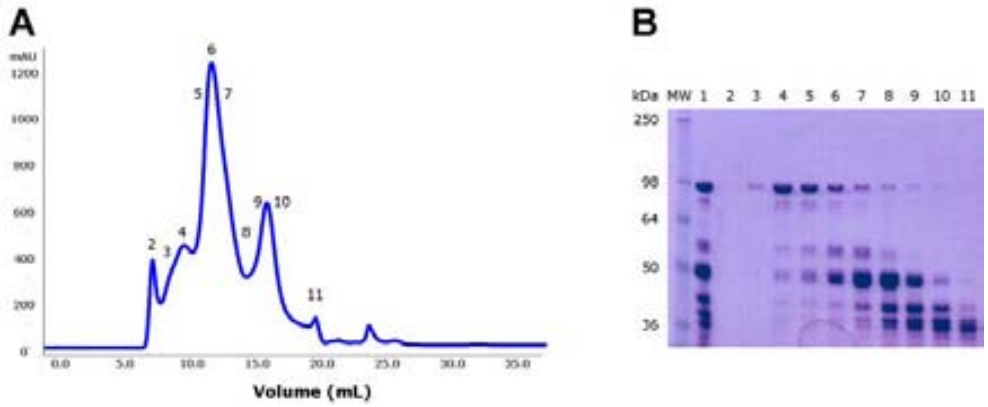


Figure III.2. Purification by GFC of the *fl*MG200 protein and analysis of the eluted fractions by SDS-PAGE. A) Elution profile of *fl*MG200 protein on a Superdex 200 10/300 GFC column. B) 12 % (v/v) SDS-PAGE analysis of the *fl*MG200 protein purified through Ni^{2+}-affinity chromatography (lane 1) and of the peak fractions collected from protein elution (lanes 2-11). Molecular weights (MW) of protein standards are indicated.

It is already well known that such heterogeneity level, as found for the *fl*MG200 protein, is adverse to the proper crystallization of protein solutions and so, new attempts to obtain a highly pure and homogeneous sample followed.

The complete *mg200* gene sequence was cloned into a pET21d expression vector that promoted the production of *fl*MG200 with a His_6-tag fused to its C-terminal. The hypothetical degradation bands generated by the production of *fl*MG200 from this construct could be sequenced from its N-terminal ends, likely revealing the exact site where the protein is being degraded.

The *fl*MG200-His_6 protein was expressed and purified following the same protocol used to produce His_9-*fl*MG200. Analogously, total protein extract and its soluble and insoluble fractions were analysed by SDS-PAGE and western blot, as well as the sample eluted with high imidazole concentration from the equilibration of the cells soluble extract with a Ni^{2+}-affinity slurry (Figure III.3A-B). Although degradation is still apparent it seems to be greatly diminished when *fl*MG200 is expressed with the His-tag fused to its C-terminal rather than to its N-terminal end (Figure III.1). As before, the protein solution obtained after the Ni^{2+}-affinity purification step was loaded into a GFC column. In this case, the lower molecular weight degradation bands were effectively separated from the full-length protein and a 95 % pure protein solution was obtained.

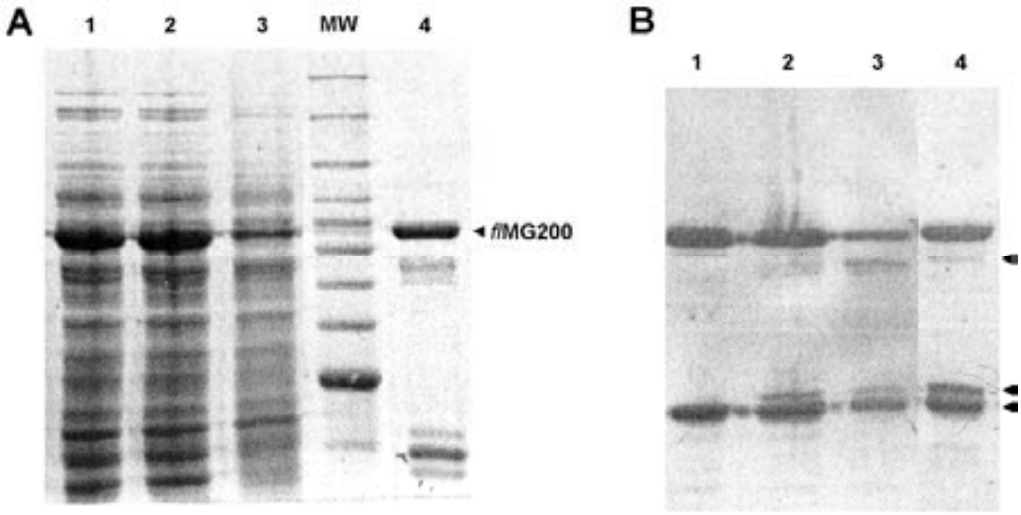


Figure III.3. SDS-PAGE and western blot analysis of IPTG-induced expression of pET21d-*fl*MG200 in *E. coli* BL21(DE3). A) CBB-stained 10 % (v/v) SDS-PAGE and (B) western blot (revealed with an anti-His-tag antibody) analysis of the total protein extract and of its soluble and insoluble fractions (lanes 1-3 in both panels, respectively) obtained from an *E. coli* BL21(DE3) pET21d-*fl*MG200 culture. Lane 4, in both panels, corresponds to the analysis of *fl*MG200 purified in a Ni^{2+}-affinity slurry. *Fl*MG200 is indicated as well as *fl*MG200 degradation products, which are marked by black arrowheads. Results obtained in the laboratory of Doctors Enrique Querol and Jaume Piñol (Institut de Biotecnologia i Biomedicina, UAB) by Jaume Piñol.

Given these promising results, *fl*MG200 production was over-scaled for further sample characterisation and crystallization studies. Gel filtration chromatograms of this sample present three protein peaks (Figure III.4A):

i) peak (a) is eluted in the column void-volume, corresponding to an aggregate of *fl*MG200 complexed with fragments resultant from its own degradation;

ii) peak (b) has an estimated molecular weight of 250 kDa corresponding to a tetrameric oligomeric state of the *fl*MG200 protein. Its SDS-PAGE analysis reveals that an almost imperceptible protein band still persists in the most pure *fl*MG200 fraction (Figure III.4B-*d*), which have a molecular weight slightly lower than it and is detected by anti-His-tag antibody meaning it is most likely a *fl*MG200 degradation product;

iii) peak (c) fractions analysis by SDS-PAGE revealed the presence of two protein bands (Figure III.4B-*e*, and *f*) which also react with the anti-His-tag antibody, being most likely *fl*MG200 degradation products.

The four protein bands detected by SDS-PAGE from the analysis of the GFC peaks were subjected to N-terminal EDMAN sequencing. Protein bands *d* and *e* could not be sequenced most probably because they are in very little quantities. On the contrary, the identity of the band that was thought to correspond to the full-length protein was confirmed and band *f* sequence match with MG200 protein amino acidic sequence starting on amino acid Leu299.

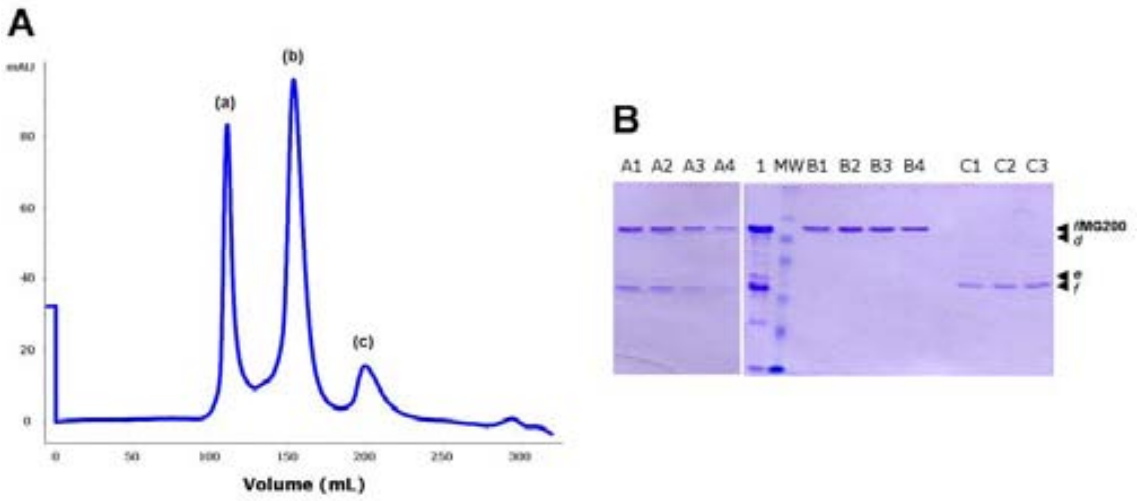


Figure III.4. Purification of the *fl*MG200 protein through a GFC column and analysis of the eluted fractions by SDS-PAGE. A) Elution profile of *fl*MG200 protein on a Superdex 200 26/60 HL column: peak (a) is eluted in the column void volume, peak (b) is eluted with a molecular weight of approximately 250 kDa (up to 4 *fl*MG200 molecules) and peak (c) is composed of two degradation products of the *f*MG200 protein. B) 12 % (v/v) SDS-PAGE analysis of peak fractions from the purification of *fl*MG200 by GFC. Fractions from peaks (a), (b) and (c) correspond to lanes A1-A4, B1-B4 and C1-C3, respectively. In lane 1 an aliquot of *fl*MG200 before purification through GFC was loaded. Absorbance was registered at 280 nm.

Pure *fl*MG200 protein solution was further analysed after being concentrated to 3-4 mg/mL or stored for several days at 4 ºC, rather by GFC or DLS. These experiments revealed that the protein solution aggregated although no precipitate could be appreciated. In order to determine if there are some compounds capable of further stabilize *fl*MG200 protein in solution a new protein batch was produced and divided into several aliquots to which different compounds were added, like glycerol and some detergents. Analysis of these aliquots by GFC indicated that no significant improvement came from the addition of these potential protein stabilizers. Glycerol was the only additive to slightly improve protein stability for longer periods of time although its polydispersity (35 %), evaluated by DLS, is still high. As the doubt persists whether sample instability was due to the partial protein degradation that contaminated the *fl*MG200 protein solution, this sample was further purified through an ion-exchange chromatography column (Figure III.5A-B). A single and sharp peak resulted from this analysis and, as confirmed by SDS-PAGE, the degradation band persists contaminating the *fl*MG200 protein solution.

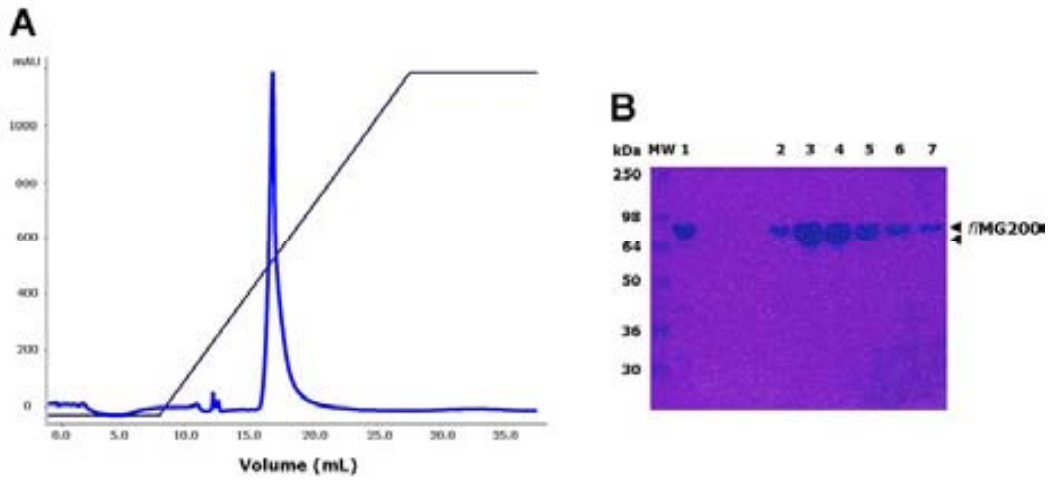


Figure III.5. Purification of the *fl*MG200 protein through an ion-exchange chromatography column and analysis of the eluted fractions by SDS-PAGE. A) Elution profile of *fl*MG200 (blue) on a MonoQ 5/50 GL column by applying a linear gradient of 0-1 M NaCl (black). Absorbance was registered at 280 nm. B) 12 % (v/v) SDS-PAGE of *fl*MG200 before purification through the ion-exchange column (lane 1) and of the peak fractions collected after protein elution (lanes 2-7). *Fl*MG200 is indicated as well as the contaminant *fl*MG200 degradation product, which is marked by a black arrowhead. Molecular weights (MW) of protein standards are indicated.

Nevertheless, studies were continued in order to obtain a more stable and homogeneous protein sample and the DSF technique was used for that purpose. *Fl*MG200 protein was determined to have a temperature of melting (*T*m) of 35 ºC in 0.02 M Tris-HCl pH 8.0,

0.2 M NaCl buffer, which is rather low for a globular protein from a human pathogen microorganism. After intense compound screening to identify some that could improve the interest protein stability, only 0.1 M potassium phosphate pH 7.0, 0.2 M NaCl buffer was found to stabilize the protein in 5 ºC, being 40 ºC the *fl*MG200 protein temperature of melting in this condition. This result was cross-checked by running the sample in a native PAGE after heating sample aliquots for several minutes at regular temperature intervals, from 25 to 60 ºC (Figure III.6). Analysis of the native polyacrylamide gel reveals that the protein starts to denature at 40 ºC being completely degraded at 45 ºC.

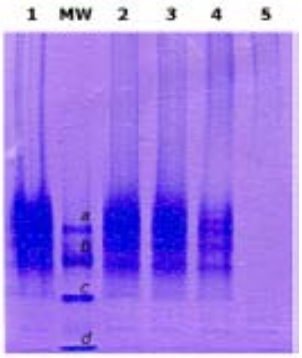


Figure III.6. Analysis of pre-heated *fl*MG200 aliquots by PAGE. CBB-stained 5 % (v/v) PAGE of *fl*MG200 aliquots treated at 25, 30, 35, 40, and 45 ºC (lanes 1 to 5, respectively) for 10 minutes. Native molecular weights (MW) for protein standards are: *a*-440, *b*-232, *c*-140 and *d*-66 kDa.

*Fl*MG200 solution was then dialyzed against the new stabilizing buffer and further analysed by CD. Qualitative analysis of the CD spectrum reveals that the protein is folded and behaves as a protein with a mixed, α-helix and β-sheet, secondary structure elements composition (Martin and Schilstra 2008) (Figure III.7A). Surprisingly, from the sample thermal denaturation following CD at 218 nm, no clear denaturation curve is apparent. Instead, there is a gain in secondary structure elements at around 43 ºC indicating a possible conformational change or protein oligomerization at around this temperature, which drives it to aggregation (Figure III.7B). This result is reinforced by the CD spectrum acquired a 55 ºC (Figure III.7A), which shows a clear change in the protein secondary structure composition at this temperature.

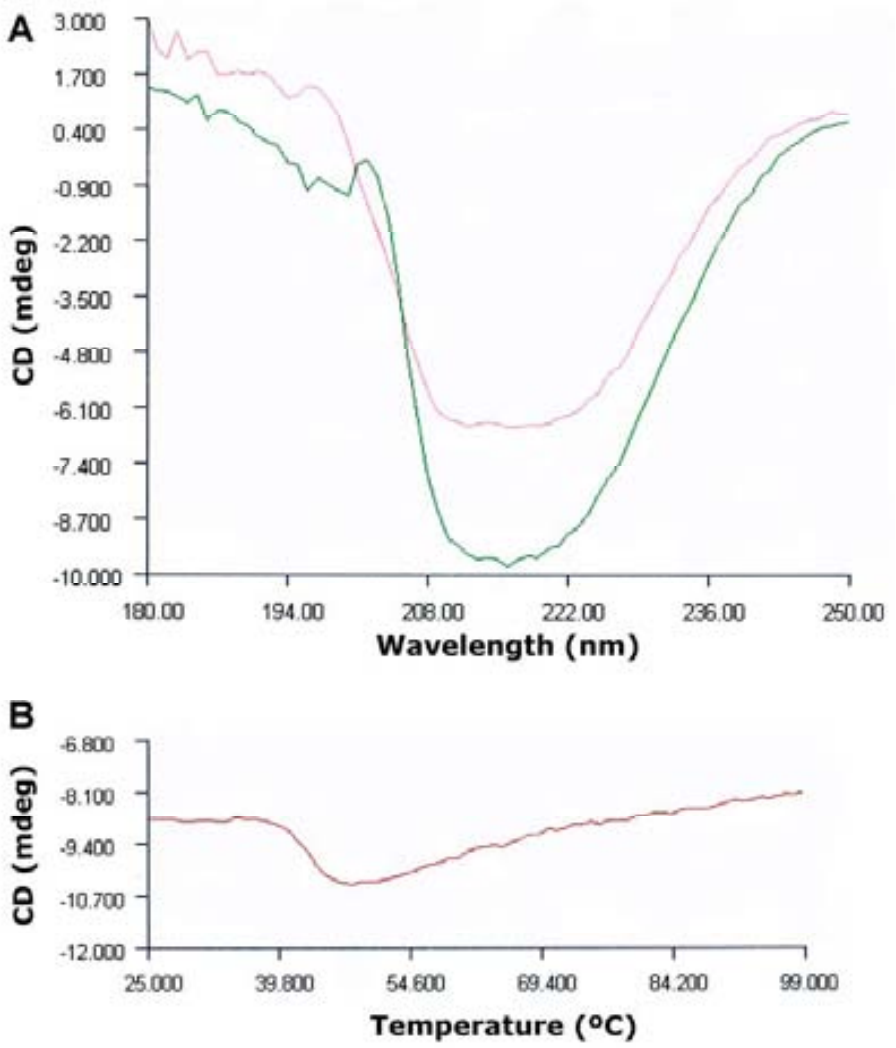


Figure III.7. Characterisation of the *fl*MG200 protein by CD spectroscopy. A) CD spectra (250-180 nm) of *fl*MG200 at 25 and 55 ºC (pink and green curves, respectively), measured at 0.1 mg/mL protein concentration. B) Thermal denaturation curve of *fl*MG200 at 0.1 mg/mL monitoring the CD signal at 218 nm from 25 to 99 ºC.

*Fl*MG200 protein solution was also analysed by isoelectric focusing (IEF), being first separated in a band strip with a pH range from 4 to 7 and after in a denaturant polyacrylamide gel (Figure III.8A). In the denaturant gel several protein spots were observed with only slightly different pI's. Several discrete bands were also observed when the sample was analysed in a native polyacrylamide gradient gel (Figure III.8B).

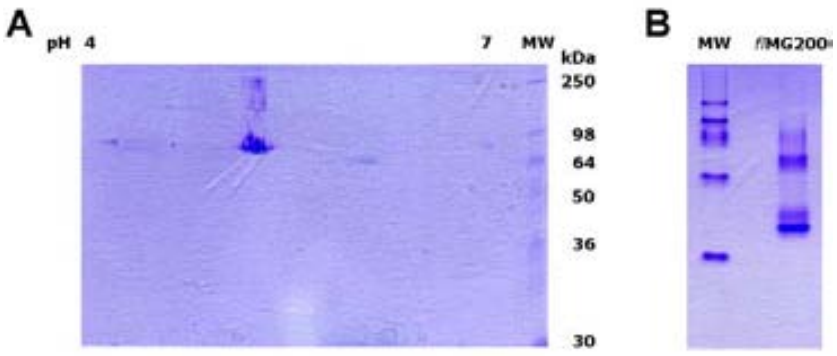


Figure III.8. *fl*MG200 analysis by IEF and PAGE. A) IEF profile of *fl*MG200 where several discrete protein spots can be observed around pH 5.0. Molecular weights (MW) of protein standards are indicated. B) CBB-stained (4-20 % (v/v)) PAGE of *fl*MG200 protein where several bands can be observed corresponding to *fl*MG200 protein variants.

These results stress out that the *fl*MG200 protein solution is highly heterogeneous. The observed heterogeneity can be both due to conformational heterogeneity, related to intrinsic disordered regions of the protein, or chemical heterogeneity, related to the protein partial degradation. This confirms the impossibility to technically purify the sample to more than 95 % with reduced polydispersity as is ideal to perform crystallization assays.

Meanwhile, all the attempts to crystallize this sample failed and heavy-precipitates appeared in more than 90 % of the screened conditions. At this stage, a new approach to the structural study of the MG200 protein was designed that will be described in the forthcoming sections of this chapter.

2.2. Expression and purification of the MG200 protein domains

Several constructs were designed to screen for soluble MG200 protein domains with improved crystallization propensity. To define the domains frontiers as accurately as possible a careful secondary structure prediction of the MG200 protein was performed using the server MeDor (Lieutaud, Canard et al. 2008). This information, together with the analysis from ClustalW multiple-sequence alignments (Larkin, Blackshields et al. 2007) for each of the protein domains, led to the design of a set of constructs with diverse domain variants as defined in Figure III.9 and with the physico-chemical properties described in Table III.1.

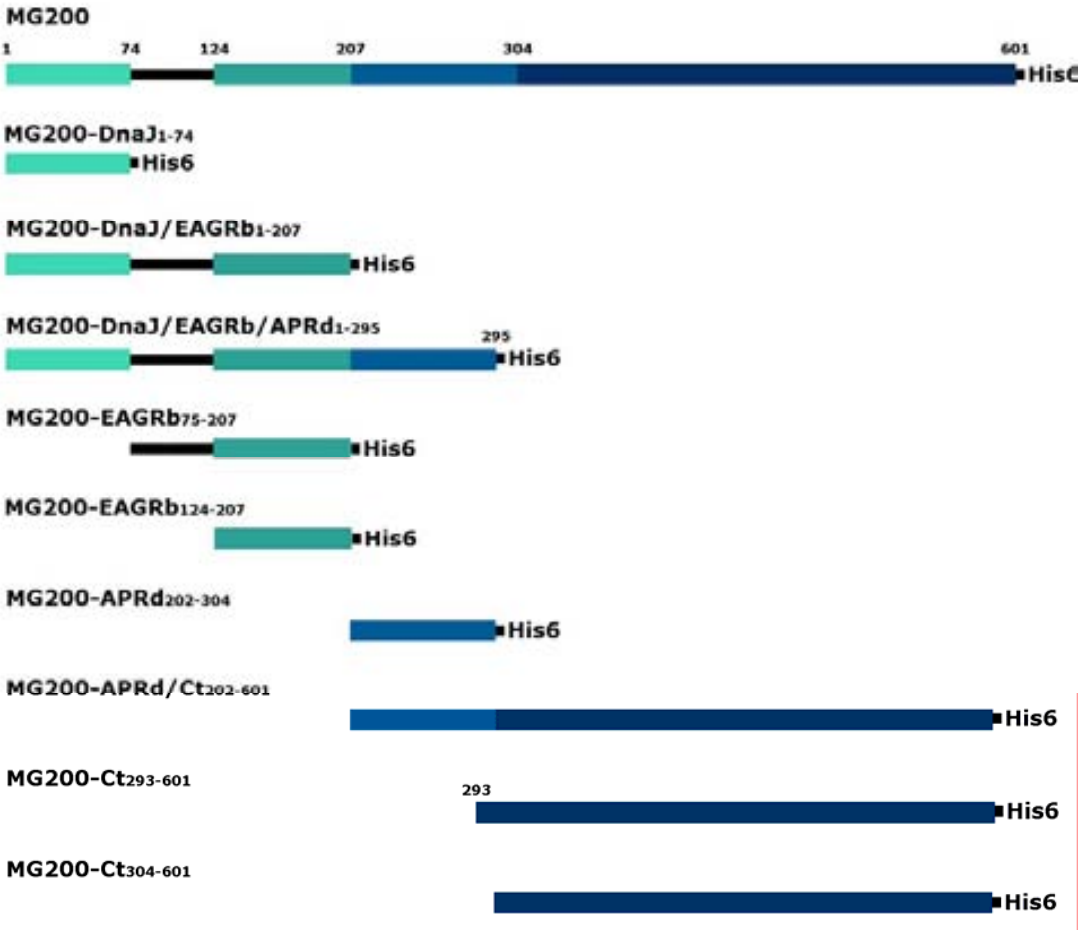


Figure III.9. MG200 domain architecture and construct design of a set of MG200 domain variants. The positions of the DnaJ, EAGRb, APRd and C-terminal domains of the full-length protein are indicated on the first lane. Residue numbers for N- and C-terminal ends are also indicated.

Table III.1. Physico-chemical properties of MG200 domain variants.

MG200 domain variants	MW (kDa)	pI	Amino acidic content (%)			Nr. residues
			Arg+Lys	Asp+Glu	Ala+Gly+Pro	
***fl*MG200**	68.5	4.9	11.5	16.0	17.0	601
MG200-DnaJ$_{1-74}$	8.5	7.9	18.9	17.6	23.0	74
MG200-DnaJ/EAGRb$_{1-207}$	24.0	5.2	15.0	18.4	18.8	207
MG200-DnaJ/EAGRb/APRd$_{1-295}$	33.5	4.6	11.2	19.0	22.0	295
MG200-EAGRb$_{75-207}$	15.6	4.8	12.8	19.8	15.1	133
MG200-EAGRb$_{124-207}$	9.8	4.4	10.7	21.4	14.3	84
MG200-APRd$_{202-304}$	11.1	3.4	1.9	22.4	28.1	103
MG200-APRd/Ct$_{202-601}$	45.0	4.7	9.5	15.0	17.1	400
MG200-Ct$_{293-601}$	35.2	5.8	11.9	12.9	13.0	309
MG200-Ct$_{304-601}$	34.0	6.4	12.1	12.4	13.0	298
Globular proteins			11.4	11.7	19.8	

A common protocol was established to express and purify all the MG200 protein domains and the yield of each expression and purification experiment is indicated on Table III.2.

Table III.2. Protein expression and purification yield for each of the MG200 domain variants.

MG200 domain variants	Expression[‡]	Yield[†]
MG200-DnaJ$_{1-74}$	S	10
MG200-DnaJ/EAGRb$_{1-207}$	S	60
MG200-DnaJ/EAGRb/APRd$_{1-295}$	NE	-
MG200-EAGRb$_{75-207}$	S	30
MG200-EAGRb$_{124-207}$	S	150
MG200-APRd$_{202-304}$	NE	-
MG200-APRd/Ct$_{202-601}$	NE	-
MG200-Ct$_{293-601}$	S	0.5
MG200-Ct$_{304-601}$	S	2

[‡] S – soluble, NE – Domain expression was not observed;
[†] Protein extraction yield in mg per liter of cell culture.

2.2.1. Expression of the MG200 APR domain

All MG200 domain variants containing the APRd, MG200-DnaJ/EAGRb/APRd$_{1-295}$, MG200-APRd$_{202-304}$ and MG200-APRd/Ct$_{202-601}$, were not expressed in the tested conditions. APRd is characterized by its high proline content and, moreover, proline residues within the domain are not organized in a recognizable pattern. This is usually associated to unstructured/disordered protein regions.

2.2.2. Expression and purification of the MG200 C-terminal domain

MG200 C-terminal domain over-expression was very poor and the produced protein was rather unstable and precipitated when taken to concentrations higher than 0.5 mg/mL or heated-up to room temperature. In an attempt to improve the domain production yield and the quality of the produced protein *E. coli* BL21(DE3) cells were co-transformed with the pET21d-MG200-Ct$_{293-601}$ or pET21d-MG200-Ct$_{304-601}$ constructs and pGro7, an expression vector for the *E. coli* GroEL and GroES chaperones which catalyze the correct folding of newly synthesized polypeptides (Yasukawa and Kanei-Ishii 1995).

The better expression and purification conditions for the C-terminal domain are the MG200-Ct$_{304-601}$ domain co-expression with the *E. coli* chaperones and its purification in presence of 10 % (v/v) glycerol. The thermal melt for the MG200-Ct$_{304-601}$ domain prepared in the above described conditions is of 40 ºC in 0.02 M potassium phosphate pH 7.0, 0.2 M NaCl buffer containing 10 % (v/v) glycerol, being equivalent to the *fl*MG200 protein temperature of melting. Moreover, molecular-weight calibration of the GFC column through which MG200-Ct$_{304-601}$ was purified suggests that it has a molecular weight of ~130 kDa, in agreement with the existence of a tetrameric form of

the domain. If this oligomeric state is also found in the quaternary structural arrangement of full-length protein the C-terminal domain is likely to be the protein tetramerization domain, which could be relevant for the protein functional role.

2.3. Crystallization of the MG200 C-terminal domain

Initial crystallization screening of the MG200 C-terminal domain resulted in more than 95 % precipitated drops even when screened at relatively low protein concentrations (3-4 mg/mL). However, crystals were obtained in nanoliter-scale drops (Figure III.10A) and these crystallization conditions were used as a starting point for crystal growth optimization. This was undertaken settling microliter-scale drops and implied the use of different crystallization techniques. The best crystals were needle-shaped clusters of crystals grown at 4 ºC (Figure III.10B) in presence of microseeds of MG200 C-terminal domain crystals. To check if the crystalline clusters were indeed protein crystals several of them were taken from the drop, washed in protein-free reservoir solution and analysed by SDS-PAGE (Figure III.10C). The denaturant polyacrylamide gel confirms that the grown needle-shaped clusters of crystals are of the MG200 C-terminal domain and efforts are still being done in order to obtain single crystals.

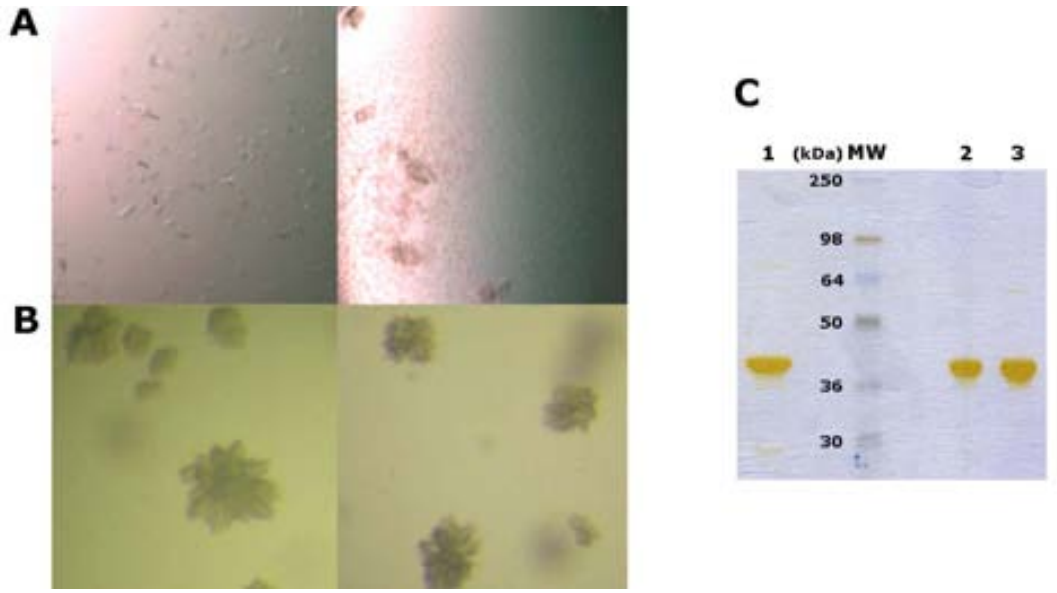


Figure III.10. MG200-$Ct_{304-601}$ crystals produced by the hanging-drop vapour-diffusion crystallization method and their analysis by SDS-PAGE. A) Nano-crystals produced in high-throughput crystallization plates. B) Small needle-shaped clusters of crystals resultant from the optimization of crystals in (A) produced by the microseeding method. C) Silver-stained-SDS-PAGE: lane 1, solution resultant from the dissolution of crystals in SDS-PAGE loading buffer; lane 2 and 3, control lanes where 1 and 2 μg of MG200-$Ct_{304-601}$ solution were respectively loaded into the gel. The molecular weights (MW) of protein standards are indicated.

2.4. MG200-DnaJ-containing domain variants characterisation

MG200-DnaJ$_{1-74}$ and MG200-DnaJ/EAGRb$_{1-207}$ domain variants were over-expressed as soluble fragments and purified to very high levels. Crystallization screening of both samples performed at the standard concentration of 10 mg/mL and at high concentrations (30-40 mg/mL) led only to clear drops where there were no signals of either crystals or precipitates, with phases being observed in just a few drops that contained alcohols in the reservoir solution. Considering that these MG200 domain variants are particularly highly charged (Table III.1) it was not surprising that they were so soluble. In order to diminish their solubility small amounts of organic solvents (5-10 % (v/v)), such as iso-propanol or MPD, were added to both samples and these were again screened. High salt concentrations were also added to the samples and screened with the same objective, reduce sample solubility. The new crystallization screening resulted in more than 97 % transparent drops, with heavy-precipitates being observed only in crystallization conditions containing high divalent metal ions concentrations.

To check if the MG200-DnaJ$_{1-74}$ and MG200-DnaJ/EAGRb$_{1-207}$ domain variants were folded they were analysed by NMR spectroscopy. For this, each of the domain variants was produced in ^{15}N-rich medium.

The ^{15}N-Heteronuclear Single Quantum Coherence (HSQC) spectrum of MG200-DnaJ$_{1-74}$ revealed eighty-five independent peaks that are dispersed within the all spectrum area and have uniform intensity and homogeneous and sharp appearance, all features of a folded protein (Figure III.11A). Also, CD spectroscopy proved that this domain variant corresponds to a folded domain (Figure III.11B), presenting a spectrum with two minimums, at 208 and 222 nm, characteristic of pure α-helical proteins. This is also in conformity with the folding described for MG200-DnaJ$_{1-74}$ domain homologues whose three-dimensional structure consists of four α-helices, two that are larger and organized in an antiparallel fashion surrounded by two smaller ones.

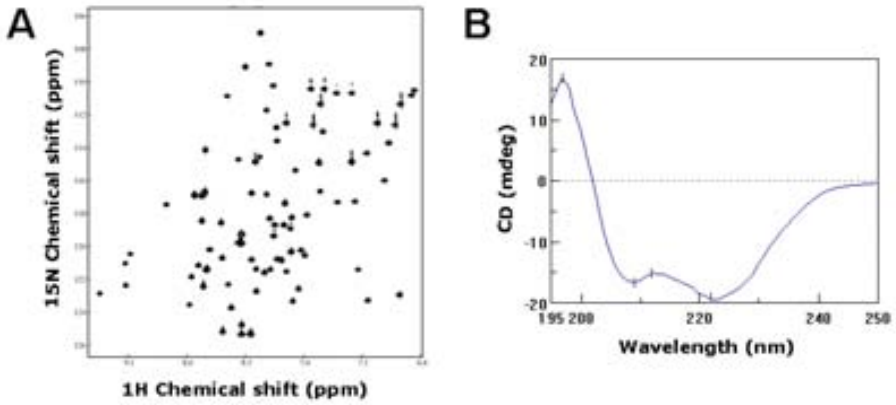


Figure III.11. MG200-DnaJ$_{1-74}$ characterisation by NMR and CD spectroscopies. A) Two-dimensional spectrum of [^{1}H, ^{15}N]-MG200-DnaJ$_{1-74}$ at 800 MHz and recorded at 25 ºC. The sample concentration was of 0.24 mM in 0.02 M potassium phosphate pH 6.5, 0.09 M NaCl and 10 % (v/v) D_2O. B) CD spectrum of MG200-DnaJ$_{1-74}$ acquired at 25 ºC from 250-195 nm at a concentration of 0.2 mg/mL in 0.02 M potassium phosphate pH 6.5, 0.1 M NaCl.

The MG200-DnaJ/EAGRb$_{1-207}$ ^{15}N-HSQC spectrum (Figure III.12A) is not as nice as the one obtained for the MG200-DnaJ$_{1-74}$ domain, presenting broad peaks of different intensity. Moreover, peaks accumulate mainly in the 8.1-8.6 ppm region of the spectrum instead of being well disperse and no peak is defined in the 10 ppm region. This region is where peaks corresponding to structured tryptophane residues appear due to the particular environment of the nitrogen atom in its particular side chain. Given these, the MG200-DnaJ/EAGRb$_{1-207}$ NMR spectrum is that of a protein that is not totally structured although a more conclusive analysis is hard to extract from the spectrum given that the MG200-DnaJ/EAGRb$_{1-207}$ domain size is in the upper limit for optimum NMR measurements.

Temperature is a parameter known to affect molecular motion and conformational dynamics of proteins in solution. In an attempt to further analyse the MG200-DnaJ/EAGRb$_{1-207}$ domain variant by NMR spectroscopy two spectra were acquired at 34 and 7 ºC. At 34 ºC MG200-DnaJ/EAGRb$_{1-207}$ NMR spectrum fingerprint is identical to the one measured at 25 ºC (Figure III.12B) indicating there are no major structural changes in the domain conformation or dynamics at this temperature. However, at 7 ºC (Figure III.12C), although no major changes are observed and peak dispersion did not improved, four new peaks appeared in the chemical shift region of the spectrum characteristic for tryptophane residues. Superposition of MG200-DnaJ$_{1-74}$ and MG200-DnaJ/EAGRb$_{1-207}$ two-dimensional NMR spectra (Figure III.12D) reveals that there is a

good match between the peaks registered for the DnaJ domain alone (MG200-DnaJ$_{1-74}$) and for the DnaJ domain included in the MG200-DnaJ/EAGRb$_{1-207}$ domain, indicating that the DnaJ domain fold is essentially the same in both domain variants. Therefore, MG200-DnaJ/EAGRbox$_{1-207}$ NMR spectrum peaks that do not match with those corresponding to the DnaJ domain are from residues complementary to this domain in the MG200-DnaJ/EAGRbox$_{1-207}$ variant. Some of these peaks are sharp and homogeneous and are expected correspond to another structured region found within the MG200-DnaJ/EAGRbox$_{1-207}$ domain variant. Furthermore, these peaks appear preferentially in the 10 ppm region of the spectrum, most likely corresponding to tryptophane residues. In MG200-DnaJ/EAGRbox$_{1-207}$ domain variant tryptophanes are concentrated at the end of the domain, between residues 169 and 201, in a region that corresponds to the EAGR box, indicating that this region could be folded. To confirm it two new MG200 domain variants were cloned including:

i) the sequence complementary to the DnaJ domain in MG200-DnaJ/EAGRbox$_{1-207}$, residues 75-207;

ii) the sequence thought to correspond to the EAGR box by multiple-sequence alignment with homologous EAGR boxes and by secondary structure prediction, residues 124-207.

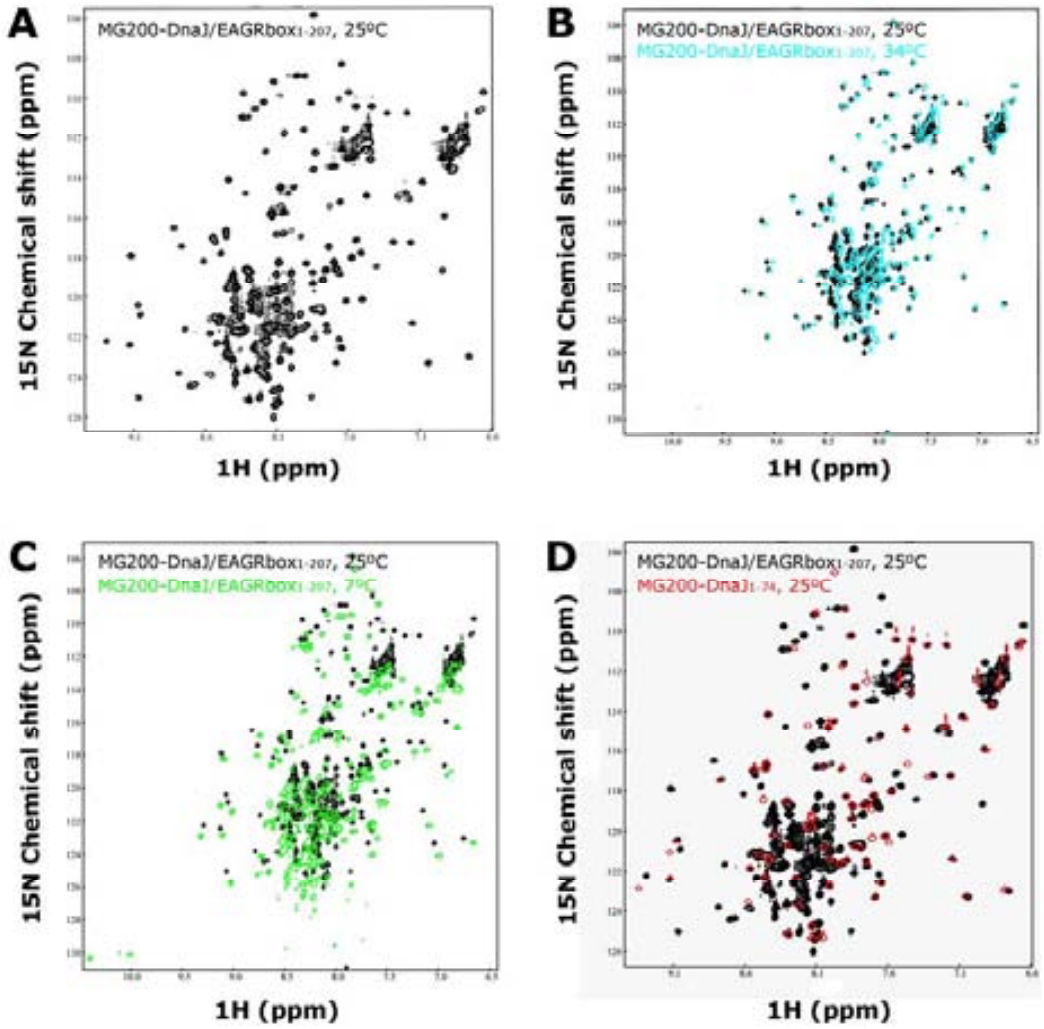


Figure III.12. [^{1}H, ^{15}N]-HSQC spectra of the MG200-DnaJ/EAGRbox$_{1-207}$ domain variant. Conventional 800 MHz [^{15}N, ^{1}H]-HSQC spectra of MG200-DnaJ/EAGRbox$_{1-207}$ in 0.02 M potassium phosphate pH 6.5, 0.09 M NaCl and 10 % (v/v) D_2O at a 0.14 mM concentration acquired at (A) 25 ºC (black), (B) 34 ºC (blue) and (C) 7 ºC (green). D) Superposition of the [^{15}N, ^{1}H]-HSQC spectra of MG200-DnaJ/EAGRbox$_{1-207}$ (black) and MG200-DnaJ$_{1-74}$ (red) both acquired at 25 ºC at concentrations of 0.14 and 0.24 mM, respectively.

2.5. MG200-EAGRb domain variants characterisation

Production of the MG200-EAGRb$_{75-207}$ and MG200-EAGRb$_{124-207}$ domain variants yielded highly pure and soluble protein solutions. These domains were analysed by one-dimensional NMR spectroscopy, although this experiment is generally crowded with overlapping signals when performed on a proteic sample and only qualitative information can be extracted from it. The MG200-EAGRb$_{75-207}$ NMR proton spectrum (Figure III.13A) indicates that this domain variant has significant unstructured/disordered regions within its sequence, revealed by the poor dispersity and resolution of the observed peaks. On the contrary, proton peaks on the MG200-EAGRb$_{124-207}$ spectrum (Figure III.13B) are more disperse, appearing as sharp and

independent peaks. This is a good indication that the domain has defined secondary structure elements. Given this, MG200-EAGRb$_{124-207}$ was produced in ^{15}N-rich medium and further analysed by two-dimensional NMR spectroscopy (Figure III.13C). In this spectrum 74 differentiated peaks were identified, which were rather intense, sharp and dispersed, indicating the domain is indeed folded.

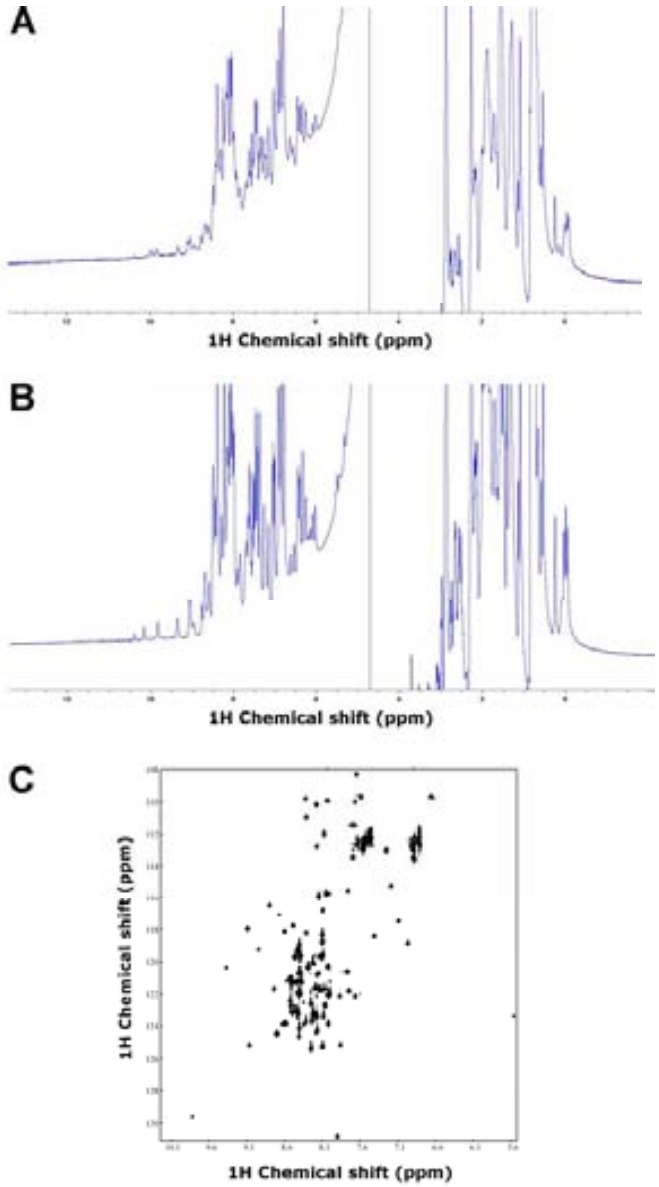


Figure III.13. MG200-EAGRb$_{75-207}$ and MG200-EAGRb$_{124-207}$ NMR spectra acquired at 600 MHz and 25 ºC. NMR proton spectra of (A) MG200-EAGRb$_{75-207}$ and (B) MG200-EAGRb$_{124-207}$ recorded at 0.38 mM and 1.20 mM protein concentrations, respectively. C) [^{15}N, ^{1}H]-HSQC spectrum of MG200-EAGRb$_{124-207}$ measured at a protein concentration of 2 mM in 0.02 M potassium phosphate pH 6.5, 0.09 M NaCl and 10 % (v/v) D_2O.

Superposition of MG200-DnaJ$_{1-74}$, MG200-DnaJ/EAGRb$_{1-207}$ and MG200-EAGRb$_{124-207}$ two-dimensional NMR spectra (Figure III.14A) reveals that there is a very good match between the peaks found on the MG200-DnaJ$_{1-74}$ and MG200-EAGRbox$_{124-207}$ spectra with those on MG200-DnaJ/EAGRbox$_{1-207}$. This indicates that both DnaJ and EAGR box domains maintain the same fold when expressed alone or when being part of

a larger polypeptide. Furthermore, this also excludes the possibility of the existence of meaningful interactions between the two domains in the MG200-DnaJ/EAGRbox$_{1-207}$ domain variant. Superposition of two-dimensional NMR spectra of MG200-DnaJ$_{1-74}$ in absence and presence of MG200-EAGRbox$_{75-207}$, not labelled with ^{15}N, show that there is no difference on the peak fingerprint on both measurements (Figure III.14B), confirming that the two domains do not interact.

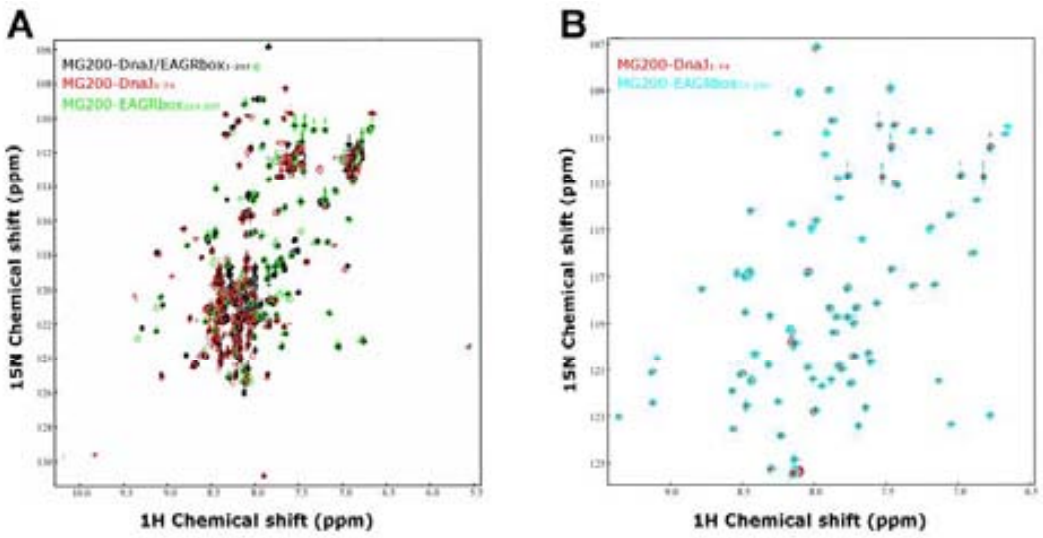


Figure III.14. Comparison of two-dimensional NMR spectra of several of the MG200-N-terminal domain variants acquired at 800 MHz and 25 ºC. A) Superposition of [^{1}H, ^{15}N]-HSQC spectra of MG200-DnaJ/EAGRb$_{1-207}$ (black), MG200-DnaJ$_{1-74}$ (red) and MG200-EAGRb$_{124-207}$ (green) at protein concentrations of 0.14, 0.24 and 2 mM, respectively. B) Superposition of [^{1}H, ^{15}N]-HSQC spectra MG200-DnaJ$_{1-74}$ at a concentration of 0.12 mM in absence (red) and presence (cyan) of 0.36 mM of MG200-EAGRb$_{75-207}$. Samples were in 0.02 M potassium phosphate pH 6.5, 0.09 M NaCl and 10 % (v/v) D_2O.

2.6. MG200-EAGRb$_{124-207}$ crystal structure

2.6.1. Crystallization and preliminary X-ray diffraction analysis

Initial crystallization screening of MG200-EAGRb$_{124-207}$ resulted in a very high percentage of transparent drops. However, the sample crystallized with a very thin needle-shape when taken to a concentration of 30 mg/mL (Figure III.15A). These crystals were optimized in order to obtain quality-diffracting single crystals (Figure III.15B) and the best crystallization condition determined is 0.1 M sodium citrate pH 4.5 and 22 % (w/v) PEG 2000. Crystals were obtained only in the pH range 3.5-4.5 likely because this fragment is abnormally enriched in charged residues, mainly aspartics and glutamics (see domain physico-chemical properties on Table III.1).

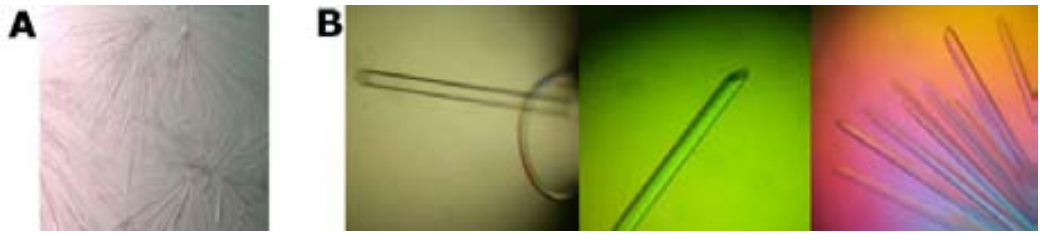


Figure III.15. MG200-EAGRb$_{124-207}$ crystals. Typical crystals grown in (A) nanoliter-scale drops and (B) microliter-scale drops (crystals approximate dimensions: 0.05 × 0.05 × 0.7 mm).

MG200-EAGRb$_{124-207}$ crystals, grown to their maximum size in one week, were frozen in liquid-nitrogen in presence of 20 % (v/v) glycerol and diffracted at the ESRF ID14eh4 beam line. The best diffracting MG200-EAGRb$_{124-207}$ native crystals, up to 2.9 Å resolution, belonged to the $P3_121$ trigonal space group with unit cell lengths of a=b=81.0 Å and c=73.3 Å. The crystal contained two subunits per asymmetric, which gives a solvent content volume of 62 %.

2.6.2. Heavy-atom derivative crystals

Phasing methods for structure solution require a heavy-atom to be present in enough sites in the molecule and at a high enough occupancy to give a clear anomalous signal. A series of heavy-atom compounds were selected for crystal derivatization by the soaking process. After soaking, diffraction data were measured at the ESRF beam lines ID14eh4, ID23eh2 or ID29 (Table III.3). From the extensive search for MG200-EAGRb$_{124-207}$ heavy-atom derivative crystals no derivative was found. At this stage, a series of EAGR box methionine single mutants were prepared.

Table III.3. Summary of the heavy-atom derivative crystals data collection.

Heavy-atom	Chemical Formula	Soaking Conditions	Resolution (Å)	Completeness (%)	R_{merge} (%)	Derivative
Osmium	K_2OsCl_6‡	10 / 23 h	5.0	-	-	X
	K_2OsCl_6‡	5 / 48 h	4.4	93.5 (85.2)	9.3 (24.5)	X
Platinum	K_2PtCl_4‡	10 / 30 h	4.0	100.0 (100.0)	14.2 (34.9)	X
	K_2PtCl_4‡	5 / 24 h	4.5	77.3 (74.1)	6.8 (11)	X
Mercury	Thimerosal‡	2 / 30 h	3.7	100.0 (100.0)	16.2(39.9)	X
	$Hg(CH_3COO)_2$‡	0.5 / 30 h	3.6	100.0 (100.0)	14.4 (44.2)	X
	EtHgChloride§	10 / 4 h	3.5	99.9 (99.2)	10.6 (29.4)	X
Leader	$Pb(NO_3)_2$‡	2 / 20 h	3.7	-	-	X
Gold	$K_2(CN)_2Au$†	2 / 8 h	3.4	99.9 (100.0)	11.6 (44.0)	X
	$K_2(CN)_2Au$§	5 / 2 h	3.5	98.1 (97.0)	11.6 (47.0)	X
Ytterbium	$Yb(NO_3)_3$†	4 / 6 h	3.5	99.7 (98.0)	10.0 (33.0)	X
	$Yb(NO_3)_3$†	2 / 48 h	3.5	99.7(97.8)	14.1 (43.6)	X
	$Yb(NO_3)_3$‡	5 / 48 h	4.3	95.0 (83.3)	9.9 (16.4)	X
Lutetium	$Lu(CH_3COO)_3$†	20 / 8 h	3.3	100.0 (99.9)	10.5 (39.4)	X
	$Lu(CH_3COO)_3$†	4 / 24 h	3.3	100.0 (100.0)	11.3 (49.2)	X
	$Lu(CH_3COO)_3$†	2 / 50 h	3.4	100.0 (100.0)	11.6 (47.0)	X
	$Lu(CH_3COO)_3$‡	4 / 30 h	3.9	98.3 (94.1)	7.4 (44.9)	X
Samarium	$Sm(CH_3COO)_3$‡	5 / 48 h	3.8	97.5 (86.3)	7.3 (19.0)	X
	$Sm(NO_3)_3$§	5 / 24 h	3.2	96.6 (83.6)	9.2 (10.3)	X

Diffraction data collected at the ESRF beam lines § ID23eh2, † ID14eh4 and ‡ ID29.
Soaking conditions: heavy-atom soaking concentration (mM) / soaking time.

2.6.3. Methionine single mutants production and crystallization

When no methionine is present in the interest protein sequence, variant proteins can be produced with methionine residues (Leahy, Erickson et al. 1994; Skinner, Zhang et al. 1994) to obtain phases by using SeMet-labelled protein crystals. As this was the case for the MG200-EAGRb$_{124-207}$ domain variant, secondary structure prediction programs were used to identify residues that were likely in well-ordered regions of the protein. According with this analysis a methionine residue was introduced in the EAGR box sequence, as single mutants, in positions corresponding to Ile140, Leu156 and Leu172. Each MG200-EAGRb$_{124-207}$ methionine variant was successfully produced in SeMet-rich medium and purified from bacterial cells following the protocol established for the native protein.

2.6.4. Structure determination and refinement

MG200-EAGRbox$_{124-207}$ crystals were obtained only for the SeMet140 and SeMet156 variants in around the same conditions determined for the native protein. Crystals from SeMet156 (but not from SeMet140) provided a significant diffraction anomalous signal that gave phases up to 3.1 Å resolution by applying the SAD method (Table III.4). The overall phasing figure of merit, as determined by the program SHELXD

(Schneider and Sheldrick 2002), was of 0.65, rather high, presumably on account of the high occupancy of the two Se atoms found. Structure refinement was performed on high resolution data, to 2.9 Å, from a native MG200-EAGRb$_{124-207}$ crystal.

Table III.4. Data collection, phasing and refinement statistics.

	Native form	SeMet derivative
Data Collection		
Wavelength (Å)	0.979	0.979
Space group	$P3_121$	$P3_121$
Unit cell parameters	a = b = 81.0 Å c = 73.3 Å α = β = 90º γ = 120º	a = b = 80.9 Å c = 73.6 Å α=β=90º Γ=120º
Resolution (Å)	30.0-2.9 (3.0-2.9)	30-3.1 (3.2-3.1)
No. of observations	210131	162530
No. of unique reflections	6450	5289
<I>/<σ(I)>	32.7 (5.2)	24.8 (6.9)
Completeness (%)	99.8 (96.5)	100 (99.8)
R_{merge} (%)	6.7 (49.8)	10.3 (44.4)
Phasing Statistics		
<Dano>/<σ(Dano)>		1.18 (0.88)
Dano completeness		98.5 (99.6)
Refinement		
Resolution (Å)	25.00-2.90 (2.97-2.90)	
No. of reflections	6120 (425)	
Asymmetric unit content	Dimer	
R_{factor}	20.3 (33.3)	
R_{free}	26.0 (37.4)	
Average B_{factor} (Å^2)	35.5	
Bonds (Å)	0.019	
Angles (º)	1.7	

Values in parentheses are for the highest resolution shell.

2.6.5. Overall structure

The final MG200-EAGRb$_{124-207}$ structure includes coordinates for residues Gln149 to Glu203 from both subunits in the crystal unit cell. Missing from the final model are residues 124 to 148, at the N-terminal end, and residues 204 to 207 and the His_6-tag residues that followed, at the C-terminal end, for which just poor electron density was observed or completely absent. A relatively large volume of significant, though diffuse, electron density remained uninterpreted in the final model. This density, which does not follow the two-fold molecular symmetry, presumably corresponds to the C-tail of a symmetry-related neighbour molecule (Figure III.16). The absence of model to explain this extra density contributed to the difficulty in obtaining even lower $R_{factors}$.

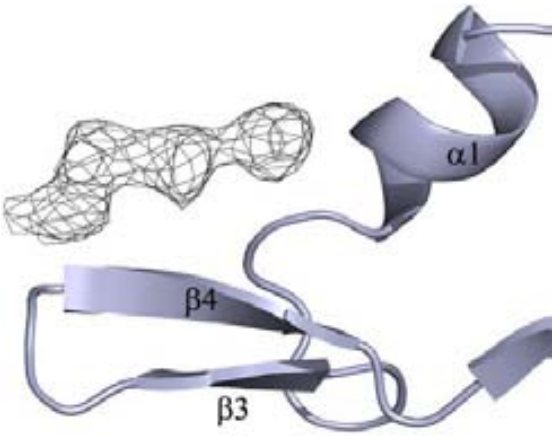


Figure III.16. Unexplained electron density map volume in the vicinity of β3-β4 hairpin and helix α1. Extra electron density found near key secondary structure elements of one of the MG200-EAGRb$_{124-207}$ subunits. Attempts to model it were not satisfactory.

Each MG200-EAGRb$_{124-207}$ subunit within the crystal asymmetric unit folds into a three-stranded antiparallel β-sheet (β1-β6-β5) flanked by a small α-helix (α1). Strands β5 and β6 are connected by a short β-turn forming a long hairpin, which is preceded by another protruding β-hairpin formed by strands β3 and β4 (Figure III.17). The N- and C-ends of the domain are both pointing away from the compact core of the subunit.

The most characteristic feature of the EAGR box sequences is the high content in aromatic residues, from which they received their name. In the structure, these residues form a very tight hydrophobic cluster composed by residues: Leu156, Val164, Trp169, Leu172, Tyr178, Trp187, Trp189, Tyr192, Phe193, Trp199 and Trp 201 (Figure III.18), which locates between the β-sheet and the flanking α-helix α1.

At a structural level, the most remarkable feature of the MG200-EAGRb$_{124-207}$ is the presence of an intra-domain symmetry axis which relates the two hairpins β3-β4 (Gly179-Trp189) and β5-β6 (Gly191-Trp201). Given their particular organization within the structure they will be referred as the wings of the subunit (wing 1 and wing 2 for the β3-β4 and β5-β6 hairpins, respectively) from now on.

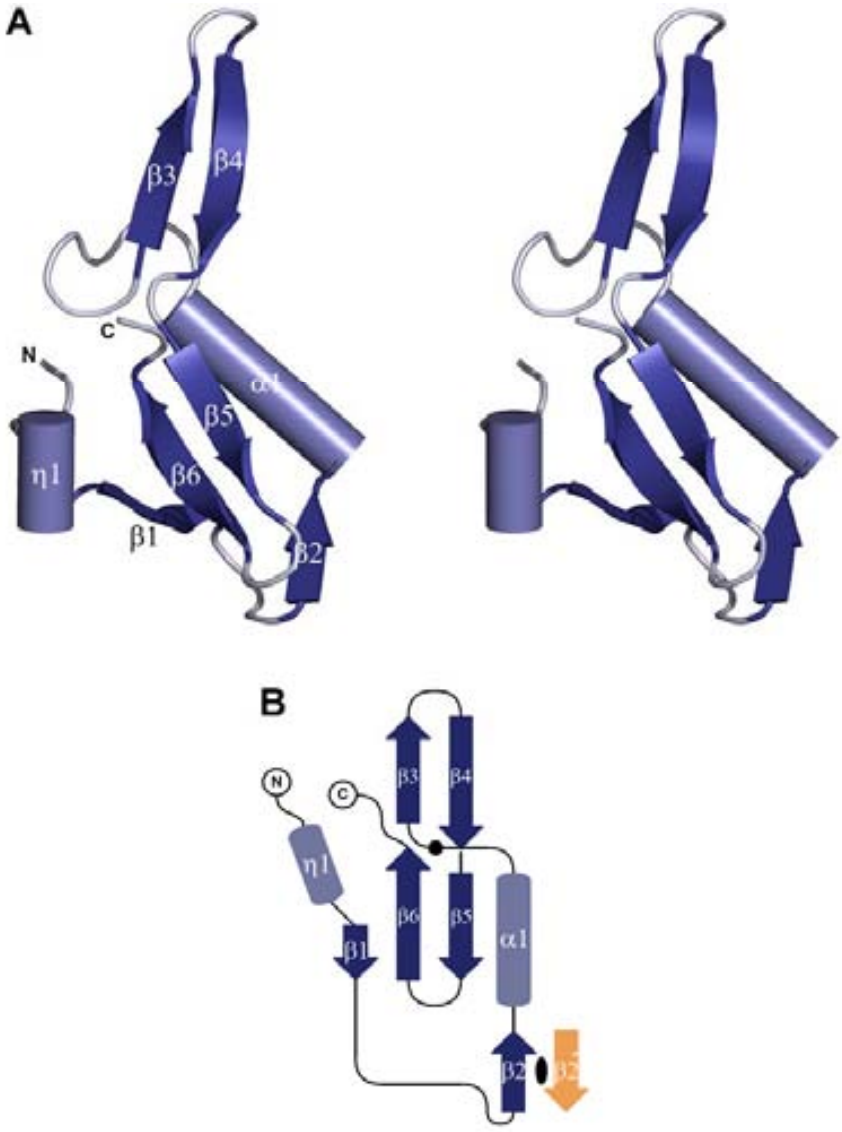


Figure III.17. MG200-EAGRb$_{124-207}$ overall struture. A) Stereogram of the MG200-EAGRb$_{124-207}$ structure (cartoon representation). The view is down the intra-domain symmetry axis, which relates the two hairpins β3-β4 and β5-β6 (wings 1 and 2, respectively). B) Topology diagram of MG200-EAGRb$_{124-207}$ with helices depicted as light blue cylinders and β-strands as dark blue arrows. The black circle indicates the intra-domain symmetry axis that relates the two wings. The black oblong dot indicates the symmetry axis that relates the two EAGR box subunits found in the crystal asymmetric unit. In orange the β2 strand from the symmetry related subunit that is labelled with a star.

The intra-domain symmetry found within each subunit of the crystal asymmetric unit has different rotation angles, 135.8º and 145.1º. The superposition of all the main chain atoms of the eleven equivalent residues of the wings gives rmsd of 0.27 and 0.62 Å for the two subunits. The superposition of the main chains results also in the

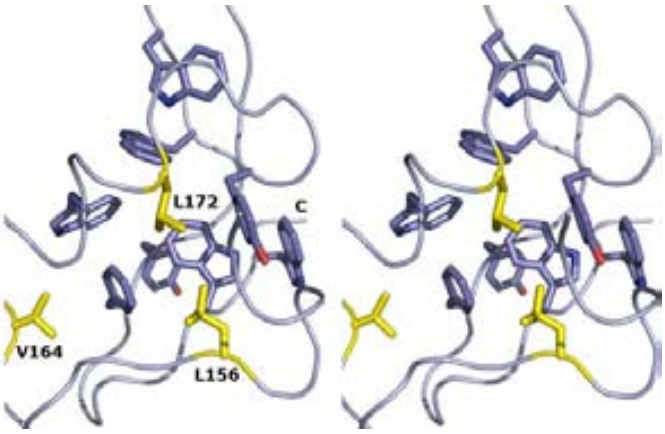


Figure III.18. Hydrophobic cluster within an MG200-EAGRb$_{124-207}$ subunit (stereo view). The eleven hydrophobic residues forming the tight hydrophobic cluster found at the core of each MG200-EAGRb$_{124-207}$ subunit are represented in stick mode and with carbon atoms coloured in blue and yellow for the aromatic and aliphatic side chains, respectively.

accurate superposition of the side chains conformations, in particular for the four conserved aromatic residues found in each wing (Figure III.19). The two wings appear to correspond to a sequence duplication that is well conserved among the EAGR box sequences available (Figure III.20).

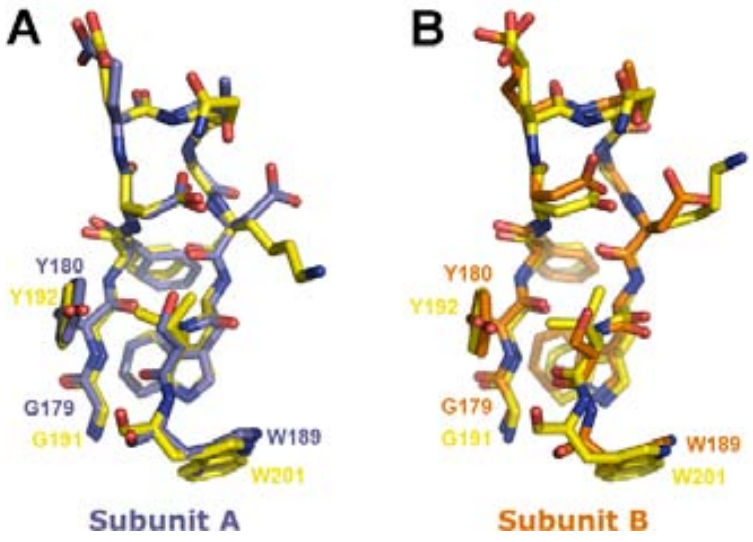


Figure III.19. Structural comparison between the wing structures found within each MG200-EAGRb$_{124-207}$ subunit. A) Superposition of the wing structures in the MG200-EAGRb$_{124-207}$ subunit A with rmsd of 0.27 Å and carbon atoms coloured in blue and yellow for residues Gly179 to Trp189 (wing 1) and Gly191 to Trp201 (wing 2), respectively. B) Superposition of the wing structures in the MG200-EAGRb$_{14-207}$ subunit B with rmsd of 0.63 Å and carbon atoms coloured in orange and yellow for residues Gly179 to Trp189 (wing 1) and Gly191 to Trp201 (wing 2), respectively.

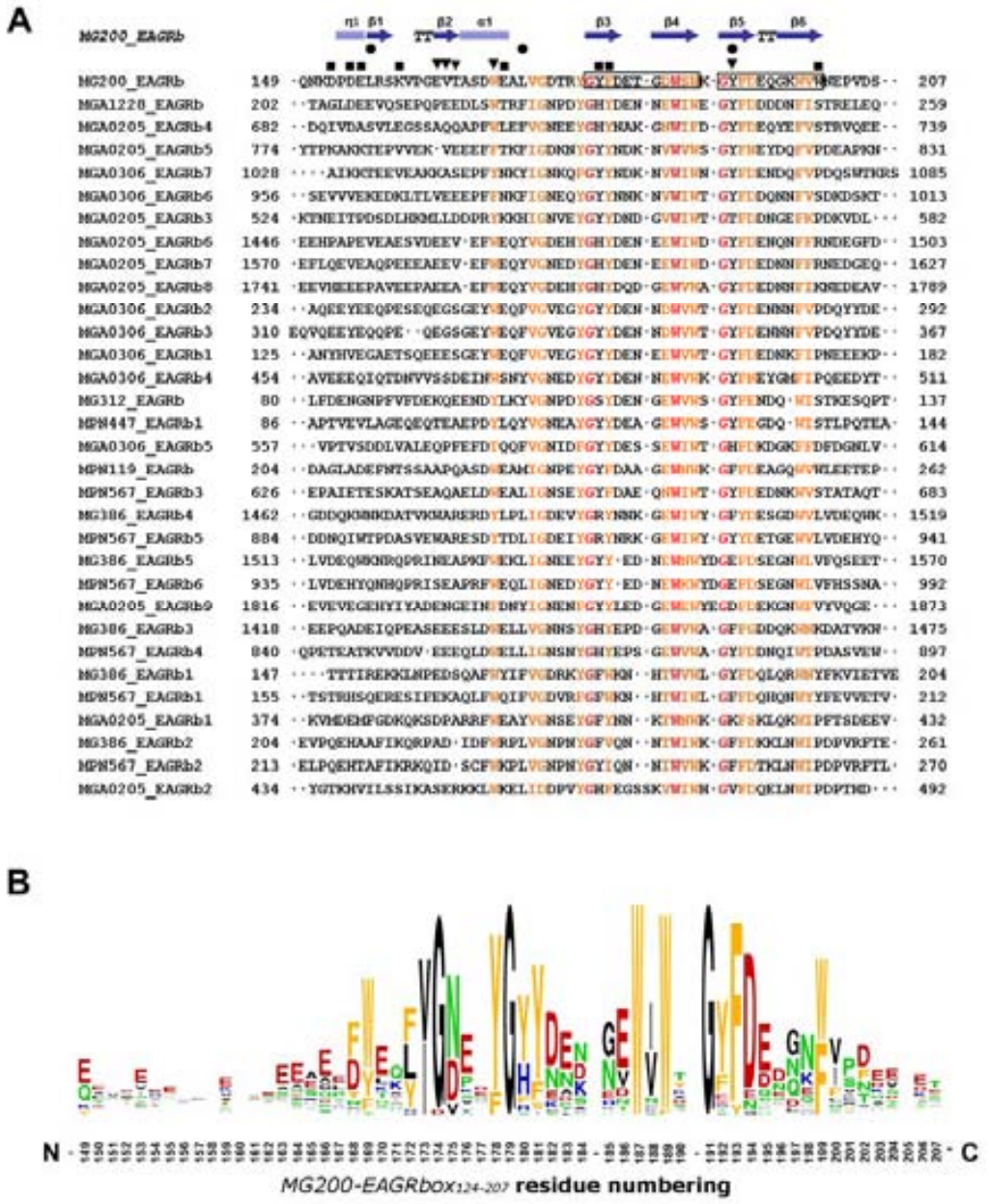


Figure III.20. Comparison of primary sequence relationship between all the EAGR box sequences currently available in the literature. A) EAGR boxes multiple-sequence analysis using the ClustalW program (Larkin, Blackshields et al. 2007). The aligned EAGR boxes are from ORFs MG200, MG312 and MG386 from *M. genitalium*, ORFs MPN119, MPN447 and MPN567 from *M. pneumoniae* and ORFs MGA0205, MGA0306 and MGA1228 from *M. gallisepticum*. The number that appears just after the gene number indicates the location of the EAGR box within the protein sequence. Residue numbering refers to the complete protein sequences as derived from the corresponding UniProt entries. Fully and strongly conserved residues are colored in red and orange, respectively. Sequence repeats are boxed in MG200. Squares and triangles are for residues involved in crystal contacts and the dimer interface, respectively. Circles indicate residues that have been mutated in the MG200-EAGRb$_{124\text{-}207}$ in this work. B) MG200-EAGRb$_{124\text{-}207}$ sequence logo (performed at http://weblogo.berkeley.edu). The size of the character depicts the relative proportion of the corresponding amino acid at a given site, the larger the logo the lesser the variability. The high conservation of glycine and aromatic residues, in agreement with the domain name, is apparent mostly in the central part of the box. The abundance of acidic amino acids, in particular glutamics, in many EAGR boxes is also evident.

2.6.6. Domain methionine variants analysis

The MG200-EAGRb$_{124-207}$ structure analysis allows rationalizing the different behaviour of the three produced methionine mutants I140M, L156M and L172M. MG200-EAGRb$_{124-207}$ with a methionine residue in position 172 did not crystallized most likely because, being central to the hydrophobic cluster, might significantly alter the structure of the domain. The SeMet140 variant crystallized but crystals did not provide consistent anomalous signal as position 140 corresponds to a disordered region in the structure, outside the globular region of MG200-EAGRb$_{124-207}$ that was part of the construct used to express the domain. In the case of the MG200-EAGRb$_{124-207}$ SeMet156 variant the mutation was in a region located at the entrance of β1 and on the edge of the hydrophobic cluster, which seems to have enough plasticity to allow the residue replacement without requiring major changes in the structure.

2.6.7. Dimer organization

The MG200-EAGRb$_{124-207}$ crystal structure reveals a compact homo-dimer. The two subunits found in the crystal asymmetric unit are related by a non-crystallographic quasi-binary axis of 177.6 º (Figure III.21A-B). The two subunits superpose well with rmsd of 0.65 Å for all the C^{α} atoms of the two subunits (Figure III.21C). The low rmsd value obtained from the overall superposition of the two subunits confirms the constancy of the determined structure and emphasizes the fact that the largest deviation between the two subunits corresponds to wing 1 and is mainly due to the different values of the intra-domain rotations in the two subunits.

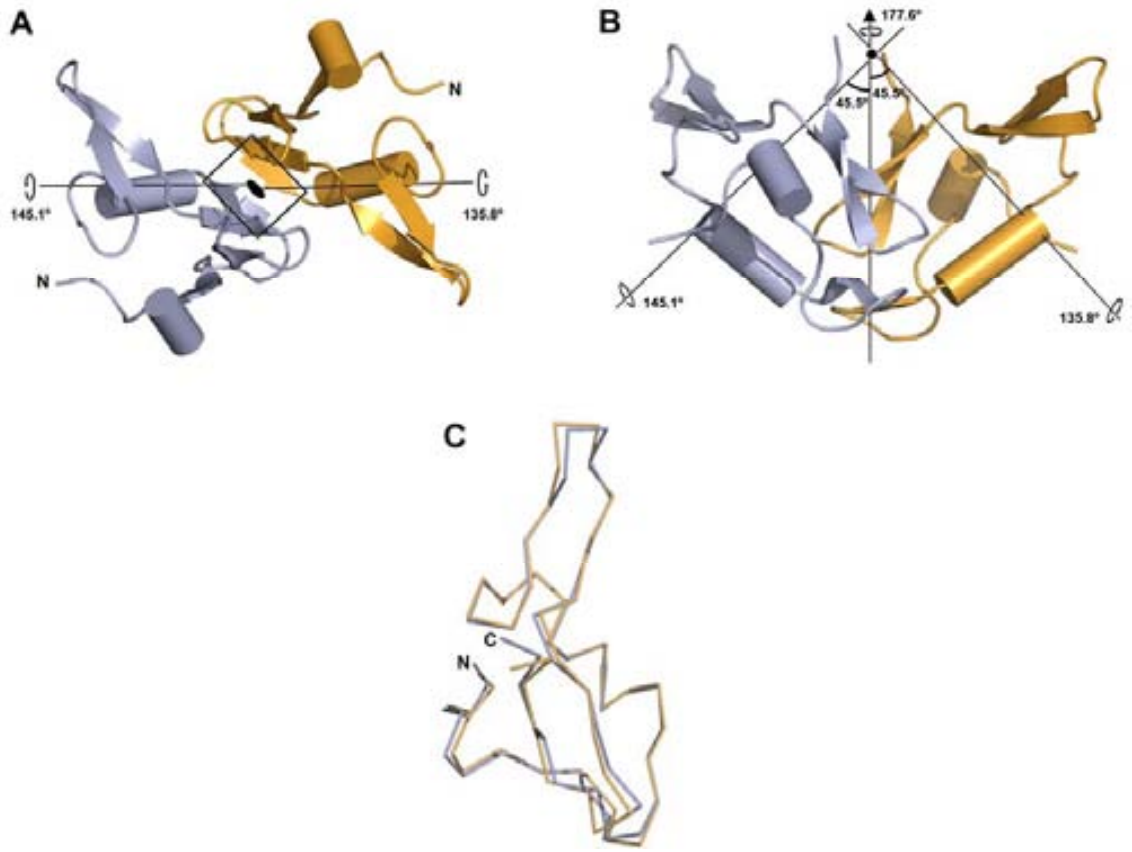


Figure III.21. MG200-EAGRb$_{124-207}$ dimer organization. Overall views of the MG200-EAGRb$_{124-207}$ dimer (A) down and (B) perpendicular to the inter-subunits symmetry axis, which is represented as an oblong dot and an arrow, respectively. The inter-subunits axis is coplanar with the two intra-domain axes, which are depicted as continuous lines. C) Superposition of the two MG200-EAGRb$_{124-207}$ subunits found in the crystal asymmetric unit with rmsd of 0.65 Å.

The dimer interface is mainly formed by two levels of interactions between subunits along the inter-subunits symmetry axis (Figure III.22):

i) the N-end or lower level, involves the main chain from strand β2 (Gly162 to Ala166), which form a five-hydrogen bonds two-stranded antiparallel β-sheet with the corresponding β2 strand in the second subunit (Figure III.22B);

ii) the C-end or upper level, involves the side chain interactions from helix α1 (Trp169-Glu170) and from strand β5 (Tyr192-Glu195) (Figure III.22C).

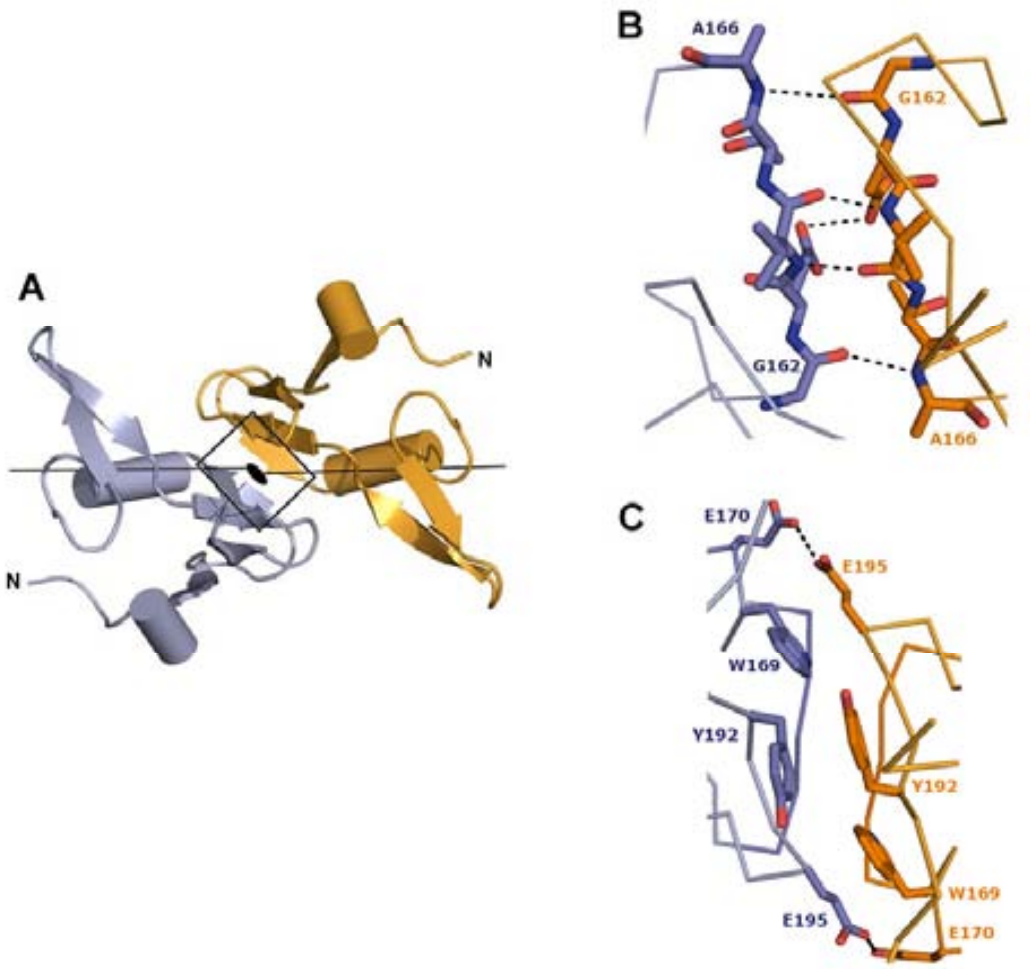


Figure III.22. Overall arrangement of the MG200-EAGRb$_{124-207}$ dimer and details of the inter-subunits interface. A) Dimeric organization of MG200-EAGRb$_{124-207}$ in the crystal down the inter-subunits symmetry axis, which is represented as an oblong dot. Interactions at the interface can be described as organized at two levels: B) N-end level, involving mainly the main chain from strand β2 (Gly162 to Ala166); C) C-end level, involving the side chains from Trp169 and Glu170 from helix α1 and also Tyr192 and Glu195 from strand β5. The β2 strands from the two subunits form a two-stranded antiparallel β-sheet with five hydrogen bonds. Both (B) and (C) are close-up views of the dimer interface down the inter-subunit symmetry axis.

Interactions bridging the Oγ atoms, from both the side chains of Glu170 with Glu195 and of Glu163 from the two subunits, have likely been enhanced by the low pH of the crystallization condition.

The dimer surface area buries 1070 Å^2 of the solvent accessible surface from each MG200-EAGRb$_{124-207}$ subunit. This area is mainly constituted by hydrophobic interactions between the aliphatic moieties of residues Trp169, Tyr192 and Glu195, which extend the intra-domain hydrophobic cluster in each subunit throughout the dimer interface (Figure III.18 and III.23).

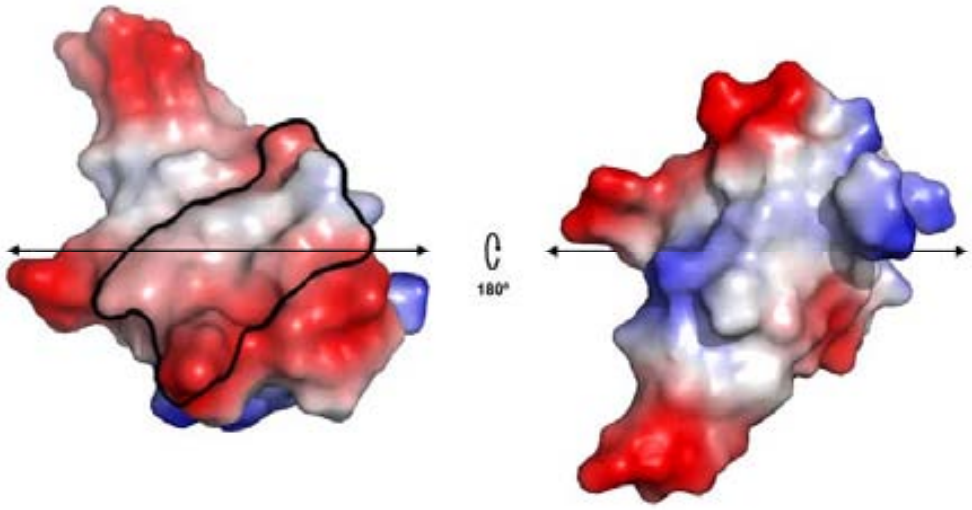
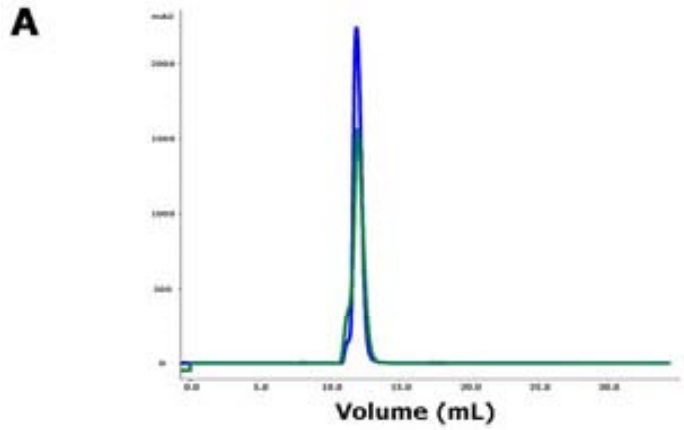


Figure III.23. Electrostatic surface potential of a MG200-EAGRb$_{124-207}$ subunit. Two views were represented to give a complete electrostatic surface representation of one MG200-EAGRb$_{124-207}$. The inter-subunits symmetry axis is shown as the horizontal arrow and the area buried at the dimer interface, with a clear hydrophobic character, is also indicated.

Besides existing as a dimer in the crystal asymmetric unit MG200-EAGRb$_{124-207}$ was also demonstrated to behave as a dimer in solution by GFC and DLS assays (Figure III.24A-B). Residues at the dimer interface, such as Tyr192 (which is positioned along the inter-subunits axis), are poorly conserved (Figure III.20) what seems to be suggesting that not all EAGR boxes dimerize. In order to interfere with the dimer formation a MG200-EAGRb$_{124-207}$ Y192D mutant was designed. However, this variant was also found to exist as a dimer in solution (Figure III.24A-B).



B

DLS parameters

MG200-EAGRb$_{124-207}$	RH (nm)	PDI (%)	MW (kDa)	Intensity (%)	Mass (%)
Native	4.1	4.9	20	98	100
Y192D	4.0	5.2	23	95	100

Figure III.24. Analysis of the oligomeric state of MG200-EAGRb$_{124-207}$, native and mutant (Y192D), in solution. A) GFC fractionation of the MG200-EAGRb$_{124-207}$ native (blue) and mutant (Y192D, green) proteins in a Superdex 75 10/300 column. Both variants eluted at 12.3 mL column volumes (indicating both have a relative molecular weight of ~24 kDa). B) Comparative DLS parameters for the two MG200-EAGRb$_{124-207}$ variants, which have an identical behaviour in solution, existing as dimers. The measurements were carried out at 20 ºC on 10 mg/mL protein samples.

The inter-subunits symmetry axis makes with the two intra-domain symmetry axes an angle of approximately 45.5º. Moreover, the three axes are essentially co-planar (Figure III.21A-B). The particular organization found for the intra- and inter-subunits axes of the MG200-EAGRb$_{124-207}$ structure results in the unbalance of interactions between the two wings. Thus, while wing 1 remains mostly exposed to the solvent, wing 2 is involved in the hydrophobic core as well as in interactions in the dimer. If sequence and structural similarities between both wings is taken into account, it seems that wing 1 should have a high tendency to interact with other macromolecules, especially the solvent exposed aromatic residues Tyr180 and Phe181 (Figure III.25).

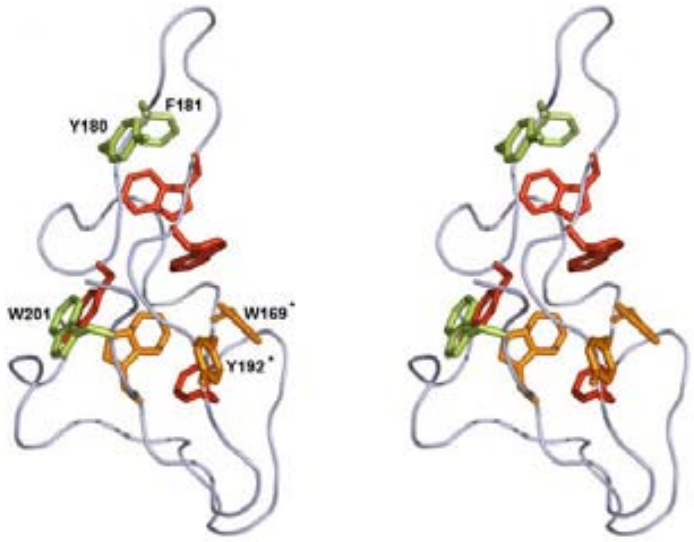


Figure III.25. Stereo view of MG200-EAGRb$_{124-207}$ with all the aromatic residues explicitly depicted in stick mode. In green the aromatic residues (Tyr180, Phe181 and Trp201) that remain exposed to the solvent in the dimer. The buried aromatic residues appear well (in orange above 55%) or very well (in red above 85%) conserved among EAGR boxes. Residues Trp169 and Tyr192 (labelled with a star) become buried only in the dimer (see Figure III.22C).

2.6.8. Crystal packing interactions

The main packing interactions found for the MG200-EAGRb$_{124-207}$ crystal dimer are precisely between wing 1, from both molecules, and symmetry related dimers (Figure III.26A). In particular, Phe181 interacts with Tyr180 from a neighbour dimer and the aliphatic part of the side chain from Glu170 interacts with Phe181 from a different symmetry related molecule. More packing interactions are apparent for wing 1 coming from the presence of a large volume of extra electron density (see section III.2.6.5. and Figure III.26B).

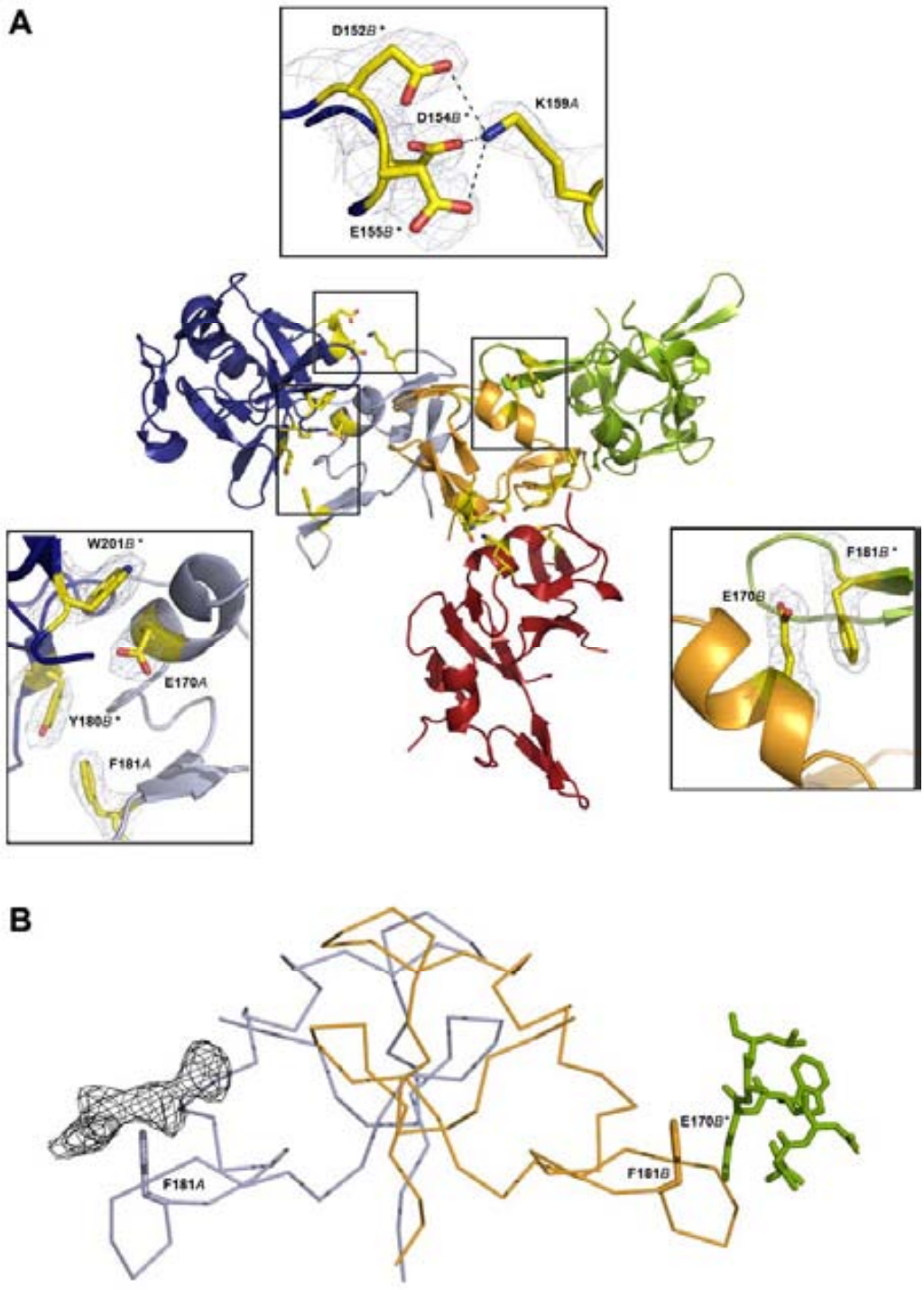


Figure III.26. Crystal packing interactions and wing 1 reactivity. A) MG200-EAGRb$_{124\text{-}207}$ dimer subunits found in the crystal asymmetric unit are coloured in light blue and orange as in Figure III.21. In turn, the three symmetry related interacting dimers are shown in red, dark blue and green. Close-up views of the most significant crystal contacts are shown together with the corresponding (2Fo-Fc) electron density for the residues involved, which are also explicitly shown in sticks mode and atom-coloured representation. Residues from the neighbour dimers are labelled with a star. Solvent exposed aromatic residues, Tyr180 and Phe181 from wing 1, which are expected to be reactive, participate in the dominant crystal interactions. B) Unexplained extra electron density, found near wing 1 of one of the subunits, likely corresponds to the C-terminal His$_6$-tag of a neighbour molecule in the crystal. For the second subunit the same protein region presents crystal packing interactions. These observations would be in agreement with a high tendency, for wing 1, to participate in inter-molecular interactions.

The reactive nature of wing 1 is now clear given the above described evidences but, despite it, the strong acidic character of wing 1 (composed among others by Asp182, Glu183, and Asp186; Figure III.27) is expected to interfere with the formation of extensive wing 1-wing 1 interactions. On the contrary, on wing 2 a lysine residue (Lys198) instead of a glutamic acid (Asp186 in wing 1) might facilitate dimerization.

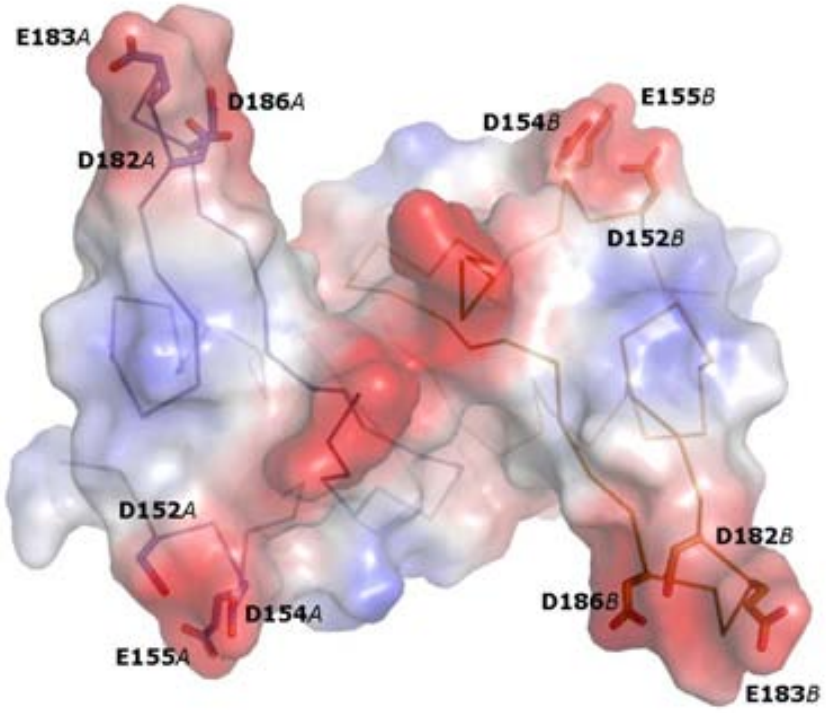


Figure III.27. MG200-EAGRb$_{124-207}$ dimer electrostatic surface potential. View down the inter-subunits symmetry axis. Subunits are coloured in light blue and orange as in Figure III.21. The molecular surface is coloured according to the local electrostatic potential as calculated with the program PYMOL (DeLano 2002). Each subunit shows at least two main patches of negatively charged residues: the first is mainly composed by the acidic residues Asp152, Asp154 and Glu155, which are key residues of crystal packing interactions; and the second, composed by Asp182, Glu183 and Asp186, confirms the strong acidic character of wing 1.

Besides the contacts between the wings, which are contributing to the crystal packing, there is at least one salt bridge, between Lys159 and the acidic residues Asp152, Asp154, and Glu155 (Figure III.26 and III.27), that is also forming part of key crystal contacts.

2.6.9. Primary sequence conservation among EAGR boxes

EAGR boxes multiple-sequence alignment, complemented with the structural information from the MG200-EAGRb$_{124-207}$ (Figure III.20), shows that there is a high degree of conservation after helix α1, in particular for glycine and aromatic (mainly tryptophane) residues from the wings (strands β3, β4 and β5, β6; Figure III.20). Surprisingly, only residues Gly179, Trp187, and Gly191 are fully conserved among all EAGR box sequences. Glycine residues 179 and 191, respectively found at the entrance of strands β3 and β5, are accurately related by the intra-domain symmetry and present main chain conformational angles favourable only for glycine (with phi and psi values of about -175° and -175°) meaning that the replacement of either Gly179 or/and Gly191 by any other amino acid would introduce steric clashes with side chains from either Trp187 or/and Trp199, respectively (Figure III.28). Trp187 itself is a fully conserved residue and Trp199 is conservatively replaced by a phenylalanine (in about 50% of the cases, Figure III.20). Their configurations are both stabilized by the stacking of the aromatic rings with the planar peptide bonds of the two strands of the corresponding hairpin and also by the face to edge interactions with tryptophanes 189 and 201, respectively. The aromatic side chains of the four tryptophanes (Trp187 with Trp189 and Trp199 with Trp201) form a caging effect that seems to allow glycines 179 and 191 to act as the hinges of the rigid body movements observed for the wings.

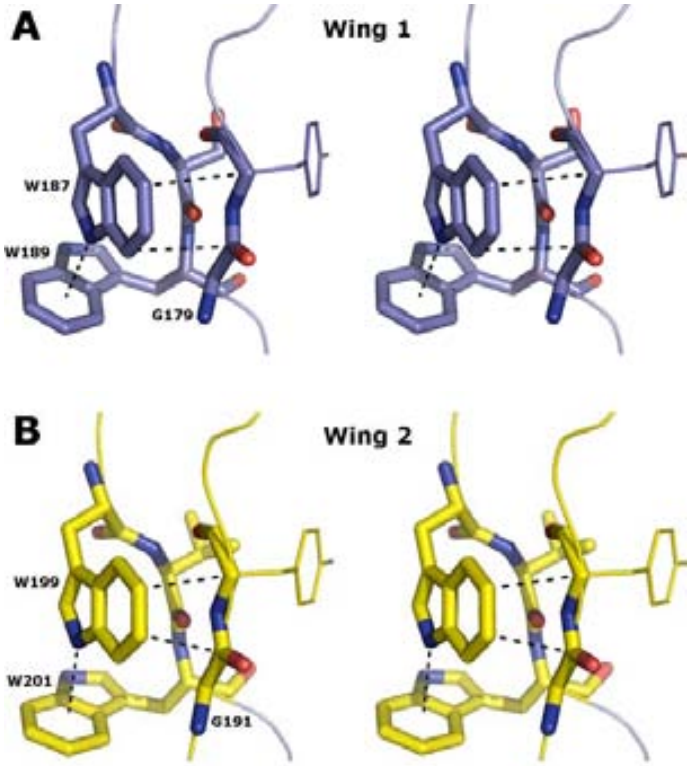


Figure III.28. Environment of the two fully conserved glycine residues in EAGR boxes. Caging defined around (A) Gly179, at the entrance of strand β3 (wing 1), and (B) Gly191, at the entrance of strand β5 (wing 2), accurately related by the intra-domain symmetry. The caging defined by the side chains of the tryptophanes allows the glycines to act as hinges for the rigid body movements of the wings.

A phylogenetic tree was built with the program MEGA4 (Tamura, Dudley et al. 2007) for the EAGR box sequences available. From the tree retrieved by the program two major trends are apparent (Figure III.20 and III.29):

i) EAGR boxes from orthologue proteins present the highest degree of similarity indicating a relatively recent common origin;

ii) EAGR boxes from multi-EAGRbox-containing proteins present, in general, a high degree of similarity between the different boxes found in the protein, suggesting the occurrence of internal duplication events.

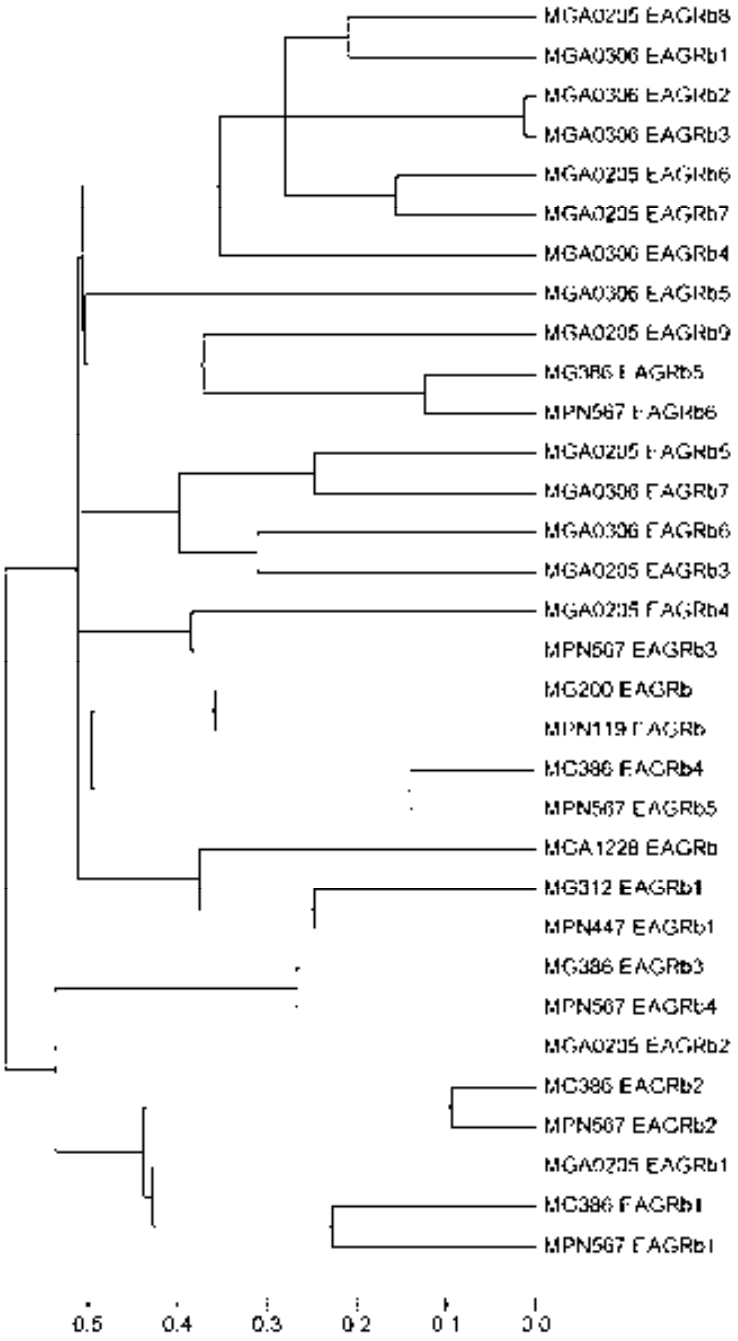


Figure III.29. Phylogenetic tree of EAGR boxes. Phylogenetic analysis was carried out by the Neighbour-Joining method using the program MEGA4. EAGR boxes from orthologue proteins of different mycoplasmas present a high degree of similarity, reflecting a recent common origin. Multiple EAGR boxes within a given protein present higher similarity, suggesting internal duplication events.

2.6.10. Structural relationship of MG200-EAGRb$_{124-207}$ with RegA

The Dali server (Holm, Kaariainen et al. 2008) was used to search for MG200-EAGRb$_{124-207}$ structural homologues. The program retrieved only a single hit, with the low Z-score of 0.7, for the translational regulator protein RegA from bacteriophage T4 (PDB accession code 1REG). The RegA protein is a 122 residues polypeptide known to bind to a messenger RNA region that includes the initiator codon (Kang, Chan et al. 1995).

MG200-EAGRb$_{124-207}$ and RegA present a similar fold maintaining the following secondary structure elements: α1, β1 and hairpins β3-β4 and β5-β6 (which correspond to wings 1 and 2, respectively). However, despite the topological similarity only eight residues are fully conserved between the two proteins, one of them corresponding to Gly179 in MG200-EAGRb$_{124-207}$, and surprisingly, none of them corresponding to the aromatic residues characteristic of the EAGR boxes (Figure III.30A).

Superposition of the MG200-EAGRb$_{124-207}$ and RegA structures gives a relatively high rmsd of 3.1 Å for just 44 equivalent residues (Figure III.30B). In RegA the wing structures can still be related with each other by a 145º rotation although the similarity between them is low (they even have different lengths). Besides this low symmetry the domain misses the proper wing hinges, given that a cysteine is found in the position corresponding to wing 2 Gly191 in MG200-EAGRb$_{124-207}$, and there is no caging for the only conserved glycine, suggesting that the wing structures in RegA are not mobile.

However, having a closer look into the RegA region that is equivalent to the reactive wing 1 from MG200-EAGRb$_{124-207}$, it is apparent that it also remains exposed to the solvent in RegA and that there are even experimental evidences indicating that it could participate in the inter-molecular interactions proposed for this protein.

A possible evolutionary connection between the MG200-EAGRb$_{124-207}$ and RegA structures is hard to establish because, despite the topological similarities, very low sequence and structural relationships were found between them.

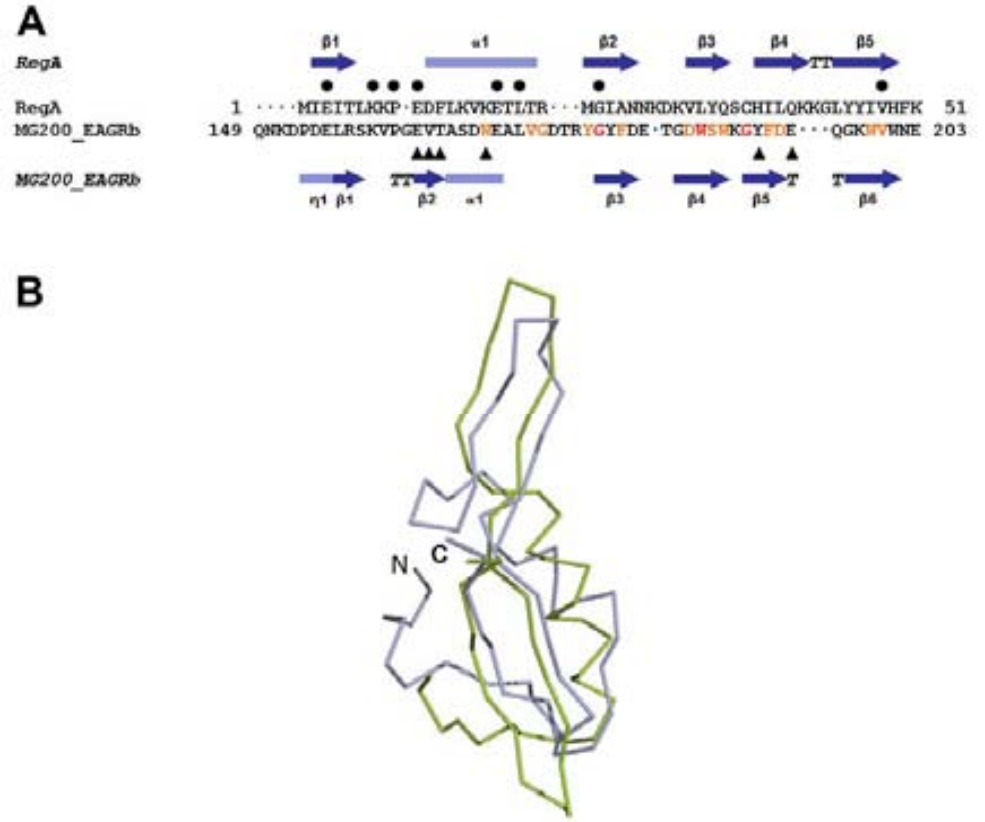


Figure III.30. Structural comparison between MG200-EAGRb$_{124-207}$ and its remote homologue protein RegA. A) Structure-based sequence alignment of MG200-EAGRb$_{124-207}$ and RegA (PDB accession code 1REG), its only structural homologue known. The secondary structure assignments for RegA and for the MG200-EAGRb$_{124-207}$ are depicted at the bottom and at the top of the alignment, respectively. Black circles mark residues conserved in both structures and triangles are for residues in the MG200-EAGRb$_{124-207}$ dimer interface. Surprisingly, none of the aromatic residues conserved among EAGR boxes is present in RegA. In fact, only the conserved EAGR box residue Gly179 is also present in RegA. B) Overall superposition of the MG200-EAGRb$_{124-207}$ (blue) and RegA (green) structures.

2.6.11. EAGR boxes possible functional roles

M. genitalium MG200 protein orthologue from *M. pneumoniae*, TopJ protein or MPN119 (Cloward and Krause 2009), do not have any known partner as was recently suggested in a recent work where the *M. pneumoniae* proteome organization was unravelled (Kuhner, van Noort et al. 2009). However, as suggested by the description of the MG200-EAGRb$_{124-207}$ structure, this domain appears to be a platform for interactions with other macromolecules where the intra- and inter-subunits symmetries combines to achieve large complexes with (quasi)-symmetrical interactions.

Examination, by Transmission Electron Microscopy (TEM), of cells bearing either a deletion of the EAGR box, in the MG312 protein (Burgos, Pich et al. 2007), or deletions

of larger regions including the EAGR boxes, in MG200 and MG386 terminal organelle proteins (Pich, O. Q.; Burgos, R.; Piñol, J.; unpublished data), revealed that there is no apparent structural defects in the terminal organelle organization of these mutants (Figure III.31). Interestingly, examining the motility of the same mutant cells determined that these cells present severe mobile handicap. Therefore, though the EAGR boxes do not seem to be directly involved in the organization of the terminal organelle they play a critical role in the motility mechanism. Nowadays there are two accepted models for motility in mycoplasmas and they are both based in the conformational flexibility of the terminal organelle electron-dense core, which is capable of pushing the cells forward (Henderson and Jensen 2006; Seybert, Herrmann et al. 2006).

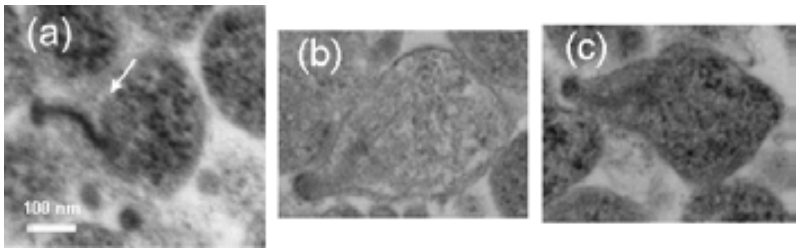


Figure III.31. TEM of *M. genitalium* cells. A) *M. genitalium* wild-type strain, (B) *mg200* and (C) *mg386* transposon mutants that contain genes disrupted just before or within the EAGR box sequence. The white arrow indicates the terminal organelle. There are no apparent changes in the terminal organelle ultrastructure organization of the mutants in comparison with the wild-type cells. All images are shown at the same magnification.

In this scenario the observed variability between the EAGR box wings together with their oligomerization potential took to the proposal of two oligomerization models (Figure III.32) from where it is apparent that EAGR boxes seem especially well suited to provide the plasticity needed in the elements of the terminal organelle to function as part of a dynamic machinery. Moreover, given that the rotation angle between the wings is not perfect the formation of polymers disposed like a spiral can be hypothesized that would take to the formation of limited size ultrastructures has those found in the terminal organelle cytoskeleton.

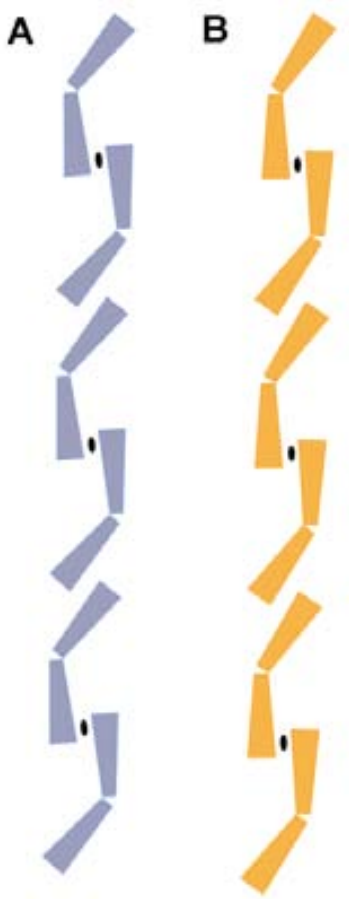


Figure III.32. EAGR boxes oligomerization models. In these models the reactive nature of wing 1 is taken into account. Oligomerization models where the intra-domain rotations are of (A) 135.8º and (B) 145.1º. In the specific case of the MG200-EAGRb$_{124-207}$ the proposed models are ruled out by the negatively charged wing 1 but open new possibilities for its interaction with other components of the terminal organelle.

2.7. Possible interactions between MG200 protein domains

To investigate whether interactions between the several MG200 protein domains are plausible, several pairs of domains were analysed by GFC, DLS or NMR spectroscopy. A summary of these experiments are provided on Table III.5.

The obtained results indicate that no interactions are expected between the MG200 protein domains. However, several of these domains can form homo-oligomers. To further validate these results some of these domains were cross-linked with gluteraldehyde.

Table III.5. Matrix summarizing the results for the interactions between MG200 domain pairs.

MG200 * domains	**$DnaJ_{1-74}$**	**DnaJ/ EAGRb 1-207**	**EAGRb 124-207**	**$Ct_{304-601}$**	***fl*MG200**
$DnaJ_{1-74}$	Dimer (GFC / DLS)	ND	No interaction (NMR)	ND	ND
DnaJ/ EAGRb 1-207		Dimer (GFC / DLS)	ND	No interaction (GFC)	ND
EAGRb 124-207			Dimer (GFC / DLS / Crystal)	ND	No interaction (GFC)
$Ct_{304-601}$				Tetramer (GFC)	ND
***fl*MG200**					Tetramer (GFC)

* The matrix diagonal corresponds to the quaternary association of each protein domain in solution. The techniques used for the determination of protein-protein interactions are indicated between parentheses. N.D.- Not-determined

Cross-linking assays proved MG200-DnaJ-containing domains have propensity to form oligomers, at least homo-dimers (Figure III.33A-B) and also that MG200-$Ct_{304-601}$ can form oligomers but have a higher tendency to precipitate (see lane 4 and 5 on Figure III.33C).

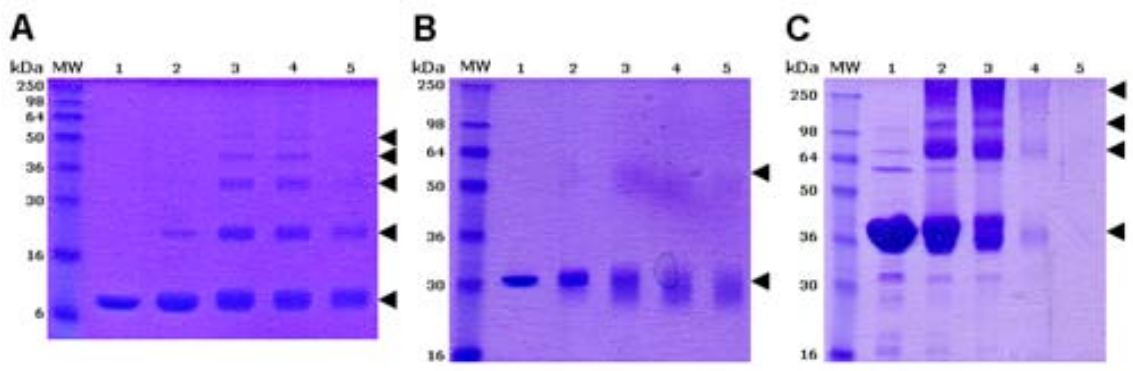


Figure III.33. Chemically induced cross-linking of soluble MG200 protein domains mediated by 25 % (v/v) gluteraldehyde and analysed by SDS-PAGE. A) CBB-stained 15 % (v/v) SDS-PAGE of MG200-$DnaJ_{1-74}$. CBB-stained 12 % (v/v) SDS-PAGE of (B) MG200-DnaJ/$EAGRb_{1-207}$ and (C) MG200-$Ct_{304-601}$. The samples were equilibrated against the cross-linking agent for 0 min (lane 1), 5 min (lane 2), 15 min (lane 3), 30 min (lane 4) and 60 min (lane 5). The molecular weight (MW) of protein standards are indicated as well as the positions of the oligomeric states of each domain (marked with black arrowheads).

2.8. MG200 protein oligomerization

The biophysical analysis of each of the MG200 protein globular domains together with the previous results obtained for the full-length protein allow to propose a model for the protein quaternary structure organization (Figure III.34A). The molecule would be a tetramer with a 222 (D2) symmetry, with the C-terminal domain forming the core of the tetramer, as suggested by the GFC results obtained for the MG200-$Ct_{304-601}$ domain, and both DnaJ domain and EAGR box disposed along one of the inter-subunits two-fold axis, with the amino acidic sequences between them existing as linkers or unstructured regions.

Considering the capability of the MG200-$DnaJ_{1-74}$ domain to form homo-oligomers (see Figure III.33A) and the tendency of the full-length protein to form large aggregates it can be hypothesized that MG200 tetramers can stick together by their N-extremes being able to form fibre-like structures, as been also suggested by preliminary EM electron-negative staining studies of this sample (Figure III.34B). Moreover, having into account all the considerations described in section III.2.6 with respect to the MG200-$EAGRb_{124-207}$ structure, it becomes apparent that this can be a strategic protein domain with sufficient plasticity, given by the predicted wing movements and the combination of the domain intra- and inter-subunits symmetries, to allow interactions with other macromolecules. These interactions, which would most likely be with other terminal organelle proteins, could eventually take to the formation of large supra-molecular complexes with quasi-symmetrical interactions as those observed in the mycoplasmas terminal organelle (Henderson and Jensen 2006; Seybert, Herrmann et al. 2006).

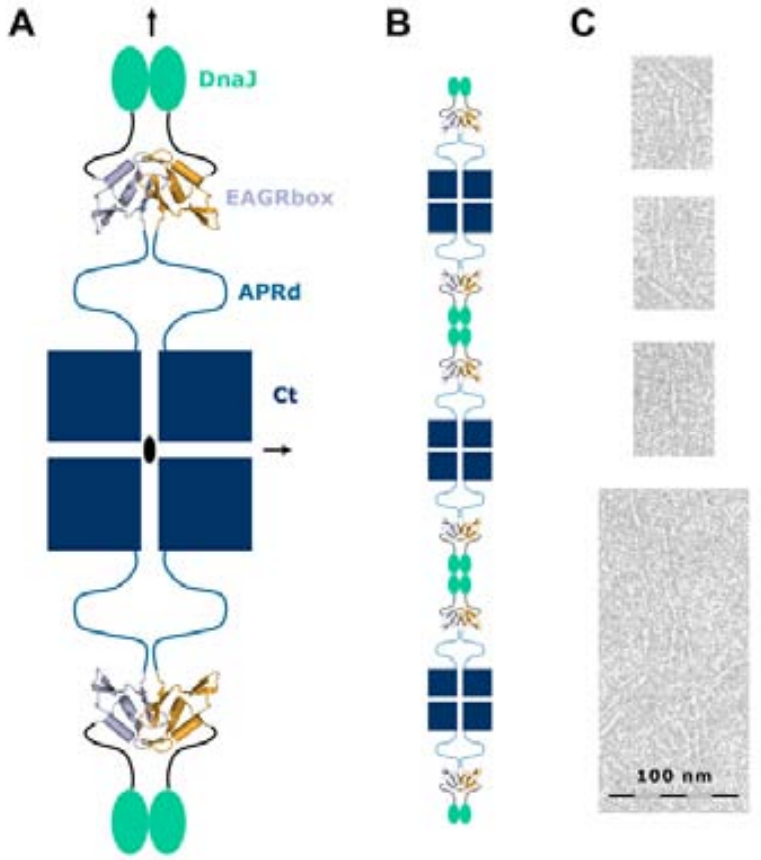


Figure III.34. Quaternary structure organization of the MG200 protein. (A) Hypotetical tetrameric organization of the MG200 protein with 222 (D2) symmetry. The C-terminal domain (Ct) would form the core of the tetramer while both DnaJ domain and EAGR box would be disposed along one of the inter-subunits two-fold axis. The APRd and the amino acidic sequence between the DnaJ domain and the EAGRb are thought to be unstructured. (B) Scheme of the possible fibre-like structures formed by the MG200 protein. (C) The tetrameric organization model for the MG200 protein is in accordance with the elongated structures, of ~30 nm, observed in EM micrographs of this sample (upper panels). Fibre-like structures can be also observed in the same preparation (lower panel). All images are shown at the same magnification. EM micrographs were obtained by Mercè Ratera from Dr. Ignacio Fita laboratory.

3. Experimental Procedures

In this section some experimental details will be given that concern the work performed specifically with the MG200 protein. General methods will be described on Chapter IV.

3.1. Cloning of the *fl*MG200 protein

The nucleotide sequence encoding for the *fl*MG200 protein was cloned between the *Nde*I/*Bam*HI restriction sites of the pET19b (Novagen, Madison, WI, USA) expression vector and also between the *Nde*I/*Xho*I restriction sites of the pET21d (Novagen, Madison, WI, USA) vector for heterologous expression in *E. coli*. The preparation of both constructs, pET19b-*fl*MG200 and pET21d-*fl*MG200, was undertaken in the laboratory of Doctors Enrique Querol and Jaume Piñol (Institut de Biotecnologia i de Biomedicina, UAB) by Oscar Q. Pich and Alícia Broto.

3.2. Expression and purification of the *fl*MG200

The *fl*MG200 protein was purified from a 2 L culture of *E. coli* BL21(DE3) modified with the pET19b-*fl*MG200 plasmid. Protein expression was induced with 0.5 mM IPTG when cells reached an optical density at 600 nm (OD_{600}) of 0.6 and cells were then incubated for 20 h at 16 ºC. The total protein cell extract was obtained in 0.02 M Tris-HCl pH 7.5, 0.5 M NaCl, 0.02 M imidazole, 1 mM EDTA disodium salt dihydrate, 0.5 mM DTT, 1 mg/mL lysozyme and a cocktail of protease inhibitors from Roche Diagnostics (Mannheim, Germany)). The recombinant *fl*MG200 protein was purified through a Ni^{2+}-affinity column (HisTrap™ HP 5 mL; GE Healthcare Life Sciences, Uppsala, Sweden) that was developed with a buffer system consisting of two solutions composed by 0.02 M Tris-HCl pH 7.5, 0.5 M NaCl, 1 mM EDTA disodium salt dihydrate, 0.5 mM DTT and 1 mM of the serine proteases inhibitor phenylmethylsulfonyl fluoride (PMSF) containing or not 0.5 M imidazole. Further purification steps consisted in loading the protein eluted from the affinity chromatography in a GFC column (Superdex 200 10/300; GE Healthcare Life Sciences, Uppsala, Sweden) equilibrated in 0.02 M Tris-HCl pH 7.5, 0.2 M NaCl, 1 mM EDTA disodium salt dihydrate, 0.5 mM DTT and 1 mM PMSF buffer and/or in an ion-exchange chromatography column (MonoQ 5/50 GL column; GE Healthcare Life Sciences, Uppsala, Sweden) which was developed in a gradient of 0-1 M NaCl.

E. coli BL21(DE3) cells containing the pET21d-*fl*MG200 plasmid were cultured in ZYM 5052 auto-inducing medium (Studier 2005). The high-density culture obtained was processed and purified in the same way as to obtain the *fl*MG200 protein from *E. coli* BL21(DE3) pET19b-*fl*MG200 cell cultures. The total protein yield was of 4 mg of recombinant protein per liter of culture. Protein aliquots were stored at -80 ºC in presence of 10 % (v/v) glycerol.

3.3. Study of the *fl*MG200 protein stability in solution

3.3.1. Differential scanning fluorimetry experiments

DSF measurements on *fl*MG200 protein solution were settled-up in a 96-well plate as follows: 5 µL of *fl*MG200 at 4 mg/mL in 0.02 M Tris-HCl pH 7.5, 0.2 M NaCl, 1 mM EDTA disodium sal dihydrate, 0.5 mM DTT and 1 mM DTT buffer were mixed with 7.5 µL 30× SYPRO orange dye (prepared in protein buffer) and 12.5 µL of the compound to be screened two times concentrated. Screened compounds comprised distinct buffer compositions with a broad pH range (from 3.5-11) and several salts at two distinct concentrations, 0.01 and 0.2 M. One of the plate wells was settled for negative control by adding initial protein buffer instead of protein solution to the mix. The temperature scan, from 25 to 85 ºC with 1 ºC/min increments, was performed in the real-time PCR BioRad iCycler5 instrument (Niesen, Berglund et al. 2007).

3.3.2. Circular dichroism measurements

CD scannings of the *fl*MG200 protein at 0.1 mg/mL in 0.02 M potassium phosphate pH 7.0, 0.2 M NaCl buffer were acquired over a wavelength range of 250-185 nm at 25 and 55 ºC in a 1 mm path-length cuvette.

CD thermal melt curves were measured at 218 nm and the CD signal was registered from 25 to 99 ºC applying a thermal gradient speed of 1 ºC/min.

3.3.3. Isoelectric focusing

To perform IEF of *fl*MG200 protein, 10 µg of protein solution in 0.02 M potassium phosphate pH 7.0 buffer were added to IEF loading buffer (7 M urea, 2 M thiourea, 2 % (w/v) CHAPS, 80 mM DTT and 2 % (w/v) ampholytes or 1 % (w/v) glycine). The first

gel dimension consisted on running the sample in a pH 4-7 gel strip and the second dimension on a 12 % (v/v) SDS-PAGE.

3.4. Cloning of the MG200 protein domain variants

PCR products were generated using *Pfu* DNA polymerase (Fermentas, Vilnius, Lithuania), the *mg200* gene as a template and the primers specified on Table III.6 for each domain variant to be cloned. The resulting PCR fragments were purified and double-digested with *Nco*I and *Xho*I restriction enzymes and ligated into the pET21d (Novagen, Madison, WI, USA) expression vector, which had been similarly digested.

Table III.6. Primer sets used in the amplification of *mg200* gene regions.

Final MG200 protein domain variants	Primer name	Primer sequence and direction	PCR product length (bp)
MG200-DnaJ$_{1-74}$*	MG200_F1‡	5' GATC*CCATGG*CTGAACAGAAACG	231
	MG200_R74†	3' CTATTCATACCAAAACTACCC*GAGCTC*TTGG	
MG200-DnaJ/EAGRb$_{1-207}$	MG200_F1	5' *GATCCCATGGCTGAACAGAAACG*	630
	MG200_R207	3' GCTTGGTCAACTAAGA*GAGCTC*TTGG	
MG200-EAGRb$_{75-207}$	MG200_F75	5' GATC*CCATGG*TTGATGGTGAACCTGC	405
	MG200_R207	3' GCTTGGTCAACTAAGA*GAGCTC*TTGG	
MG200-EAGRb$_{124-207}$	MG200_F124	5' GATC*CCATGG*CTAAGCAAGAACAACCTG	258
	MG200_R207	3' GCTTGGTCAACTAAGA*GAGCTC*TTGG	
MG200-DnaJ/EAGRb/APRd$_{1-295}$	MG200_F1	5' GATC*CCATGG*CTGAACAGAAACG	891
	MG200_R295	3' GGGTTGATTTCGATGATTC*GAGCTC*TTGG	
MG200-APRd$_{202-304}$	MG200_F202	5' GATC*CCATGG*CTAACGAACCAGTTGATTCTGAAACC	315
	MG200_R304	3' CCTACTAGAAAACAAACTGTTGTAATTG*GAGCTC*TTGG	
MG200-APRd/Ct$_{202-601}$	MG200_F202	5' GATC*CCATGG*CTAACGAACCAGTTGATTCTGAAACC	1206
	MG200_R601	3' GGAGAGGGTTCTTGGGTAATCA*GAGCTC*TTGG	
MG200-Ct$_{293-601}$	MG200_F293	5' GATC*CCATGG*CTACTAAGGATGATCTTTTGTTTGAC	921
	MG200_R601	3' GGAGAGGGTTCTTGGGTAATCA*GAGCTC*TTGG	
MG200-Ct$_{304-601}$	MG200_F304	5' GATC*CCATGG*CTAACCTCACTACCTATGAACAAGTTG	888
	MG200_R601	3' GGAGAGGGTTCTTGGGTAATCA*GAGCTC*TTGG	

* Numbers in subscript refer to the MG200 amino acidic sequence numbering;
‡ F - forward primer, † R - Reverse primer.

3.4.1. MG200-EAGRb$_{124-207}$ mutagenesis

The vector pET21d-MG200-EAGRb$_{124-207}$ was used as template for PCR-based site-directed mutagenesis. Two complementary oligonucleotides, for each mutation to perform (Table III.7), were synthesized to generate the following single mutants: I140M, L156M and L172M. Mutation of a single oligonucleotide, from ATT to ATG in the case of I140M and from TTG to ATG in the case of L172M, resulted in a codon change. To obtain the L156M mutant two oligonucleotides had to be changed, from TTA to ATG, in order to achieve the correct codon change.

Table III.7. Sets of mutagenic primers used for PCR-generated MG200-EAGRb$_{124-207}$ mutants.

Mutant location*	Primer name	Primer sequence and direction
I140M	I141M_F‡	5' TGTTGAGCAAACC*ATG*AAAAAGGTGCAAC
	I141M_R†	3' ACAACTCGTTTGG*TAC*TTTTTCCACGTTG
L156M	L156M_F	5' AAAGACCCAGATGAA*ATG*CGTTCTAAGGTCC
	L156M_R	3' TTTCTGGGTCTACTT*TAC*GCAAGATTCCAGG
L172M	L172M_F	5' TAGTGATTGGGAAGCA*ATG*GTTGGTGATACTGG
	L172M_R	3' ATCACTAACCCTTCGT*TAC*CAACCACTATGATCC
Y192D	Y192D_F	5' GGAGTTGGAAGGGT*GAC*TTTGATGAACAGG
	Y192D_R	3' CCTCAACCTTCCCA*CTG*AAACTACTTGTCC

* Numbering refers to the MG200 amino acidic sequence numbering;
‡ F - forward primer, † R - Reverse primer

Another single mutation, Y192D, was introduced in the native MG200-EAGRb$_{124-207}$ by overlap-extension PCR with primers containing the appropriate target substitutions (Table III.7). In this case, the TAC codon had to be changed to GAC to achieve the correct codon change.

3.5. Expression and purification of the MG200 protein domain variants

A standard protocol was established to express and purify the newly cloned MG200 domain variants. Each domain was expressed 1L *E. coli* BL21(DE3) cell cultures grown in ZYM 5052 auto-inducing media for 20 h. In the case of the MG200 C-terminal domain variants each construct was co-transformed with the pGro7 vector (Takara Bio Inc., Shiga, Japan) in *E. coli* BL21(DE3) cells. These were cultured at 37 ºC in 2 L of YT2× media until reach an OD_{600} of 0.4. Afterwards the culture was induced with 0.02 % (w/v) L-arabinose for 30 min (to induce GroES-GroEL expression) and cooled-down to 20 ºC. The recombinant protein expression was then induced with 1 mM IPTG for 16 h at 20 ºC.

Total protein cell extracts were obtained in 0.02 M potassium phosphate pH 7.0, 0.5 M NaCl, 0.02 M imidazole buffer and processed as described on section IV.4.1. All the MG200 protein domain variants were purified using Ni^{2+}-affinity followed by gel filtration chromatographies. GFCs were performed on a Superdex 75 16/60 column (GE Healthcare Life Sciences, Uppsala, Sweden) equilibrated in 0.02 M potassium phosphate pH 7.0, 0.1 M NaCl buffer.

3.6. Nuclear Magnetic Resonance measurements

NMR spectra were acquired for the soluble MG200 N-terminal domain variants. The experiments were settled for highly pure and concentrated protein domains labelled with

the ^{15}N isotope in thin-walled glass tubes and performed at 7, 25 or 34 ºC. Sample volumes were of 150-300 µL at concentrations ranging from 0.12-1.20 mM. Protein solutions were in 0.02 M potassium phosphate pH 6.5, 0.1 M NaCl buffer.

3.7. Crystallization of the soluble MG200 protein domain variants

3.7.1. Crystallization of the MG200 C-terminal domain

Crystals were obtained in several crystallization conditions from both MG200 C-terminal domain variants on initial crystallization trials performed at 20 ºC in nanoliter-scale drops. The conditions were A3, A7 and B6 from the JBScreen classic 9 (Jena Bioscience, Jena, Germany) commercial screen: 0.1 M HEPES pH 7.5, 0.2 M sodium citrate and 10-25 % (v/v) iso-propanol. MG200-$Ct_{293-601}$ and MG200-$Ct_{304-601}$ were screened at 3 and 4 mg/mL, respectively.

Crystallization conditions were reproduced and optimized at microliter-scale. The best crystallization condition was obtained at 4 ºC for the MG200-$Ct_{304-601}$ domain variant in a hanging-drop vapour-diffusion experiment by mixing equal volumes of protein and reservoir solution and microseeds from crystals of the same protein domain variant. The final crystallization condition was 0.1 M HEPES pH 7.5, 0.4 M sodium citrate and 8 % (v/v) iso-propanol. The protein was at a concentration of 2.5 mg/mL in 0.02 M Tris-HCl pH 7.0, 0.1 M NaCl, 1 mM EDTA disodium salt dihydrate, 0.5 mM DTT and 10 % (v/v) glycerol buffer.

3.7.2. Crystallization of MG200-EAGRb$_{124-207}$

Crystallization drops from initial assays performed at 20 ºC with the MG200-EAGRb$_{124-207}$ domain concentrated to 30 mg/mL were almost exclusively transparent, precipitate-free, except for condition 40 from the Index commercial screen (Hampton Research, CA, USA) where thin needles were observed. The initial crystallization condition, composed of 0.1 M citric acid pH 3.5 and 25 % (w/v) PEG3350, was optimized to 0.1 M sodium citrate pH 4.5 and 22 % (w/v) PEG 2000.

MG200-EAGRb$_{124-207}$ crystals grew within 7 days at 20 ºC by the hanging-drop vapor-diffusion method. Equal microliter-volumes of protein and reservoir solution were mixed together and crystals were exclusively obtained in the acidic pH range of 3.5-4.5.

3.8. MG200-EAGRb$_{124-207}$ structure determination

3.8.1. X-ray diffraction data collection and processing

MG200-EAGRb$_{124-207}$ contains no methionine that could be used to prepare SeMet-containing crystals from this domain variant. To overcome this handicap an extensive search of heavy-atom derivative crystals was undertaken.

Heavy-atom compounds used to derivatize MG200-EAGRb$_{124-207}$ crystals were selected in the same way as proceeded to derivatize the MG438 protein and the protocol used to prepare the heavy-atom derivative crystals was also identical (section II.3.4.). Native and derivative crystals were frozen in liquid nitrogen in presence of 20 % (v/v) glycerol. X-ray diffraction data were collected at 100 K at the ESRF beam lines ID14eh4, ID23eh2 and ID29 and data were processed with the program packages DENZO and SCALEPACK (Otwinowski 1997). The HKL2MAP program package (Pape 2004) was used to search for heavy-atoms in the derivative crystals.

3.8.2. Structure determination and refinement

SeMet-labelled MG200-EAGRb$_{124-20}$ crystals belonging to the trigonal space group $P3_121$ contained two protein subunits per asymmetric unit. Initial phasing was performed with SAD data from a MG200-EAGRb$_{124-207}$ SeMet156 crystal at 3.1 Å resolution. Two Se atoms were found with SHELXD (Schneider and Sheldrick 2002) and refinement of the heavy-atoms positions was carried out with the same program. A partial automatic model building was performed with the program ARP/wARP 6.0 (Perrakis, Morris et al. 1999) followed by refinement performed in several REFMAC5 cycles (Murshudov, Vagin et al. 1997), including TLS and restrained B_{factor} refinement, intercalated with manual rebuilding with COOT (Emsley and Cowtan 2004). The diffraction data set used for refinement was from a MG200-EAGRb$_{124-207}$ native crystal at 2.9 Å resolution. The final model was refined to a R_{factor} of 18.5 and a R_{free} 24.0 and all the residues included in the model lie in the most favored region of the Ramachandran plot (Laskowski 1993).

Structure factors and coordinates were deposited in the RCBS Protein Data Bank under accession code number 3N9F.

CHAPTER IV

Materials and Methods

CHAPTER IV. MATERIALS and METHODS

1. List of material and equipments

- Cartesian Honeybee pippeting robot (Hamilton, Reno, NV, USA)
- Columns and slurries for protein chromatography were from GE Healthcare Life Sciences (Uppsala, Sweden)
- Crystal Phoenix pipetting robot (ARInstruments, Sunnyvale, CA, USA)
- Enzymes for DNA restriction and for standard DNA procedures were from Fermentas (Vilnius, Lithuania)
- Eppendorf Mastercycler® PCR machine (Eppendorf, Hauppauge, NY, USA)
- Gene Genius Bio Imaging System (Syngene, Cambridge, UK)
- GFX™ PCR DNA and gel band Purification Kit (GE Healthcare, Buckinghamshire, UK)
- iCycler5 real-time PCR instrument (BioRad, Hercules, CA, USA)
- Immobilon-PVDF Membrane (Millipore, Billerica, MA, USA)
- IPGphor™ IEF System (GE Healthcare Life Sciences, Uppsala, Sweden)
- Jasco J-810 spectropolarimeter (Easton, MD, USA)
- Mini-protean Tetra-Cell electrophoresis system (BioRad Laboratories, Hercules, CA, USA)
- Mini Trans-Blot Electrophoretic Transfer Cell (BioRad Laboratories, Hercules, CA, USA)
- Nano-drop® spectrophotometer (Thermo Scientific, Wilmington, DE, USA)
- NMR spectrometers 600 and 800 MHz Bruker Digital Avance (San Francisco, CA, USA)
- Oligonucleotide primers for PCR amplification were synthesized by Roche Diagnostics (Mannheim, Germany)
- Perfect DNA™ 1 kbp Ladder (Novagen, Madison, WI, USA)
- QIAprep® Spin Miniprep Kit (QIAGEN, Hilden, Germany)
- SeeBlue® pre-stained molecular weight standards (Invitrogen, Carlsbad, CA, USA)
- SYBR® Safe (Invitrogen, Carlsbad, CA, USA)
- SYPRO® Orange (Invitrogen, Carlsbad, CA, USA)

- Zetasizer Nano DLS instrument (Malvern, Worcestershire, UK)

2. Bacterial strains and vectors

Bacterial strains used for plasmidic DNA amplification and cloning:

- ***E. coli* XL1 Blue** *rec*A1, *end*A1, *gyr*A96, *thi*-1, *hsd*R17 (r_K^-, m_K^+), *sup*E44, *rel*A1, *lac*, [F', *proAB*, *lac*$I^qZ\Delta M15$::Tn10(tet^r)].
- ***E. coli* DH5α** *rec*A1, *end*A1, *gyr*A96, *thi*-1, *hsd*R17(r_K^-, m_K^+), glnV44, *rel*A1, *deo*R, *nup*G, [F^-, Φ80d*lacZ*ΔM15, Δ(*lacZYA-argF*)U169, λ^-].

Bacterial strains used for recombinant protein expression

- ***E. coli* BL21(DE3)** F^-, *omp*T, *hsd*S_B(r_B^-, m_B^-), *dcm*, *gal*, λ(DE3)
- ***E. coli* BL21(DE3) pLysS** F^-, *omp*T, *hsd*S_B(r_B^- m_B^-), *dcm*, *gal*, *lon*, λ(DE3), pLysS (cm^R). pLysS plasmid is chloramphenicol resistant and encodes for the T7 phage lysozyme, an inhibitor for T7 polymerase which reduces and almost eliminates expression from transformed T7 promoter containing plasmids when not induced.

Vectors for protein expression :

- **pET19b**, commercial vector (Novagen, Madison, WI, USA) used for recombinant protein expression in *E. coli* that adds a His_9-tag to the N-terminal of the interest protein followed by an enterokinase target site. The vector is ampicillin resistant.
- **pET21d**, commercial vector (Novagen, Madison, WI, USA) used for recombinant protein expression in *E. coli* that adds a His_6-tag to the C-terminal of the interest protein. The vector is ampicillin resistant.
- **pGro7**, commercial vector (Takara, Bio Inc., Shiga, Japan) used for expression of the *E. coli* GroES and GroEL chaperones. The vector is chloramphenicol resistant.

3. Molecular biology procedures

3.1. Microbiologic methods

3.1.1. Bacterial culture media composition

Solid medium preparation

Luria-Bertani (LB) *broth/agar* (1L):

10 g tryptone, 5 g yeast extract, 10 g NaCl and 15 g agar. Autoclave, let the medium cool down to ~60 ºC, add the appropriate supplements and distribute 30-35 mL in Petri dishes.

Liquid media preparation

LB broth (1 L):

10 g tryptone, 5 g yeast extract and 10 g NaCl (autoclave).

YT2× (1L):

16 g tryptone, 10 g yeast extract and 5 g NaCl (autoclave).

ZYM 5052 auto-inducing medium for protein production in high-density shaking cultures (Studier 2005)

Stock solutions:

- 1 M $MgSO_4$ prepared in water
- 1000× trace metals solution (1 L) - 8mL 5 M HCl, 5 g $FeCl_2.4H_2O$, 184 mg $CaCl_2.2H_2O$, 64 mg H_3BO_3, 18 mg $CoCl_2.6H_2O$, 4 mg $CuCl_2.2H_2O$, 340 mg $ZnCl_2$, 605 mg $Na_2MoO_4.2H_2O$ and 40 mg $MnCl_2.4H_2O$ (filter through 0.22 μm filter).
- 50× 5052 solution - 25 % (v/v) glycerol, 2.5 % (w/v) glucose and 10 % (w/v) lactose (autoclave).
- 50× M solution - 1.25 M Na_2HPO_4, 1.25 M KH_2PO_4, 2.5 M NH_4Cl and 0.25 M Na_2SO_4 (filter through 0.22 μm filter).

Final ZYM 5052 auto-inducing media composition (1 L):

- Autoclave 8 g of tryptone and 4 g of yeast extract in water.
- Add to the previously autoclaved media 1.6 mL 1 M $MgSO_4$, 1.6 mL 1000× trace metals solution, 16 mL 50× 5052 solution, 16 mL 50× M and the appropriate antibiotics.

Procedure for induction:

- A pre-culture must be grown overnight at 37 ºC. After pre-inoculate the culture media let cells grow for 3 h at 37 ºC and then start to slowly diminish the

temperature from 37 to 20 ºC. Cells can be incubated for 36 h in these conditions previous to cell harvesting.

Minimal media for expression of SeMet-labelled proteins (Doublie 1997).

Stock solutions

- 5× salt solution (1L) - 28 g K_2HPO_4, 8 g KH_2PO_4, 4 g $(NH_4)_2SO_4$, 1 g tri-sodium citrate and 0.4 g $MgSO_4.7H_2O$ (autoclave).
- 2 M glucose solution (100 mL, filter through 0.22 µm filter).
- 5× amino acid solution (0.5 L) - 0.1 g of Ala, Arg, Asn, Asp, Cys, Gln, Glu, Gly, His, Pro, Ser, Trp, Tyr and 0.25 g of Ile, Leu, Lys, Phe, Thr and Val. Amino acids can be dissolved at 60 to 80 ºC with stirring and the pH of the final solution must be adjusted to 7.5 (filter through 0.22 µm filter).
- 10 mg/mL SeMet solution (should be prepared fresh each time).
- 4 mg/mL thiamine solution prepared in water (filter through 0.22 µm filter).
- 4 mg/mL thymine solution prepared in water (filter through 0.22 µm filter).

Final media composition (1L):

- 200 mL 5× salt solution, 200 mL 5× amino acid solution, 8 mL 4 mg/mL thiamine, 8 mL 4 mg/mL thymine, 16 mL 2 M glucose solution, 5 mL SeMet solution and appropriate antibiotics.

Procedure for induction:

- In order to condition the bacterial cells to the minimal media, overnight culture should be done with methionine minimal media for ~12 h. Other expression factors such as induction temperature and length should be kept the same as for the protein expression in regular media, although these variables can be modified in an attempt to increase yield.

^{9}M medium for production of proteins marked with ^{15}N

Medium composition (1 L):

- Mix 4 g $Na_2HPO_4.2H_2O$, 3 g KH_2PO_4, 0.5 g NaCl, 1 g $^{15}NH_4Cl$ and autoclave.

- Add 20 mL 20 % (w/v) glucose, 2 mL 1 M $MgSO_4$, 1 mL 0.1 mM $ZnCl_2$, 50 µL 1 M $CaCl_2$, 2 mL 0.9 % (w/v) biotin and 1 mL 0.5 % (w/v) thiamine (filter each solution though a 0.22 µm filter).

3.1.2. Antibiotics and supplements

- 1000× ampicillin stock solution - 100 mg/mL ampicillin was prepared in water and filtered through a 0.22 µm filter.
- 1000× chloramphenicol stock solution - 34 mg/mL chloramphenicol was prepared in 96 % (v/v) ethanol.
- Isopropyl-β-D-thiogalactoside (IPTG) stock solution - 1M IPTG was prepared in water and filtered through a 0.22 µm filter. This compound was used to induce the expression of genes that were under the control of the *lacZ* gene promotor.

3.1.3. *E. coli* culture conditions

E. coli cells were cultivated in LB/agar plates and, after growing at 37 ºC, an isolated single colony was inoculated into 4 mL LB media and allowed to grow overnight until a supersaturated culture was obtained. After, this pre-culture was inoculated into fresh LB media (1/200 dilution factor) and grown at 37 ºC with shaking (250 rpm) until reach an OD_{600} of 0.6-0.8. Cells were then equilibrated at the final expression temperature, induced with 0.1-1 mM IPTG and incubated with shaking at the final expression temperature from 3 to 16h. All media contained the appropriate antibiotic depending on the bacterial strain and the expression vector to be used. Protein over-expression was undertaken in cell cultures of 0.5 to 4 L.

3.1.4. Heat-shock transformation of bacterial cells

Bacterial competent cells are thawed on ice. For each sample to be transformed 50 µL of cells are used in a sterilized and pre-chilled eppendorf tube. Add 1 µL of the sample (except when otherwise indicated), mix gently and incubate on ice for 20 min. Heat-pulse the transformation reactions for 45 sec at 42 ºC, place on ice for 2 min, add 0.5 mL LB media pre-heated at 37 ºC and incubate the reactions for 1 h at 37 ºC. Transformation reaction products are spread on LB/agar plates containing the appropriate antibiotics and colonies allowed to grow overnight at 37 ºC (Sambrook 2001).

3.2. Recombinant DNA technology

3.2.1. Extraction of plasmidic DNA

The QIAprep® Spin MiniPrep Kit (QIAGEN, Hilden, Germany) was used for extraction of up to 20 μg of high-pure plasmid used later for sequencing and cloning. Bacterial cultures are lysed and the lysates cleared by centrifugation. These are then applied to the QIAprep columns and plasmid DNA elution performed according to the manufacturer's guidelines.

3.2.2. DNA quantification

NanoDrop® spectrophotometer was used to quickly measure DNA absorbance at 260 nm in microliter-scale drops and to calculate DNA concentrations following Thermo Scientific (Wilmington, DE, USA) indications.

3.2.3. DNA electrophoresis in agarose gels

Agarose gel electrophoresis is a standard technique to separate DNA molecules or fragments by size by applying an electric potential. The agarose powder was mixed with the TAE 1× (0.04 M Tris-acetate pH 8.5, 1 mM EDTA disodium salt dihydrate) electrophoresis running buffer to the desired concentration (1 % (w/v) agarose unless otherwise indicated), and then heated in a microwave oven until completely melted. After cooling the solution to about 60 ºC it was poured into a casting tray containing a sample comb and allowed to solidify at room temperature. Samples containing DNA mixed with 6× loading buffer (0.24 M Tris-acetate pH 8.0, 0.06 M EDTA, 0.6 % (w/v) SDS, 0.003 % (w/v) bromophenol blue, 0.003 % (w/v) xylene cyanol and 12 % v/v) glycerol) were loaded into the sample wells and a current was applied. To estimate the size of the DNA fragments separated in the gel the commercial Perfect DNA™ 1 kbp Ladder (Novagen, Madison, WI, USA) was used as molecular weight marker.

Gels were stained with SYBR® Safe (Invitrogen, Carlsbad, CA, USA) or ethidium bromide (when very low DNA quantities add to be detected) to allow DNA visualization after electrophoresis through a transilluminator.

3.2.4. DNA purification from agarose gels

To obtain highly pure DNA fragments from PCR amplification reactions or plasmid DNA restriction reactions samples were loaded in a 1 % (w/v) agarose gel and bands

with the correct size were excised from it. The excised gel fragments were subjected to DNA extraction following the manufacturer's instructions for the use of the GFX™ PCR DNA and gel band Purification Kit (GE Healthcare, Buckinghamshire, UK).

3.2.5. DNA fragments amplification by PCR

DNA fragments for the preparation of new constructs or cloning into different expression vectors were amplified by Polymerase Chain Reaction (PCR) with the usage of proper primer pairs, designed according to the fragment to be amplified and the vector where the fragment has to be inserted. PCR was also used to incorporate mutations by site-directed mutagenesis.

A standard PCR reaction was carried out in a final volume of 50 µL containing 40 µM dNTPs, 0.1 µM of each primer, 10 ng DNA template and 0.05 U/µL *Pfu* DNA polymerase (Fermentas, Vilnius, Lithuania) in 0.02 M Tris pH 8.8, 0.01 M KCl, 0.01 M $(NH_4)_2SO_4$, 2 mM $MgSO_4$ and 0.1 % (v/v) Triton X-100 reaction buffer. PCR was performed in the Eppendorf Mastercycler® PCR machine (Eppendorf, Hauppauge, NY, USA) and started with an initial denaturation step at 94 °C for 2 min, followed by 5 cycles at 94 °C, 55 °C and 72 °C, 25 more cycles at 94 °C, 62 °C and 72 °C and by a final extension step of 10 minutes at 72 °C. Duration of each denaturation, annealing and elongation steps depended on the size of the product to amplify, taking into account that *Pfu* DNA polymerase takes 2 min to amplify a 1 kb fragment.

3.2.6. Restriction enzymatic reactions

PCR amplification products and plasmidic DNA were digested with restriction enzymes purchased from Fermentas (Vilnius, Lithuania) according to the manufacturer's guidelines. Digestion products were purified from agarose gels.

3.2.7. Ligation enzymatic reactions

Pure, double-digested PCR products and vectors were mixed together and the ligation reaction was settled as recommended by Fermentas (Vilnius, Lithuania). Typically, reactions were performed on a final volume of 10 µL and incubated at 20 ºC for 16 h.

Ligation products (2-10 µL) were transformed in *E. coli* DH5α for plasmid DNA amplification. Plasmid DNA was then extracted from cells and used to verify the

correctness of the DNA sequence and to transform bacterial cells dedicated to protein expression.

3.2.8. Site-directed mutagenesis

Mutagenic oligonucleotide primers were designed to introduce single amino acid changes in the native sequence (Hutchison, Phillips et al. 1978).

Mutagenic PCR reaction mixture was carried out in a final volume of 50 µL containing 1× reaction buffer (0.02 M Tris pH 8.8, 0.01 M KCl, 0.01 M $(NH_4)_2SO_4$, 2 mM $MgSO_4$ and 0.1 % (v/v) Triton X-100), 10 ng dsDNA template, 0.1 µM of each forward and reverse mutagenic olignucleotide primers, 40 µM dNTPs mix and 0.05 U/µL *Pfu* DNA polymerase (Fermentas, Vilnius, Lithuania). After an initial denaturation step at 95 ºC for 30 sec, the samples were subjected to 16 cycles of 95 ºC for 30 sec, 55 ºC for 1 min and 68 ºC for 12 min (last step duration can change according to the size of the amplification fragment).

The PCR product was placed on ice to cool the reaction to ~37 ºC. *Dpn*I (Fermentas, Vilnius, Lithuania) was added to the amplification reaction and incubated for 1 h at 37 ºC. After digestion, 10 µL of the *Dpn*I-treated DNA was transformed in *E. coli* XL1 Blue supercompetent cells. The obtained transformants were sequenced in the Genomic platform of the SCT-UB to check for positive mutant colonies.

3.2.9. DNA sequencing

Sequence verification of new constructs and mutant vectors was performed in the Transcriptomics Platform of the SCT-UB.

4. Protein production and general analysis methods

4.1. Preparation of total protein extracts

Cells were harvested by centrifugation at 4000 *g* for 20 minutes at 4 ºC, resuspended in lysis buffer (20 mL per 1 L of culture) of 0.02 M Tris-HCl pH 7.5 (unless otherwise indicated), 0.5 M NaCl, 0.02 M imidazole and 1 mg/mL lysozyme supplemented with a cocktail of protease inhibitors from Roche Diagnostics (Mannheim, Germany). The extract was incubated for 10 min at room temperature and cells were mechanically disrupted by sonication (5 cycles of 20-30 sec at 250 W). The lysate was then

centrifuged at 39000 *g* for 30 minutes (4 ºC) and the cleared lysate, recovered from the cell debris, was filtered through a 0.22 μm filter.

4.2. Protein electrophoresis in polyacrylamide gels

4.2.1. Protein electrophoresis in denaturing conditions (SDS-PAGE)

Polyacrylamide gel electrophoresis (Laemmli and Favre 1973) was preformed in a slab gel following standard procedures. The samples were denatured before loading into the gel by heating at 97 ºC for 5 min in the presence of 5 % (w/v) SDS and 0.025 % (w/v) bromophenol blue.

Molecular weight SeeBlue® pre-stained protein standards (Invitrogen, Carlsbad, CA, USA) were myosin (250 kDa), BSA (98 kDa), glutamic dehydrogenase (64 kDa), alcohol dehydrogenase (50 kDa), carbonic anhydrase (36 kDa), myoglobin (30 kDa), lysozyme (16 kDa), aprotinin (6 kDa) and insulin, B chain (4 kDa).

4.2.2. Protein electrophoresis in native conditions (PAGE)

Electrophoresis of native proteins were performed on 5 % (v/v) polyacrylamide gels using the same buffer system as for denaturalising PAGE lacking SDS. Gels were run in a cold chamber to avoid heating.

4.3. Isoelectric Focusing

IEF, or electrofocusing, is a technique for separating different molecules by their electric charge differences. It is a type of zone electrophoresis usually performed in a gel, which is based on the fact that a molecule's charge changes with the pH of its surroundings.

Proteins, as amphoteric molecules, carry positive, negative or zero net charges depending on the pH of their local environments. The overall charge of a particular protein is determined by the ionizable acidic and basic side chains of it constituent amino acids and prosthetic groups. The net charge on a protein is the algebraic sum of all its positive and negative charges. Thus, there is a specific pH for every protein at which the net charge it carries is zero. This isoelectric pH value, termed pI, is a characteristic physico-chemical property of every protein.

In electrophoresis, the net charge of a protein determines the direction of its migration (electrophoretic motility). At pH values below the pI of a particular protein it will migrate towards the cathod. Conversely, at pH values above its pI a protein will move towards the anode. IEF carried out under non-denaturing conditions is a high resolution technique that allows separating proteins that differ in their pI values by only 0.02 pH units. Due to the technique high resolution protein samples which appear to be homogeneous when tested by other means can often be separated into several components by IEF. Such micro-heterogeneity can be indicative of differences in primary structure, conformational isomers, differences in composition and number of prosthetic groups or partial denaturation.

Samples for IEF were in a typical biochemical buffer, almost salt free. Small sample quantities (5-30 μg) were added to IEF loading buffer (7 M urea, 2 M thiourea, 2 % (w/v) CHAPS, 0.08 M DTT and 2 % (w/v) ampholytes or 1 % (w/v) glycine). The sample was loaded in the sample holder of the IPGphor IEF system (GE Healthcare Life Sciences (Uppsala, Sweden), it was covered with a gel strip with the proper pH range and with mineral oil. The run was then performed according to the manufacturer's instruction manual. After running the first gel dimension the strip was washed for 15 min in equilibration buffer (0.05 M Tris-HCl pH 8.8, 6 M urea, 30 % (v/v) glycerol and 0.025 % (w/v) bromophenol blue) containing 2 % (w/v) SDS and washed for another 15 min in equilibration buffer containing 0.5 mM iodacetamide. The gel strip was then loaded in SDS-PAGE to run the second gel dimension. The gel was stained in CBB.

4.4. Polyacrylamide protein gels staining

4.4.1. Polyacrylamide protein gels staining in CBB

Staining of polyacrylamide protein gels with CBB is a quantitative and protein specific method that can detect 0.3-1.0 μg of protein per band. A completely blue gel is obtained after submerge the gel for at least 15 min in a 0.1 % (w/v) CBB R-250 solution prepared in 7 % (v/v) acetic acid and 15 % (v/v) ethanol. The gel is then distained in 7 % (v/v) acetic acid and 15 % (v/v) ethanol in order to observe the protein bands.

4.4.2. Silver staining of polyacrylamide protein gels

The silver staining of polyacrylamide protein gels method is more sensitive than the CBB staining method and was used in cases where relatively low amount of protein had to be detected, as in the case of protein solutions proceeding from protein crystals.

The gel is incubated for 1 h in a fixative solution containing 50 % (v/v) methanol, 12 % (v/v) acetic acid and 0.02 % (v/v) formaldehyde. It follows a 1 h washing in 50 % (v/v) ethanol, changing the washing solution every 20 min. The gel is then pre-treated with 0.4 mM sodium thiosulfate for 1 min, washed in deionised water 2 times for 20 sec each and incubated in 5 mM silver nitrate solution containing 0.03 % (v/v) formaldehyde for another 20 min. After rinse the gel 2 times on deionised water the water is drained and the developer solution (0.4 M sodium carbonate containing 0.02 % (v/v) formaldehyde) added. When the desired contrast is obtained the gel is quickly washed 2 times in deionised water and the stop solution (50 % (v/v) methanol and 12 % (v/v) acetic acid) added. A solution of 50 % (v/v) methanol is finally used to destain the gel background and reinforce protein bands visualization.

4.5. Electroblot onto PVDF membranes

Samples are separated by SDS-PAGE. Meanwhile, the PVDF membrane is rinsed for a few seconds in 100 % (v/v) methanol for activation, washed in MiliQ water and stored in transfer buffer (0.025 M Tris pH 8.3, 0.192 M glycine, 20 % (v/v) methanol and 0.05 % (w/v) SDS). The gel is sandwiched between the activated PVDF membrane sheet and several sheets of 3MM paper (Millipore), assembled in a wet blotting apparatus and transferred at 4 ºC for 2 h at 80 V. The membrane is then stained for 1 min in 0.1 % (w/v) CBB R-250 prepared in 50 % (v/v) methanol, destained for 10 min (50 % (v/v) methanol, 10 % (v/v) acetic acid) and washed with water. The protein bands electrobloted onto the PVDF membrane were used for N-terminal EDMAN sequencing.

4.6. Edman sequencing

Edman sequencing of protein bands were undertaken in the Proteomics platform of the SCT-UB.

4.7. Protein quantification

4.7.1. Bradford assay for protein quantification

To measure the protein concentration in a protein extract or pure protein solution the Bradford dye-binding assay is performed as follows: the Bradford reagent is 5 times diluted in MiliQ water, 20 µL of protein solution are added to 1 mL of the diluted reagent and mixed and the protein-induced absorbance shift of Coomassie blue dye is measured at 595 nm (Bradford 1976). A standard protein concentration curve is prepared, using a dilution series (0.1-1.0 mg/ml) of known BSA concentration dissolved in the same buffer of the interest protein, to quantify the protein samples.

4.7.2. Protein quantification on NanoDrop® spectrophotometer

Proteins that contain Trp, Tyr residues or Cys-Cys disulphide bonds absorb at 280 nm making absorbance spectroscopy a fast and convenient method for the quantification of purified protein preparations. Measuring absorbance at 280 nm of a protein solution in a concentration range from 0.1 to 100 mg/mL has the advantage that no additional reagents are needed and a standard protein concentration curve is not required.

A 2 µL protein solution drop is applied onto the NanoDrop® spectrophotometer lower measurement pedestal. Following, the opposite pedestal is brought into contact with the liquid sample and a liquid bridge is formed with a precise path-length of 1 mm. The sample is then excited by the UV light and the light emitted from the sample is captured by the spectrophotometer. The Beer-Lambert equation, $A = E \times b \times c$ (where $\underline{A}$ denotes the absorbance value, $\underline{E}$ the extinction coefficient, $\underline{b}$ the path-length and $\underline{c}$ the analyte concentration) is used to correlate absorbance with protein concentration.

4.8. Protein purification chromatographic techniques

In order to study a protein structure the protein must be isolated and purified to homogeneity. The high purity level is generally achieved after several chromatographic purification steps. As each step usually results in some degree of product loss, a purification strategy was defined to reach the highest purification level in the fewest steps. These were selected depending on the target protein size, charge and solubility. Chromatographic methods were applied using automated FPLC equipment.

Immobilized metal ion affinity chromatography

The first purification step involved Ni^{2+}-based immobilized metal ion affinity chromatography of the soluble cytoplasmatic extract. The target protein, which contained a histidine tag, bound specifically to the column beads. Affinity columns coupled with Ni^{2+} ions (HisTrap™ HP; GE Healthcare Life Sciences, Uppsala, Sweden) were equilibrated with 0.02 M imidazole to avoid unspecific binding and improve sample purity on the very fist step of protein purification. Affinity column development was performed according to the manufacturer's guidelines (GE Healthcare Life Sciences, Uppsala, Sweden).

Ion-exchange chromatography

Ion-exchange chromatography allows separating proteins based on charge. Mono Q 5/50 GL column (GE Healthcare Life Sciences, Uppsala, Sweden) is prepared for anion exchange, containing a positively charged stationary phase that attracts negatively charged molecules. Elution of the target protein is done by applying a salt concentration gradient in the column (0-1 M NaCl), which resulted in a change or neutralization of the charged functional groups of the protein. The samples to be purified by this method were dialysed against a buffer with low salt content, 0-50 mM NaCl.

Gel filtration chromatography (size-exclusion)

GFC was used as the last purification step, the polishing step used to further purify each protein and evaluate its homogeneity and oligomerization state. In this chromatographic method the larger proteins are separated from the smaller ones since larger molecules travel faster through the cross-linked polymer in the chromatography column. Eluate was collected in a series of tubes separating proteins based on elution time.

4.9. Protein cross-linking with gluteraldehyde

In order to characterize the possible oligomeric forms of a purified protein in solution 10 μL drops of 0.5-1 mg/mL protein solution were equilibrated from 5 min to 2 h against 40 μL drops of acidified 25 % (v/v) gluteraldehyde (Fadouloglou, Kokkinidis et al. 2008). The protein drops were then analyzed by SDS-PAGE and the resulting gel stained with CBB.

5. Biophysical methods for protein characterisation

Crystal structure determination of interest proteins require a complete understanding of how they behave in solution in relation to its stability, conformation, aggregation or complex formation as a function of buffer conditions, temperature and time. In this work several biophysical techniques were explored aiming to obtain a more complete understanding of the proteins in study.

5.1. Differential scanning fluorimetry

Protein samples must be stable over long periods of time in order to determine activities, interactions or structures. In the optimal conditions protein denaturation is prevented while proteins are stored or during freezing. There are a large number of parameters that affect protein stability such as temperature, buffer composition and pH, salt or detergents, ligands that bind to the proteins in a specific manner and, sometimes, protein concentration. The temperature of melting of a protein is a good parameter to evaluate protein stability in solution given that it depends on the Gibbs free energy of unfolding, which is temperature dependent. In order to determine the conditions in which the proteins are most stable in solution melting temperature variations can be measured in presence of buffers with different compositions.

In the DSF method, protein unfolding is monitored in the presence of a fluorescent dye in a real-time PCR instrument. SYPRO® Orange (Invitrogen, Carlsbad, CA, USA) is the most widely used probe to this kind of assays due to its high signal-to-noise ratio. The fluorescence intensity is plotted as a function of temperature generating a sigmoidal curve (transition curve). The temperature of melting corresponds to the transition curve inflection point (Niesen, Berglund et al. 2007).

DSF meaurements were taken in a 96-well plate where small molecules like buffers and salts, or compounds like substrates, inhibitors or co-factors (that are thought to interact with the protein of interest) are screened (Vedadi, Niesen et al. 2006). Some wells were dedicated to control, where the initial protein buffer was tested to check interference with the dye or for reference, where the protein melting temperature was measured with respect to initial buffer conditions.

The experiment was settled up as a compound screen where 5 µL of protein solution within a concentration range from 1 to 10 mg/mL were mixed with 7.5 µL 30× SYPRO orange dye, previously prepared in the interest protein buffer, and 12.5 µL of the compound to be screened two times concentrated. After plate sealing with an optical foil seal the plate was centrifuged tor 1 min at room temperature and placed in the real-time PCR BioRad iCycler5 instrument (Niesen, Berglund et al. 2007). The temperature scan was from 25 to 85 ºC, at 1 ºC/min. Measurement accuracy may depend on optimization of parameters such as protein and dye concentration.

5.2. Dynamic light scattering

DLS was routinely used as a tool to evaluate monodispersity in pure protein solutions and to estimate the size of molecules and particles in solution in a fast and accurate manner. DLS is a technique used to determine the hydrodynamic radius of proteins from the Stokes-Einstein equation, $R_h = k_B T/6\pi\eta D_T$ (where $\underline{k_B}$ denotes the Boltzmann constant, $\underline{T}$ the absolute temperature, $\underline{\eta}$ is the viscosity of the solvent in the same temperature, $\underline{D_T}$ the diffusion coefficient and $\underline{R_h}$ the experimental measure of the maximal radius of a hydrated object).

DLS measurements were performed in protein solutions under different conditions using a Malvern Instruments (Worcestershire, UK) Zetasizer Nano. Samples were measured in a 1 cm path-length quartz cuvette and three measurements were performed on each sample with an average of ten scans for measurement.

5.3. Circular dichroism spectroscopy

Circular dichroism (CD) spectroscopy is a well established technique to investigate protein folding properties and to determine protein stability (Greenfield 2006).

The optimal buffer to perform CD measurements must not contain optically active compounds and be as transparent as possible. The protein solution must be concentrated to 0.1-0.5 mg/mL and filtered through a 0.22 µm pore-sized filter previous to the measurement to reduce light scattering. Spectra acquisition were performed at 25 ºC in a 1 mm path-length cuvette over a wavelength range of 250-185 nm (this range was changed in cases where non-optimal buffers had to be used). Six replicate spectra were collected for each sample and averaged.

Protein thermal melt curves can be also acquired with CD spectroscopy to determine the protein temperature of melting. A complete spectrum, within the range of 250-185 nm, was measured to check for good sample stability. To obtain the protein melting curve the wavelength was fixed at 218 nm (because at this wavelength there is a CD signal contribution from both α-helices and β-sheets secondary structure elements) and CD signal was registered from 20 to 99 ºC applying a thermal gradient speed of 1 ºC/min.

5.4. Nuclear magnetic resonance spectroscopy

Multi-dimensional Nuclear Magnetic Resonance (NMR) spectroscopy can be applied to the study of the structure and dynamics of biological polymers.

NMR is a phenomenon which occurs when the nuclei of certain atoms are immersed in a static magnetic field and exposed to a second oscillating magnetic field. Some nuclei experience this phenomenon depending on whether they possess a property called spin. These nuclei can absorb or emit radiofrequency energy at a specific frequency, depending on its local chemical environment within the structure of the molecule. The frequency of the radiofrequency energy emitted by each nucleus is reported in units of 'chemical shift', which is proportional to Hz (the most familiar unit of frequency). NMR spectroscopy uses this phenomenon to study physical, chemical and biological properties of matter when it interacts with the electromagnetic radiation. One-dimensional NMR spectroscopy is routinely used to study chemical structure while two-dimensional spectra are used to determine the structure of more complex molecules.

To collect a spectrum, magnetization is transferred into the sample and between nuclei using pulses of electromagnetic energy (radiofrequency); the process is described with so-called pulse sequences. When magnetization is transferred through the chemical bonds the information is used to assign the different chemical shifts to a specific nucleus.

NMR spectra are acquired on aqueous samples of highly pure protein solutions in a thin-walled glass tube. Typically, 300-600 μL of sample at a protein concentration of 0.1-3 mM are needed. The protein source can be either natural or recombinant but in the latter case isotopic labelling is possible and highly recommended. NMR spectroscopy of

proteins from natural sources is restricted to utilizing NMR solely on protons because the most abundant isotopes of C, O and N have no net nuclear spin, which is the physical property exploited this technique. However, the less common isotopes, ^{15}N and ^{13}C, have net nuclear spin that makes them suitable for nuclear magnetic resonance and, therefore, labelling of the target proteins with these compounds opens up possibilities for doing more advanced experiments which also detect or use these nuclei. As NMR spectra are measured for compounds dissolved in a solvent and signals will be observed for it, spectra are recorded in a deuterated solvent. However, deuteration is not complete and signals for the residual protons can be observed.

To acquire the proteins NMR spectra several parameters were settled which include the width of the spectrum, number of data points in the spectrum and the receiver gain. The measurements were performed in 600 or 800 MHz Bruker Digital Avance NMR spectrometer (San Francisco, CA, USA). A first experiment consisted on measuring a one-dimensional [^{1}H]-NMR spectrum for each sample where the number of signals corresponded to the different types of protons, the chemical shifts were related with the protons type of environment and the dynamics of proton environments was given by the shape of the peaks.

In NMR experiments, ideally, each nucleus in the molecule experiences a distinct chemical environment having a distinct and measurable chemical shift. However, in large molecules, such as proteins, the number of resonances is typically too high and a one-dimensional spectrum inevitably has coincidental overlaps. Therefore, multi-dimensional experiments were performed to correlate the frequencies of distinct nuclei. The additional dimensions decrease the chance of overlap and have larger information content since they correlate signals from nuclei within a specific part of the molecule.

The HSQC experiment is used frequently in NMR spectroscopy applied to the study of proteins. The resulting spectrum is two-dimensional with one axis for ^{1}H and the other for a heteronucleus (an atomic nucleus other than a proton), most often ^{13}C or ^{15}N. The spectrum contains a peak for each unique proton attached to the heteronucleus being considered. Thus, if the chemical shift of a specific proton is known the chemical shift of the coupled heteronucleus can be determined and vice versa. In the ^{15}N-HSQC one signal is expected for each amino acid residue (except for proline which has no amide-

hydrogen due to the cyclic nature of its backbone). Tryptophane and certain other residues with N-containing side chains give rise to additional signals. The ^{15}N-HSQC is referred to as the fingerprint of a protein because each protein has a unique pattern of signal positions. The HSQC experiment is also useful for detecting interactions with ligands, such as other proteins or drugs. By comparing HSQC spectra of free and ligand-bound protein it is possible to find changes in the chemical shifts of the peaks, which are likely to occur at the binding interface.

Two-dimensional HSQC spectra were measured in ^{15}N-labelled protein samples and qualitatively analysed. For a folded protein the peaks were usually sharp, well dispersed and most of the individual peaks could be distinguished. HSQC, being a relatively quick experiment, was useful to determine if relatively small proteins (< 200 amino acids) were folded and to screen candidates for structure determination. When the number of peaks did not match the number of protein residues (considering the special case of prolines and amide-containing amino acids) it indicated that the sample was heterogeneous, disordered, unfolded or that it could exist in multiple conformations.

6. Protein crystallization

6.1. Crystallization methods

Growing of high-quality single crystals is the basis of X-ray crystal structure determination and it is considered this process major bottleneck.

The crystallization of macromolecules is a multi-parametric process that consists on bringing the initial highly concentrated solution into a supersaturation state that will force the macromolecules into the solid state, the crystal. There are a number of well established methods for proteins and nucleic acids crystallization but, by far, vapour-diffusion is the most widely used. This method is based on water (or other volatile agent) diffusion between a drop of 'sitting' or 'hanging' macromolecule solution mixed with the appropriate precipitating agent and a larger reservoir solution inside a closed container.

Initial crystallization screenings were performed using the sitting-drop vapour-diffusion method and commercial sparse-matrix screens to determine preliminary crystallization conditions (Cudney, Patel et al. 1994). Crystallization trials were settled either manually

in 24-well plates (microliter-scale drops were settled) or using the crystallization robots Cartesian Honeybee (Hamilton, Reno, NV, U.S.A.) or Crystal Phoenix (ARI, Sunnyvale, CA, USA), where 96-well plates were used and nanoliter-scale drops were settled. Robots were used on the Automated Crystallography Platform of the Barcelona Scientific Park. When no promising leads were found after the initial crystallization trials several rounds of action were followed, such as alter crystallization temperature, sample concentration, drop volume or by adding small organic compounds to the drops. At this stage, if the protein did not crystallized the strategy was to go back to the protein solution and perform an optimum solubility screen to determine the buffer conditions where the protein is more homogeneous (Jancarik, Pufan et al. 2004), change the established expression and purification protocol or, inclusively, change the protein construct.

Other crystallization methods were only tested in the crystal optimization stage. The microbatch method involves the mixing of drops of a protein solution with precipitants under a layer of hydrophobic oil, which prevents evaporation of the drops (Chayen 1997).

In the microfluidic crystallization method nanoliter-scale drops containing protein, precipitant and additive solutions in variable ratios, are formed in the flow of immiscible fluids inside microfluidic channels. The solutions are loaded in syringes connected into converging channels (the flow of the solutions is established by driving syringes with syringes pumps) and after the last drop is settled the syringes are disconnected and the flow is stopped (Hansen, Classen et al. 2006).

Protein crystallization in capillars by counter-diffusion is settled in a box with a capillary holder and cover. A layer of buffered agarose gel is set at the bottom of the box, the capillaries, which are used as protein chambers, are inserted through the hole of the capillary holder at a given depth in the gel and finally the salt solution is poured on top of the gel. In the counter-diffusion technique mass transport is controlled only by diffusion and sedimentation and convection are avoided by working in small dimension gelled systems (Ng, Gavira et al. 2003).

6.2. Crystal growth, manipulation and soaking

Generally, crystals obtained on initial crystallization screenings are of poor quality and unsuitable for optimal data collection. When one or more crystallization 'hits' were found follow-up experiments were done that consisted in refining the crystallization variables and introduce some new ones in order to produce diffraction-quality single crystals. Crystal growth optimization steps consisted in varying the concentrations of the crystallization condition components, slight pH changes, using additives, switching to a buffer or precipitant with similar composition or using different crystallization methods. Moreover, crystals were subjected to post-crystallization methods such as annealing, dehydration or cross-linking (Heras and Martin 2005). These techniques provided, in some cases, unquestionable improvements on crystal quality.

Some crystals were modified after growing in order to incorporate some ligands (to test protein-ligand binding) or heavy-atoms (for phasing purposes). To introduce a new compound into the crystal lattices, crystals were immersed (soak) into harvesting solutions containing the desired ligand or heavy-atom compound.

In the present work cryogenic data collection from crystals mounted in nylon-loops was used in all cases and crystals were cryoprotected before flash-cooling in liquid nitrogen. In this process, the crystals were immersed and transferred into crystallization solution drops with increasing concentrations of cryoprotectant, typically glycerol.

6.3. Single crystal X-ray diffraction

6.3.1. Data collection, processing and scaling

X-ray data collection has become much easier and fast after the important technological advances of the recent years. In the last experimental step towards protein crystal structure determination the quality and accuracy of the measured intensities is of critical importance to obtain a precise macromolecular atomic model (Dauter 1999).

Data acquisition was generally performed in a synchrotron light-source on one of the ESRF (Grenoble, France) beam lines dedicated to protein crystallography. Data collection was sometimes performed in an automatic manner by the beam line controlling software, decisions being made at the level of crystal exposure time and rotation range per frame, crystal-to-detector distance and number of images to collect to

assure data completion (parameter that greatly depends on the radiation damage affecting the crystal), which relates to crystal lattice geometry.

The MOSFLM data processing software (Leslie 1992) was used for initial inspection of the first two collected orthogonal images to determine crystal quality and effective resolution limit. After indexing and determination of the unit cell dimensions and the possible space group, a data collection strategy was established to maximize both the resolution and completeness of the dataset, which was then indexed and integrated with the DENZO program from the HKL2000 software package (Otwinowski 1997). During integration the position of each Bragg reflection in the image is predicted, after subtracting the X-ray background, and the intensity error estimated. The integration profiles were analysed to match the size and shape of diffraction spots.

Data quality was judged after merging the results at the scaling step that was performed with Scalepack from the HKL2000 software package (Otwinowski 1997). Statistics on data quality and completeness were assessed and the global R_{merge} factor, which gives the average ratio of the spread of intensities of the multiple measured symmetry-equivalent reflections to the estimated value of the reflection intensity, analysed. In the step of merging equivalent intensities some outliers, reflections with wrongly measured intensities or that did not agree with their equivalents, were identified and rejected.

6.3.2. Structural data solving and refining

The analysis of X-ray diffraction data from protein crystals and their interpretation in terms of a model of the macromolecule constitute a multistep process that consists in structure solution (scaling, heavy-atom location and phasing), density modification, model building and refinement. Some of these steps will be briefly treated in the text below in the case they were used to solve the protein structures described in this thesis.

In order to visualize the structure of the protein of interest some phase information must be obtained (phase problem) and the structure can be solved. The three main methods for solving the phase problem are:

i) introducing highly dispersive atoms, Multiple Isomorphous Replacement (MIR);

ii) introducing anomalous dispersive atoms, Multiple-wavelength Anomalous Dispersion (MAD);

iii) by Molecular Replacement (MR), using a previously determined structural model of a homologous protein.

Initial phases can be calculated with any of these methods and together with the experimental amplitudes the electron density can also be calculated and a structural model built.

6.3.2.1. Molecular replacement method

When a molecular model is available for a protein that is homologous to the protein of interest the phase problem can be solved by the MR method (Rossmann 1962). In this method the protein of interest is considered to have the homologous protein structure, having into account that homologous proteins share highly similar protein fold. In MR the structure of the homologous protein is transferred from its crystal packing to the crystal of the protein under investigation. The known molecule has to be positioned in the unit cell of the crystal of the protein with unknown structure by the determination of its correct orientation (rotation) and accurate positioning (translation), which are calculated from the so called rotation and translation functions. The results of these functions are analysed in terms of the correlation coefficient between the experimental Patterson function and the Patterson function calculated with the structural homologue. After the correct rotation and translation of the structural homologue an electron density map is calculated combining the experimental amplitudes of the structure factors with the phases obtained for the homologue structure.

6.3.2.2. Multiple isomorphous replacement method

In the MIR method X-ray scatterers of high atomic number (heavy-atoms) are introduced into specific sites of the crystal unit cell, either by soaking the heavy-atom compound into the crystal or by co-crystallising it together with the protein. Introducing atoms by soaking into protein crystals is one of the technique bottlenecks because heavy-atoms can disrupt the macromolecule structure or crystal packing of the native protein. Crystals must be isomorphous in respect to the native protein crystals. The isomorphous replacement is achieved by diffusion of the heavy-atom complexes through the protein crystal solvent channels where some amino acidic side chains can

coordinate with them. In some cases, as for metalloproteins, endogenous metal atoms can be replaced by these heavier metal atoms. Given that these have a higher scattering power in comparison with C, H, N, O and S, they can introduce differences in the diffraction pattern intensity of the derivative crystal in relation to that of the native. The intensity differences measured between both diffraction spectra are used to calculate a Patterson map (map of vectors of the heavy-atoms relative position) from which their position in the unit cell can be determined. In these conditions structure factors including the phases can now be obtained and, consequently, also the electron density map.

To perform a MIR experiment, native protein crystals were soaked in solutions of diverse heavy-atom compounds and an initial control of the crystal isomorphism was analysed by comparing the native and derivative crystals unit cell dimensions. Data were collected for both crystals and the Patterson function applied to determine the heavy-atoms coordinates. These positions were further refined, the protein phase angles determined and finally the electron density map was calculated.

6.3.2.3. Multiple-wavelength Anomalous Dispersion method

In the MAD method changes in the diffraction intensities can be induced by taking advantage of the atoms physical properties (Hendrickson, Smith et al. 1988). For heavy-atoms, the energy of an absorbed X-ray photon promotes the transition from a ground state orbital to an excited level and the corresponding electronic transition leads to anomalous scattering. The characteristic energy for such a transition, called absorption edge, depend on atomic orbital levels and are element specific. In this way, if the wavelength of the X-ray radiation is near the heavy-atom absorption edge, a fraction of the radiation is absorbed by the heavy-atom and reemitted with different phase. Information about the phase of the scattered X-rays can be derived from the anomalous scattering. The majority of the heavy-atoms used in protein crystallography have absorption peaks at energies (wavelengths) that could be easily achieved with synchrotron radiation and the success of the experiment greatly rely on the ordered presence of these special atoms on the protein crystal.

To perform the MAD experiment data were collected from a heavy-atom protein derivative crystal at two or more wavelengths. The typically used radiation energies are:

i) a radiation energy that maximizes the absorption component of the anomalous scatterer,
ii) a radiation energy that minimizes the dispersion component of the anomalous scatterer,
iii) a radiation energy remote with respect to the others.

The different data sets were combined and the differences between them analysed in order to calculate the amplitudes distribution and phases generated by the heavy-atoms. These phases were used as an initial approximation to the global phases allowing the calculation of the electron density for all the protein.

In both MIR and MAD methods, heavy-atoms substructures compatible with the difference Patterson function were obtained by automated inspection of difference Patterson functions performed by the SHELXD program (Schneider and Sheldrick 2002). Each potential substructure was then individually evaluated. The substructure was generally refined and used to identify further anomalously scattering atoms and the completed substructure was then used to calculate phases for the entire structure using SOLVE (Terwilliger and Berendzen 1999) or HKL2MAP (Pape 2004) software packages. The resultant electron density maps were visually examined to determine if it has the expected macromolecule features and if the process was successful.

6.3.3. Refining and building programs

The electron density function was calculated after solving the phase problem and its values were graphically represented in the so called electron density map. The interpretable map showed positive electron density regions where an initial polypeptidic skeleton was traced using TURBO/FRODO (Roussel and Cambillau 1989) or COOT (Emsley and Cowtan 2004) programs, which allowed the building of the Cα main-chain and also of the amino acids side chains. In cases where high resolution data was available the initial interpretation of the electron density map was automatically performed with the ARP/wARP 6.0 software suite (Perrakis, Morris et al. 1999). In model building some mathematical tools were introduced to facilitate the convergence between the structural model and the experimental observations provided by the diffraction spectra. The structural model is defined by the atomic coordinates (x, y, z) and a temperature factor (B_{factor}) for each atom, which is a measure of the atom thermal mobility around its equilibrium position. After initial model building at a graphics

workstation structural refinement was undertaken. At this stage, the structural parameters were adjusted in order to optimize the agreement between observations and prediction. The optimization process was performed with the program REFMAC (Murshudov, Vagin et al. 1997) and was achieved by the introduction of stereochemical restraints into the model given that amino acid structures are known with high accuracy. The final model was obtained after several refinement cycles combined with model building until little or no further improvements were obtained.

6.3.4. Structural model validation

Model validation consists in examining the data for incorrectness or bias and in confirming the suitability of the model (Dodson, Davies et al. 1998).

The final model structural information, obtained after the refinement process and that fits well the experimental observations, was written in a *pdb* file. The model consistency was determined checking if the model had a reasonable stereochemistry by analysing model bonds size and angle and its rmsd in relation to defined geometrical and thermal parameters. The planar structure of the peptide group was checked by analysing the conformation of N-Cα and Cα-C bonds of the peptidic bond and the peptidic bond torsions (Φ, *phi*, and ψ, *psi*) that are energetically limited to certain allowed regions defined in the *Ramachandran* plot.

The most accepted function to evaluate the quality of the model refinement is the crystallographic R_{factor} which measures the differences between the observed spectra and the one calculated from the obtained structural model, being a measure of agreement between model and data. Given this, the lower the factor the better the model fits the experimental data. Another parameter used to evaluate data quality is cross-validation performed by the random selection of some data at the beginning of the experiment (test set) that is not used during refinement, for which a R_{factor} is calculated. A crystallographic R_{free} is calculated that should not separate from the R_{factor} value when more parameters are added into the model.

The validated structural models were finally deposited in the Protein Data Bank database where they were, once again, revised and validated. Finally, a unique accession

code was attributed to each deposited model coordinates, which became accessible to the scientific community.

CONCLUSIONS

Starting as a sideshow of early microbiology, mycoplasmas have become central to modern computational and theoretical biology and the understanding of infectious disease. To look ahead, I am confident that the first decade of the 21st century will lead to a complete computer model of mycoplasma cell function. As the biology of the 21st century folds, I suspect that the minimal cell concept as embodied on the mycoplasma will continue to be central to the understanding of life.

Harold J. MOROWITZ
Robinson Professor, Krasnow Institute for Advanced Studies
in Molecular Biology and Pathogeneicity of Mycoplasmas, 2002

CONCLUSIONS

In a certain way, a nearly complete understanding of mycoplasmas was accomplished in the first decade of the 21st century as 'prophesized' by Morowitz, but such a simple organism was discovered to be much more complex than expected.

The discrete but accurate and unique contributions of this thesis work to a more complete understanding of mycoplasmas are now listed in form of conclusions:

Chapter II. MG438, an orphan type I R-M S subunit

The first crystal structure of a type I R-M S subunit was solved to 2.3 Å resolution by the MAD method using data from a SeMet-MG438 crystal.

The MG438 protein structure presents an overall cyclic topology with an intra-subunit two-fold symmetry axis that relates the two half parts of the molecule by a pure rotational symmetry of exactly 180º, each half containing a globular domain and a conserved helix. Moreover, this intra-subunit symmetry reflects the partial sequence repetition observed for the molecule primary sequence, where the two half parts share 23 % amino acidic sequence identity, suggesting that an internal duplication event might have occurred.

MG438 intra-subunit two-fold symmetry axis is highly polar defining two molecular faces: the **S** face, which is characterized by two patches of exposed hydrophobic residues and positively charges residues, and a **Z** face with the CRs at a lower level and a polar character, which constitutes the DNA binding face.

A very close structural homologue was reported almost simultaneously to the MG438 structure determination. This structure was from the type I R-M S subunit from the archae *Methanococcus jannaschii*, S.Mja, which share low sequence identity (16 %) with MG438. The major difference between them comes from the different disposition of the TRDs with respect to the CRs.

In MG438 large exposed hydrophobic surfaces are restricted to the **S** face, which is likely reflecting the subunit solitary character, while in S.Mja, which belongs to a complete type I R-M system, exposed hydrophobic residues exist throughout the whole molecule surface.

The modelling of a MTase ternary complex was performed starting from the structure of the type II N6-adenine DNA *Taq*I MTase structure, its DNA target sequences and the MG438 structure. The MG438 residues proposed, by the MTase model analysis, to interact with the half-sites of the non-palindromic bipartite DNA target site belong to the TRDs, in accordance with previous mutagenic experiments performed on the EcoKI type I R-M system.

The MTase model also suggests that type I MTase complexes have two-fold rotational symmetry. The two M subunits can arrange on either side of the S subunit gaining access to the methylation targets via base flipping, similarly to the mechanism followed by type II MTases.

The analysis of the first structure from a type I R-M S subunit provides a structural framework consistent with prior biochemical information available, corroborating most of the anticipated structural features but giving few clues about the functional role of this orphan protein.

Chapter III. MG200, a terminal organelle motility protein

The MG200 protein produced behaves as a tetramer in solution and tends to form aggregates. Several biophysical techniques applied to the study of the protein in solution revealed that it was highly heterogeneous, which could have prevented protein crystallization.

Three, out of four, MG200 domains were successfully over-expressed and purified. For the fourth domain, which consists in an acidic and proline-rich region predicted to have a random-coil structure, over-expression was not observed.

MG200-DnaJ$_{1-74}$ is a highly soluble domain presenting an all α-helical fold similar to the one described for its protein homologues, the J domains.
MG200 C-terminal region is a poorly-soluble globular domain for which crystals are already available. However, these crystals still need to be improved to proceed with X-ray diffraction studies.

MG200-EAGRb$_{124-207}$ is a highly soluble domain. High quality-diffraction crystals were obtained, allowing its crystal structure determination. This constitutes the first structural piece, at almost atomic resolution, from any component of the terminal organelle.

MG200-EAGRb$_{124-207}$ presents an essentially new fold with some peculiar features, such as the presence of an intra-domain symmetry axis that relates two hairpins (the domain wings), which seem to correspond to a sequence duplication that is well conserved among the EAGRb sequences available.

The MG200-EAGRb$_{124-207}$ structure forms a dimer that is stabilized by a hydrophobic core that extends throughout the contacting interface.

EAGRb sequences show a high degree of conservation, mainly for the glycine and aromatic residues from the wings. The EAGR boxes from orthologue proteins present the highest degree of similarity, indicating a relatively recent common origin. For proteins having multiple EAGR boxes there is a high degree of similarity between their own boxes, suggesting the occurrence of internal duplication events.

Among all the EAGRb sequences there are only three fully conserved residues, two glycines and one tryptophane. The glycine residues, Gly179 and Gly 191, are strategically located at the entrance of the two β-hairpins that define the domain wings and are accurately related by the intra-domain symmetry axis.

MG200-EAGRb$_{124-207}$ domain plasticity, together with the presence and organization of the intra-domain and inter-subunits symmetry axes, results in the unbalance of interactions, mainly for the highly conserved aromatic residues, suggesting that MG200-EAGRb$_{124-207}$ (in particular the wing 1) should have a high tendency to participate in protein-protein interactions.

A quaternary structure model with 222 (D2) symmetry was proposed for the MG200 protein which is in accordance with the high tendency of the protein to oligomerize and also with preliminary EM studies.

REFERENCES

REFERENCES

A

Alm, R. A. and T. J. Trust (1999). "Analysis of the genetic diversity of Helicobacter pylori: the tale of two genomes." J Mol Med 77(12): 834-46.

Argos, P. (1983). "Evidence for a repeating domain in type I restriction enzymes." Embo J 4(1351-1355).

B

Bailey, C. C., R. Younkins, et al. (1996). "Characterization of genes encoding topoisomerase IV of Mycoplasma genitalium." Gene 168(1): 77-80.

Balish, M. F. (2006). "Subcellular structures of mycoplasmas." Front Biosci 11: 2017-27.

Balish, M. F. and D. C. Krause (2006). "Mycoplasmas: a distinct cytoskeleton for wall-less bacteria." J Mol Microbiol Biotechnol 11(3-5): 244-55.

Beck, B. D., P. G. Arscott, et al. (1978). "Novel properties of bacterial elongation factor Tu." Proc Natl Acad Sci U S A 75(3): 1250-4.

Biberfeld, G. and P. Biberfeld (1970). "Ultrastructural features of Mycoplasma pneumoniae." J Bacteriol 102(3): 855-61.

Bickle, T. A. (1987). DNA restriction and modification systems. Escherichia coli and Salmonella typhimurium: Cellular and Molecular Biology. A. S. f. Microbiology. Washington, D.C.: 692-696.

Bickle, T. A. and D. H. Kruger (1993). "Biology of DNA restriction." Microbiol Rev 57(2): 434-50.

Bjorkholm, B. M., J. L. Guruge, et al. (2002). "Colonization of germ-free transgenic mice with genotyped Helicobacter pylori strains from a case-control study of gastric cancer reveals a correlation between host responses and HsdS components of type I restriction-modification systems." J Biol Chem 277(37): 34191-7.

Boggon, T. J. and L. Shapiro (2000). "Screening for phasing atoms in protein crystallography." Structure 8(7): R143-9.

Bradford, M. M. (1976). "A rapid and sensitive method for the quantitation of microgram quantities of protein utilizing the principle of protein-dye binding." Anal Biochem 72: 248-54.

Bredt, W. (1973). "Motility of Mycoplasmas." Ann. N. Y. Acad. Sci. 225: 246-250.

Bullas, L. R., C. Colson, et al. (1980). "Deoxyribonucleic acid restriction and modification systems in Salmonella: chromosomally located systems of different serotypes." J Bacteriol 141(1): 275-92.

Burgos, R., O. Q. Pich, et al. (2007). "Functional analysis of the Mycoplasma genitalium MG312 protein reveals a specific requirement of the MG312 N-terminal domain for gliding motility." J Bacteriol 189(19): 7014-23.

Burgos, R., O. Q. Pich, et al. (2008). "Deletion of the Mycoplasma genitalium MG_217 gene modifies cell gliding behaviour by altering terminal organelle curvature." Mol Microbiol 69(4): 1029-40.

C

Calisto, B. M., O. Q. Pich, et al. (2005). "Crystal structure of a putative type I restriction-modification S subunit from Mycoplasma genitalium." J Mol Biol 351(4): 749-62.

Chandonia, J. M., S. H. Kim, et al. (2006). "Target selection and deselection at the Berkeley Structural Genomics Center." Proteins 62(2): 356-70.

Chayen, N. E. (1997). "The role of oil in macromolecular crystallization." Structure 5(10): 1269-74.

Chin, V., V. Valinluck, et al. (2004). "KpnBI is the prototype of a new family (IE) of bacterial type I restriction-modification system." Nucleic Acids Res 32(18): e138.

Citti, C. and R. Rosengarten (1997). "Mycoplasma genetic variation and its implication for pathogenesis." Wien Klin Wochenschr 109(14-15): 562-8.

Cloward, J. M. and D. C. Krause (2009). "Mycoplasma pneumoniae J-domain protein required for terminal organelle function." Mol Microbiol 71(5): 1296-307.

Collaborative Computational Project, N. (1994). "The CCP4 suite: programs for protein crystallography." Acta Crystallogr D 50: 760-763.

Crick, F. H. C. (1953). "The packing of alpha-helices: simple coiled-coils." Acta Crystallog. 6: 689-697.

Cudney, R., S. Patel, et al. (1994). "Screening and optimization strategies for macromolecular crystal growth." Acta Crystallogr D Biol Crystallogr 50(Pt 4): 414-23.

D

Dandekar, T., M. Huynen, et al. (2000). "Re-annotating the Mycoplasma pneumoniae genome sequence: adding value, function and reading frames." Nucleic Acids Res 28(17): 3278-88.

Daniel, A. S., F. V. Fuller-Pace, et al. (1988). "Distribution and diversity of hsd genes in Escherichia coli and other enteric bacteria." J Bacteriol 170(4): 1775-82.

Dauter, Z. (1999). "Data-collection strategies." Acta Crystallogr D Biol Crystallogr 55(Pt 10): 1703-17.

Dauter, Z., M. Dauter, et al. (2000). "Novel approach to phasing proteins: derivatization by short cryo-soaking with halides." Acta Crystallogr D Biol Crystallogr 56(Pt 2): 232-7.

DeLano, W. L. (2002). "The PyMOL Molecular Graphics System." from http://www.pymol.org.

Dhandayuthapani, S., W. G. Rasmussen, et al. (1999). "Disruption of gene mg218 of Mycoplasma genitalium through homologous recombination leads to an adherence-deficient phenotype." Proc Natl Acad Sci U S A 96(9): 5227-32.

Dodson, E. J., G. J. Davies, et al. (1998). "Validation tools: can they indicate the information content of macromolecular crystal structures?" Structure 6(6): 685-90.

Doublie, S. (1997). "Preparation of selenomethionyl proteins for phase determination." Methods Enzymol 276: 523-30.

Dryden, D. T., L. P. Cooper, et al. (1997). "The in vitro assembly of the EcoKI type I DNA restriction/modification enzyme and its in vivo implications." Biochemistry 36(5): 1065-76.

Dryden, D. T., S. S. Sturrock, et al. (1995). "Structural modelling of a type I DNA methyltransferase." Nat Struct Biol 2(8): 632-5.

Dybvig, K. (1993). "DNA rearrangements and phenotypic switching in prokaryotes." Mol Microbiol 10(3): 465-71.

Dybvig, K. and H. Yu (1994). "Regulation of a restriction and modification system via DNA inversion in Mycoplasma pulmonis." Mol Microbiol 12(4): 547-60.

E

Emsley, P. and K. Cowtan (2004). "Coot: model-building tools for molecular graphics." Acta Crystallogr D Biol Crystallogr 60(Pt 12 Pt 1): 2126-32.

Evans, G. and R. F. Pettifer (2001). "CHOOCH: a program for deriving anomalous-scattering factors from X-ray fluorescence spectra." J. Appl. Cryst. 34: 82-86.

F

Fadouloglou, V. E., M. Kokkinidis, et al. (2008). "Determination of protein oligomerization state: two approaches based on glutaraldehyde crosslinking." Anal Biochem 373(2): 404-6.

Feldner, J., U. Gobel, et al. (1982). "Mycoplasma pneumoniae adhesin localized to tip structure by monoclonal antibody." Nature 298(5876): 765-7.

Finn, R. D., J. Mistry, et al. (2010). "The Pfam protein families database." Nucleic Acids Res 38(Database issue): D211-22.

Flaherty, K. M., D. B. McKay, et al. (1991). "Similarity of the three-dimensional structures of actin and the ATPase fragment of a 70-kDa heat shock cognate protein." Proc Natl Acad Sci U S A 88(11): 5041-5.

Fraser, C. M., J. D. Gocayne, et al. (1995). "The minimal gene complement of Mycoplasma genitalium." Science 270(5235): 397-403.

Fuller-Pace, F. V. and N. E. Murray (1986). "Two DNA recognition domains of the specificity polypeptides of a family of type I restriction enzymes." Proc Natl Acad Sci U S A 83(24): 9368-72.

G

Gann, A. A., A. J. Campbell, et al. (1987). "Reassortment of DNA recognition domains and the evolution of new specificities." Mol Microbiol 1(1): 13-22.

Gibson, D. G., G. A. Benders, et al. (2008). "Complete chemical synthesis, assembly, and cloning of a Mycoplasma genitalium genome." Science 319(5867): 1215-20.

Gibson, D. G., J. I. Glass, et al. (2010). "Creation of a Bacterial Cell Controlled by a Chemically Synthesized Genome." Science Published online.

Glass, J. I., N. Assad-Garcia, et al. (2006). "Essential genes of a minimal bacterium." Proc Natl Acad Sci U S A 103(2): 425-30.

Glover, S. W. and C. Colson (1969). "Genetics of host-controlled restriction and modification in Escherichia coli." Genet Res 13(2): 227-40.

Goedecke, K., M. Pignot, et al. (2001). "Structure of the N6-adenine DNA methyltransferase M.TaqI in complex with DNA and a cofactor analog." Nat Struct Biol 8(2): 121-5.

Greenfield, N. J. (2006). "Using circular dichroism spectra to estimate protein secondary structure." Nat Protoc 1(6): 2876-90.

Guell, M., V. van Noort, et al. (2009). "Transcriptome complexity in a genome-reduced bacterium." Science 326(5957): 1268-71.

H

Han, J. H., S. Batey, et al. (2007). "The folding and evolution of multidomain proteins." Nat Rev Mol Cell Biol 8(4): 319-30.

Hansen, C. L., S. Classen, et al. (2006). "A microfluidic device for kinetic optimization of protein crystallization and in situ structure determination." J Am Chem Soc 128(10): 3142-3.

Hasselbring, B. M., J. L. Jordan, et al. (2005). "Mutant analysis reveals a specific requirement for protein P30 in Mycoplasma pneumoniae gliding motility." J Bacteriol 187(18): 6281-9.

Hasselbring, B. M. and D. C. Krause (2007). "Cytoskeletal protein P41 is required to anchor the terminal organelle of the wall-less prokaryote Mycoplasma pneumoniae." Mol Microbiol 63(1): 44-53.

Hasselbring, B. M., C. A. Page, et al. (2006). "Transposon mutagenesis identifies genes associated with Mycoplasma pneumoniae gliding motility." J Bacteriol 188(17): 6335-45.

Hatchel, J. M. and M. F. Balish (2008). "Attachment organelle ultrastructure correlates with phylogeny, not gliding motility properties, in Mycoplasma pneumoniae relatives." Microbiology 154(Pt 1): 286-95.

Hatchel, J. M., R. S. Balish, et al. (2006). "Ultrastructure and gliding motility of Mycoplasma amphoriforme, a possible human respiratory pathogen." Microbiology 152(Pt 7): 2181-9.

Hegermann, J., R. Herrmann, et al. (2002). "Cytoskeletal elements in the bacterium Mycoplasma pneumoniae." Naturwissenschaften 89(10): 453-8.

Henderson, G. P. and G. J. Jensen (2006). "Three-dimensional structure of Mycoplasma pneumoniae's attachment organelle and a model for its role in gliding motility." Mol Microbiol 60(2): 376-85.

Hendrickson, W. A., J. L. Smith, et al. (1988). "Crystallographic structure analysis of lamprey hemoglobin from anomalous dispersion of synchrotron radiation." Proteins 4(2): 77-88.

Hennessy, F., M. E. Cheetham, et al. (2000). "Analysis of the levels of conservation of the J domain among the various types of DnaJ-like proteins." Cell Stress Chaperones 5(4): 347-58.

Heras, B. and J. L. Martin (2005). "Post-crystallization treatments for improving diffraction quality of protein crystals." Acta Crystallogr D Biol Crystallogr 61(Pt 9): 1173-80.

Herbelin, A., E. Ruuth, et al. (1994). "Mycoplasma arginini TUH-14 membrane lipoproteins induce production of interleukin-1, interleukin-6, and tumor necrosis factor alpha by human monocytes." Infect Immun 62(10): 4690-4.

Herrmann, R. and B. Reiner (1998). "Mycoplasma pneumoniae and Mycoplasma genitalium: a comparison of two closely related bacterial species." Curr Opin Microbiol 1(5): 572-9.

Highlander, S. K. and O. Garza (1996). "The restriction-modification system of Pasteurella haemolytica is a member of a new family of type I enzymes." Gene 178(1-2): 89-96.

Himmelreich, R., H. Hilbert, et al. (1996). "Complete sequence analysis of the genome of the bacterium Mycoplasma pneumoniae." Nucleic Acids Res 24(22): 4420-49.

Himmelreich, R., H. Plagens, et al. (1997). "Comparative analysis of the genomes of the bacteria Mycoplasma pneumoniae and Mycoplasma genitalium." Nucleic Acids Res 25(4): 701-12.

Holm, L., S. Kaariainen, et al. (2008). "Searching protein structure databases with DaliLite v.3." Bioinformatics 24(23): 2780-1.

Holm, L. and C. Sander (1993). "Protein structure comparison by alignment of distance matrices." J Mol Biol 233(1): 123-38.

Holt, M. R. and A. Koffer (2001). "Cell motility: proline-rich proteins promote protrusions." Trends Cell Biol 11(1): 38-46.

Hopfner, K. P., L. Craig, et al. (2002). "The Rad50 zinc-hook is a structure joining Mre11 complexes in DNA recombination and repair." Nature 418(6897): 562-6.

Hutchison, C. A., S. N. Peterson, et al. (1999). "Global transposon mutagenesis and a minimal Mycoplasma genome." Science 286(5447): 2165-9.

Hutchison, C. A., 3rd, S. Phillips, et al. (1978). "Mutagenesis at a specific position in a DNA sequence." J Biol Chem 253(18): 6551-60.

I

Ives, C. L., P. D. Nathan, et al. (1992). "Regulation of the BamHI restriction-modification system by a small intergenic open reading frame, bamHIC, in both Escherichia coli and Bacillus subtilis." J Bacteriol 174(22): 7194-201.

J

Jacob, C., F. Nouzieres, et al. (1997). "Isolation, characterization, and complementation of a motility mutant of Spiroplasma citri." J Bacteriol 179(15): 4802-10.

Jaffe, J. D., M. Miyata, et al. (2004). "Energetics of gliding motility in Mycoplasma mobile." J Bacteriol 186(13): 4254-61.

Jaffe, J. D., N. Stange-Thomann, et al. (2004). "The complete genome and proteome of Mycoplasma mobile." Genome Res 14(8): 1447-61.

Jancarik, J., R. Pufan, et al. (2004). "Optimum solubility (OS) screening: an efficient method to optimize buffer conditions for homogeneity and crystallization of proteins." Acta Crystallogr D Biol Crystallogr 60(Pt 9): 1670-3.

Janscak, P., U. Sandmeier, et al. (2001). "Subunit assembly and mode of DNA cleavage of the type III restriction endonucleases EcoP1I and EcoP15I." J Mol Biol 306(3): 417-31.

Jensen, J. S. (2006). "Mycoplasma genitalium infections. Diagnosis, clinical aspects, and pathogenesis." Dan Med Bull 53(1): 1-27.

Jensen, J. S., J. Blom, et al. (1994). "Intracellular location of Mycoplasma genitalium in cultured Vero cells as demonstrated by electron microscopy." Int J Exp Pathol 75(2): 91-8.

Jordan, J. L., H. Y. Chang, et al. (2007). "Protein P200 is dispensable for Mycoplasma pneumoniae hemadsorption but not gliding motility or colonization of differentiated bronchial epithelium." Infect Immun 75(1): 518-22.

Jurenaite-Urbanaviciene, S., J. Serksnaite, et al. (2007). "Generation of DNA cleavage specificities of type II restriction endonucleases by reassortment of target recognition domains." Proc Natl Acad Sci U S A 104(25): 10358-63.

K

Kabsch, W. and C. Sander (1983). "Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features." Biopolymers 22(2577-2637).

Kang, C., R. Chan, et al. (1995). "Crystal structure of the T4 regA translational regulator protein at 1.9 A resolution." Science 268(5214): 1170-3.

Kay, B. K., M. P. Williamson, et al. (2000). "The importance of being proline: the interaction of proline-rich motifs in signaling proteins with their cognate domains." Faseb J 14(2): 231-41.

Kelly, T. J., Jr. and H. O. Smith (1970). "A restriction enzyme from Hemophilus influenzae. II." J Mol Biol 51(2): 393-409.

Kennaway, C. K., A. Obarska-Kosinska, et al. (2009). "The structure of M.EcoKI Type I DNA methyltransferase with a DNA mimic antirestriction protein." Nucleic Acids Res 37(3): 762-70.

Kenri, T., S. Seto, et al. (2004). "Use of fluorescent-protein tagging to determine the subcellular localization of mycoplasma pneumoniae proteins encoded by the cytadherence regulatory locus." J Bacteriol 186(20): 6944-55.

Kim, J. S., A. Degiovanni, et al. (2005). "Crystal structure of DNA sequence specificity subunit of a type I restriction-modification enzyme and its functional implications." Proc Natl Acad Sci U S A 102(9): 3248-53.

Kim, S. H., D. H. Shin, et al. (2005). "Structural genomics of minimal organisms and protein fold space." J Struct Funct Genomics 6(2-3): 63-70.

Kim, K. K., H. Yokota, et al. (1999). "Four-helical-bundle structure of the cytoplasmic domain of a serine chemotaxis receptor." Nature 400(6746): 787-92.

Kneale, G. G. (1994). "A symmetrical model for the domain structure of type I DNA methyltransferases." J Mol Biol 243(1): 1-5.

Kobayashi, I. (2001). "Behaviour of restriction-modification systems as selfish mobile elements and their impact on genome evolution." Nucl. Acid. Res. 29: 3742-3756.

Kostyal, D. A., G. H. Butler, et al. (1994). "A 48-kilodalton Mycoplasma fermentans membrane protein induces cytokine secretion by human monocytes." Infect Immun 62(9): 3793-800.

Krause, D. C. (1996). "Mycoplasma pneumoniae cytadherence: unravelling the tie that binds." Mol Microbiol 20(2): 247-53.

Krause, D. C. and M. F. Balish (2004). "Cellular engineering in a minimal microbe: structure and assembly of the terminal organelle of Mycoplasma pneumoniae." Mol Microbiol 51(4): 917-24.

Krause, D. C., T. Proft, et al. (1997). "Transposon mutagenesis reinforces the correlation between Mycoplasma pneumoniae cytoskeletal protein HMW2 and cytadherence." J Bacteriol 179: 2668-2677.

Kuhner, S., V. van Noort, et al. (2009). "Proteome organization in a genome-reduced bacterium." Science 326(5957): 1235-40.

Kusano, K., T. Naito, et al. (1995). "Restriction-modification systems as genomic parasites in competition for specific sequences." Proc Natl Acad Sci U S A 92(24): 11095-9.

Kusano, K., K. Sakagami, et al. (1997). "A new type of illegitimate recombination is dependent on restriction and homologous interaction." J Bacteriol 179(17): 5380-90.

L

Laemmli, U. K. and M. Favre (1973). "Maturation of the head of bacteriophage T4. I. DNA packaging events." J Mol Biol 80(4): 575-99.

Lamont, R. F., F. Anthony, L. Myatt, L. Booth, P. M. Furr, D. Taylor-Robinson (1990). "Prodution of prostaglandinE2 by human amnion in vitro in response to addition of media conditioned by microorgnisms associated chlorioamnionitis and preterm labor." A. J. Obstet. Gynecol. 162: 819-825.

Larkin, M. A., G. Blackshields, et al. (2007). "Clustal W and Clustal X version 2.0." Bioinformatics 23(21): 2947-8.

Laskowski, R. A., MacArthur, R. W., Moss, D. S., Thornton, J. M. (1993). "PROCHECK: a program to check the stereochemical quality of protein structures." J. Appl. Cryst. 26: 283-291.

Layh-Schmitt, G., A. Podtelejnikov, et al. (2000). "Proteins complexed to the P1 adhesin of Mycoplasma pneumoniae." Microbiology 146 (Pt 3): 741-7.

Leahy, D. J., H. P. Erickson, et al. (1994). "Crystallization of a fragment of human fibronectin: introduction of methionine by site-directed mutagenesis to allow phasing via selenomethionine." Proteins 19(1): 48-54.

Lee, N. S., O. Rutebuka, et al. (1997). "KpnAI, a new type I restriction-modification system in Klebsiella pneumoniae." J Mol Biol 271(3): 342-8.

Leslie, A. G. W. (1992). "Recent changes to the MOSFLM package for processing film and image plate data" Joint CCP4 + ESF-EAMCB Newsletter on Protein Crystallography 26.

Li, Z. and G. J. Jensen (2009). "Electron cryotomography: a new view into microbial ultrastructure." Curr Opin Microbiol 12(3): 333-40.

Lieutaud, P., B. Canard, et al. (2008). "MeDor: a metaserver for predicting protein disorder." BMC Genomics 9 Suppl 2: S25.

Ligon, J. V. and G. E. Kenny (1991). "Virulence of ureaplasmal urease for mice." Infect Immun 59(3): 1170-1.

M

Martin, S. R. and M. J. Schilstra (2008). "Circular dichroism and its application to the study of biomolecules." Methods Cell Biol 84: 263-93.

McBride, M. J. (2001). "Bacterial gliding motility: multiple mechanisms for cell movement over surfaces." Annu Rev Microbiol 55: 49-75.

Meisel, A., T. A. Bickle, et al. (1992). "Type III restriction enzymes need two inversely oriented recognition sites for DNA cleavage." Nature 355(6359): 467-9.

Mernagh, D. R., L. A. Reynolds, et al. (1997). "DNA binding and subunit interactions in the type I methyltransferase M.EcoR124I." Nucl. Acid. Res. 25: 987-991.

Miyata, M. (2008). "Centipede and inchworm models to explain Mycoplasma gliding." Trends Microbiol 16(1): 6-12.

Murray, N. E. (2000). "Type I restriction systems: sophisticated molecular machines (a legacy of Bertani and Weigle)." Microbiol Mol Biol Rev 64(2): 412-34.

Murray, N. E., J. A. Gough, et al. (1982). "Structural homologies among type I restriction-modification systems." Embo J 1(5): 535-9.

Murshudov, G. N., A. A. Vagin, et al. (1997). "Refinement of macromolecular structures by the maximum-likelihood method." Acta Crystallogr D Biol Crystallogr 53(Pt 3): 240-55.

Musatovova, O., S. Dhandayuthapani, et al. (2006). "Transcriptional heat shock response in the smallest known self-replicating cell, Mycoplasma genitalium." J Bacteriol 188(8): 2845-55.

Mushegian, A. R. and E. V. Koonin (1996). "A minimal gene set for cellular life derived by comparison of complete bacterial genomes." Proc Natl Acad Sci U S A 93(19): 10268-73.

N

Naito, T., K. Kusano, et al. (1995). "Selfish behavior of restriction-modification systems." Science 267(5199): 897-9.

Nakane, D. and M. Miyata (2007). "Cytoskeletal "jellyfish" structure of Mycoplasma mobile." Proc Natl Acad Sci U S A 104(49): 19518-23.

Nakane, D. and M. Miyata (2009). "Cytoskeletal asymmetrical dumbbell structure of a gliding mycoplasma, Mycoplasma gallisepticum, revealed by negative-staining electron microscopy." J Bacteriol 191(10): 3256-64.

Ng, J. D., J. A. Gavira, et al. (2003). "Protein crystallization by capillary counterdiffusion for applied crystallographic structure determination." J Struct Biol 142(1): 218-31.

Niesen, F. H., H. Berglund, et al. (2007). "The use of differential scanning fluorimetry to detect ligand interactions that promote protein stability." Nat Protoc 2(9): 2212-21.

O

O'Neill, M., D. T. Dryden, et al. (1998). "Localization of a protein-DNA interface by random mutagenesis." Embo J 17(23): 7118-27.

O'Neill, M., L. M. Powell, et al. (2001). "Target recognition by EcoKI: the recognition domain is robust and restriction-deficiency commonly results from the proteolytic control of enzyme activity." J Mol Biol 307(3): 951-63.

Obarska, A., A. Blundell, et al. (2006). "Structural model for the multisubunit Type IC restriction-modification DNA methyltransferase M.EcoR124I in complex with DNA." Nucleic Acids Res 34(7): 1992-2005.

Otwinowski, Z., Minor, W. (1997). Processing of X-ray Diffraction Data Collected in Oscillation Mode. Methods in Enzymology, Macromolecular Crystallography, part A. J. a. R. M. S. C.W. Carter, Eds., Academic Press. 276: 307-326.

P

Pape, T., Schneider, T. R. (2004). "HKL2MAP: a graphical user interface for phasing with SHELX programs" J. Appl. Cryst. 37: 843-844.

Patel, J., I. Taylor, et al. (1992). "High-level expression of the cloned genes encoding the subunits of and intact DNA methyltransferase, M.EcoR124." Gene 112(1): 21-7.

Perrakis, A., R. Morris, et al. (1999). "Automated protein model building combined with iterative structure refinement." Nat Struct Biol 6(5): 458-63.

Pich, O. Q., R. Burgos, et al. (2006). "Mycoplasma genitalium mg200 and mg386 genes are involved in gliding motility but not in cytadherence." Mol Microbiol 60(6): 1509-19.

Proft, T., H. Hilbert, et al. (1995). "The proline-rich P65 protein of Mycoplasma pneumoniae is a component of the Triton X-100-insoluble fraction and exhibits size polymorphism in the strains M129 and FH." J Bacteriol 177(12): 3370-8.

Proft, T., H. Hilbert, et al. (1996). "The P200 protein of Mycoplasma pneumoniae shows common features with the cytadherence-associated proteins HMW1 and HMW3." Gene 171(1): 79-82.

R

Razin, S. and E. Jacobs (1992). "Mycoplasma adhesion." J Gen Microbiol 138(3): 407-22.

Razin, S., R. Herrmann (2002). Molecular Biology and Pathogenicity of Mycoplasmas. New York, Kluwer Academic / Plenum Publishers.

Razin, S., D. Yogev, et al. (1998). "Molecular biology and pathogenicity of mycoplasmas." Microbiol Mol Biol Rev 62(4): 1094-156.

Regula, J. T., G. Boguth, et al. (2001). "Defining the mycoplasma 'cytoskeleton': the protein composition of the Triton X-100 insoluble fraction of the bacterium Mycoplasma pneumoniae determined by 2-D gel electrophoresis and mass spectrometry." Microbiology 147(Pt 4): 1045-57.

Relich, R. F., A. J. Friedberg, et al. (2009). "Novel cellular organization in a gliding mycoplasma, Mycoplasma insons." J Bacteriol 191(16): 5312-4.

Rossmann, M. G., Blow, D. M. (1962). "The detection of subunits within the crystallographic asymmetric unit." Acta Cryst. 15: 24-31.

Roberts, R. J. and X. Xeng (1998). "Base flipping." Annu. Rev. Biochem. 67: 181-198.

Roussel, A. and C. Cambillau (1989). "TURBO-FRODO." In Silicon Graphics Geometry Partner Directory: 77-78.

Rottem, S. (2003). "Interaction of mycoplasmas with host cells." Physiol Rev 83(2): 417-32.

Roy, P. H. and H. O. Smith (1973). "DNA methylases of Hemophilus influenzae Rd. I. Purification and properties." J Mol Biol 81(4): 427-44.

S

Sambrook, J., Russel D. (2001). Molecular Cloning: a Laboratory Manual. New York, Cold Spring Harbor Laboratory Press.

Schluckebier, G., M. Kozak, et al. (1997). "Differential binding of S-adenosylmethionine S-adenosylhomocysteine and Sinefungin to the adenine-specific DNA methyltransferase M.TaqI." J Mol Biol 265(1): 56-67.

Schneider, T. R. and G. M. Sheldrick (2002). "Substructure solution with SHELXD." Acta Crystallogr D Biol Crystallogr 58(Pt 10 Pt 2): 1772-9.

Schouler, C., F. Clier, et al. (1998). "A type IC restriction-modification system in Lactococcus lactis." J Bacteriol 180(2): 407-11.

Seto, S., T. Kenri, et al. (2005). "Involvement of P1 adhesin in gliding motility of Mycoplasma pneumoniae as revealed by the inhibitory effects of antibody under optimized gliding conditions." J Bacteriol 187(5): 1875-7.

Seto, S., G. Layh-Schmitt, et al. (2001). "Visualization of the attachment organelle and cytadherence proteins of Mycoplasma pneumoniae by immunofluorescence microscopy." J Bacteriol 183(5): 1621-30.

Seto, S. and M. Miyata (1999). "Partitioning, movement, and positioning of nucleoids in Mycoplasma capricolum." J Bacteriol 181(19): 6073-80.

Seto, S. and M. Miyata (2003). "Attachment organelle formation represented by localization of cytadherence proteins and formation of the electron-dense core in wild-type and mutant strains of Mycoplasma pneumoniae." J Bacteriol 185(3): 1082-91.

Seto, S., A. Uenoyama, et al. (2005). "Identification of a 521-kilodalton protein (Gli521) involved in force generation or force transmission for Mycoplasma mobile gliding." J Bacteriol 187(10): 3502-10.

Seybert, A., R. Herrmann, et al. (2006). "Structural analysis of Mycoplasma pneumoniae by cryo-electron tomography." J Struct Biol 156(2): 342-54.

Sharp, P. M., J. E. Kelleher, et al. (1992). "Roles of selection and recombination in the evolution of type I restriction-modification systems in enterobacteria." Proc Natl Acad Sci U S A 89(20): 9836-40.

Shi, W., Y. Zhou, et al. (1992). "DnaK, DnaJ, and GrpE are required for flagellum synthesis in Escherichia coli." J Bacteriol 174(19): 6256-63.

Siegert, R., M. R. Leroux, et al. (2000). "Structure of the molecular chaperone prefoldin: unique interaction of multiple coiled coil tentacles with unfolded proteins." Cell 103(4): 621-32.

Sitaraman, R. and K. Dybvig (1997). "The hsd loci of Mycoplasma pulmonis: organization, rearrangements and expression of genes." Mol Microbiol 26(1): 109-20.

Sjostrom, J. E., S. Lofdahl, et al. (1978). "Biological characteristics of a type I restriction-modification system in Staphylococcus aureus." J Bacteriol 133(3): 1144-9.

Skinner, M. M., H. Zhang, et al. (1994). "Structure of the gene V protein of bacteriophage f1 determined by multiwavelength x-ray diffraction on the selenomethionyl protein." Proc Natl Acad Sci U S A 91(6): 2071-5.

Smith, M. A., C. M. Read, et al. (2001). "Domain structure and subunit interactions in the type I DNA methyltransferase M.EcoR124I." J Mol Biol 314(1): 41-50.

Stuart, D. I., M. Levine, et al. (1979). "Crystal structure of cat muscle pyruvate kinase at a resolution of 2.6 A." J Mol Biol 134(1): 109-42.

Studier, F. W. (2005). "Protein production by auto-induction in high density shaking cultures." Protein Expr Purif 41(1): 207-34.

Sugahara, M., Y. Asada, et al. (2005). "Heavy-atom Database System: a tool for the preparation of heavy-atom derivatives of protein crystals based on amino-acid sequence and crystallization conditions." Acta Crystallogr D Biol Crystallogr 61(Pt 9): 1302-5.

T

Tamura, K., J. Dudley, et al. (2007). "MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) software version 4.0." Mol Biol Evol 24(8): 1596-9.

Tao, T. and R. M. Blumenthal (1992). "Sequence and characterization of pvuIIR, the PvuII endonuclease gene, and of pvuIIC, its regulatory gene." J Bacteriol 174(10): 3395-8.

Taylor-Robinson, D. (1996). "Infections due to species of Mycoplasma and Ureaplasma: an update." Clin Infect Dis 23(4): 671-82; quiz 683-4.

Taylor, I. A., K. G. Davis, et al. (1994). "DNA-binding induces a major structural transition in a type I methyltransferase." Embo J 13(23): 5772-8.

Teichmann, S. A., J. Park, et al. (1998). "Structural assignments to the Mycoplasma genitalium proteins show extensive gene duplications and domain rearrangements." Proc Natl Acad Sci U S A 95(25): 14658-63.

Terwilliger, T. C. and J. Berendzen (1999). "Automated MAD and MIR structure solution." Acta Crystallogr D Biol Crystallogr 55(Pt 4): 849-61.

Thygesen, J., S. Weinstein, et al. (1996). "The suitability of multi-metal clusters for phasing in crystallography of large macromolecular assemblies." Structure 4(5): 513-8.

U

Uenoyama, A., A. Kusumoto, et al. (2004). "Identification of a 349-kilodalton protein (Gli349) responsible for cytadherence and glass binding during gliding of Mycoplasma mobile." J Bacteriol 186(5): 1537-45.

Uenoyama, A. and M. Miyata (2005). "Gliding ghosts of Mycoplasma mobile." Proc Natl Acad Sci U S A 102(36): 12754-8.

Uenoyama, A. and M. Miyata (2005). "Identification of a 123-kilodalton protein (Gli123) involved in machinery for gliding motility of Mycoplasma mobile." J Bacteriol 187(16): 5578-84.

V

Valinluck, B., N. S. Lee, et al. (1995). "A new restriction-modification system, KpnBI, recognized in Klebsiella pneumoniae." Gene 167(1-2): 59-62.

Vedadi, M., F. H. Niesen, et al. (2006). "Chemical screening methods to identify ligands that promote protein stability, protein crystallization, and structure determination." Proc Natl Acad Sci U S A 103(43): 15835-40.

Voelker, L. L. and K. Dybvig (1996). "Gene transfer in Mycoplasma arthritidis: transformation, conjugal transfer of Tn916, and evidence for a restriction system recognizing AGCT." J Bacteriol 178(20): 6078-81.

Voelker, L. L. and K. Dybvig (1999). "Sequence analysis of the Mycoplasma arthritidis bacteriophage MAV1 genome identifies the putative virulence factor." Gene 233(1-2): 101-7.

W

Walsh, P., D. Bursac, et al. (2004). "The J-protein family: modulating protein assembly, disassembly and translocation." EMBO Rep 5(6): 567-71.

Wolgemuth, C. W., O. Igoshin, et al. (2003). "The motility of mollicutes." Biophys J 85(2): 828-42.

X

Xu, G., J. Willert, et al. (1995). "BsuCI, a type-I restriction-modification system in Bacillus subtilis." Gene 157(1-2): 59.

Y

Yang, Z., Y. Geng, et al. (1998). "A DnaK homolog in Myxococcus xanthus is involved in social motility and fruiting body formation." J Bacteriol 180(2): 218-24.

Yasukawa, T. and C. Kanei-Ishii (1995). "Increase of solubility of foreign proteins in Escherichia coli by coproduction of the bacterial thioredoxin." J Biol Chem 270: 25328-31.