Structure prediction of P1-type ATPases and molecular dynamics simulations on their Metal Binding Domains

Karthik Arumugam

UNIVERSITÉ JOSEPH FOURIER - GRENOBLE 1

École Doctorale Chimie et Sciences du vivant

THÈSE

pour obtenir le grade de
DOCTEUR DE L'UNIVERSITÉ JOSEPH FOURIER

Spécialité : Biologie Structurale et Nanobiologie

présentée et soutenue publiquement le 12 novembre 2009

par

Karthik ARUMUGAM

Structure prediction of P1-type ATPases and molecular dynamics simulations on their Metal Binding Domains

JURY

Pr. Arthur SOUCEMARIANADIN	Président
Dr. Dorothée BERTHOMIEU	Rapporteur
Dr. Norbert GARNIER	Rapporteur
Dr. Patrice CATTY	Examinateur
Dr. Stéphane REDON	Examinateur
Dr. Serge CROUZY	Directeur de thése

LABORATOIRE DE CHIMIE ET BIOLOGIE DES MÉTAUX UMR 5249
Institut de Recherches en Technologies et Sciences pour le Vivant
Commissariat á l'Énergie Atomique (CEA), Grenoble

NANO-D - INRIA Grenoble - Rhone-Alpes
Laboratoire Jean Kuntzmann - UMR 5224 - Grenoble

Acknowledgements

I would like to thank my advisor, Dr. Serge Crouzy, for his great guidance and support throughout my research. It was a privilege to be a member of his team. I had a wonderful experience working in his lab where I enjoyed working on scientific projects and learned how to become a scientist. During the years we worked together he was much more than an advisor to me. He taught me a lot about life as well as science. Everything I have learned from him will stay with me throughout my life. This thesis would have never been possible without him.

I would like to thank my Co-supervisor Dr. Stephane Redon, Dr. Patrice Catty, Dr. Elisabeth Mintz, Dr. Florent Guillain and Dr. Michel Ferrand for their valuable advice and support during my PhD thesis.

Many thanks to Dr. Michel Vivaudou and Dr. Christophe Moreau, for his useful discussion about Hm2Kir6 project with me and allowing me to do research on the same.

I gratefully acknowledge the efforts of my dissertation committee members, Dr. Dorothee Berthomieu and Dr. Norbert Garnier. I owe my personal thanks to Prof. Arthur Soucemarianadin for his kind help during my initial time of my thesis.

I would also like to thank all my past and present member of my Lab. I am deeply indebted to the people at University Joseph Fourier and at CEA, for their help during my stay.

I gratefully acknowledge the financial support that I received from the CEA Direction des Relations Internationales (DRI) via EGIDE, and from the LCBM laboratory.

I would like to express my gratitude towards my Prof. R. Nayak and Prof. Shaila from Indian Institute of Science, India, for their guidance, encouragement and support. I especially thank my Deputy Registrar Mr. Paneerselvam for his continuous encouragement and trust on me.

A special thoughts to my friends who have made my life at Grenoble so enjoyable both inside and outside the Lab, including Prem, Sandeep, Tarun, Peng, Mahesh, Cheickna, Marine, Mouna, Pankaj, Abhinav, Robert, Shikha.

Finally I would like to thank my family. The biggest thanks which has been saved till last has to go to Amma and Appa, whose support, love and trust in me over the last 26 years has been unfaltering.

Contents

Chapter 1

Introduction

1.1 Historical background

Four billion years ago, the first living cell appeared. Mother earth had been waiting patiently for at least 10 billion years for this to happen and ever since that day, nothing has been quite the same. The cells replicated and replicated and finally a group of cells wrote this thesis.

In order to have a functional, living cell some of the basic things that are needed are: Some way of converting and utilizing energy, for instance chemical energy or the ray from the sun and self replication and self assembly. The first cells might have used RNA as the information carrier and as molecular machines. Nowadays, modern cells have DNA as the information carrier and RNA as an intermediate step towards proteins but sometimes it also functions as molecular machines.
One less obvious thing is that molecules need to be located at the right place at the right time. One thing that aids this is compartmentalisation, when molecules are kept separated by different kinds of barriers. For instance, the DNA molecule of a cell needs to be located inside the cell, if it is outside, it will rapidly get degraded.

My research has been focused on the proteins that are located in the membrane that surrounds cells and compartments. These proteins are responsible for transport through the membrane, they are involved in cell to cell signalling, they are crucial for generating and converting chemical and sun rays into useful energy. In order to be able to understand these processes in detail, a 3D structure of the protein is needed. Unfortunately, structure determination of membrane proteins is very time consuming and difficult so there are considerably fewer structures available than for water soluble proteins. There are millions of proteins with known amino acid sequence and tens of thousands of proteins with known structure but only a few hundred of these structures are membrane proteins. Therefore, theoretical approaches that give structural insight from the amino acid sequence hold great potential since computing time is very cheap. However, since computational predictions are just predictions, it is also important to validate the theoretical results with practical experiments.

1.2 From sequence to structure to function

The revolutionary discovery of the molecular structure of DNA by Watson and Crick (1953) made a tremendous impact on science [1]. Never before had it been possible to study and understand the molecular details of how genetic information was stored. It could be seen that the DNA double helix was stabilize by hydrogen bonds between Adenine (A),Thymine (T), Guanine (G) and Cytosine(C). The problem was that no one knew how the DNA sequences were translated into proteins.

The genetic code was finally broken by the Nirenberg laboratory in 1961-66 [2]. It was discovered that each DNA nucleotide triplet stands for one of the 20 different amino acids that are the building blocks of peptides and proteins. These events lead to the birth of molecular biology.

There are two main events involved in going from DNA to proteins. The first step is the transcription, copying, of DNA into complementary RNA by the enzyme RNA polymerase. This enzyme is essential for life and present in all organisms and also in some viruses. The second step is the translation of the mRNA into an amino acid that forms the protein. This step is performed by the ribosome which translates the genetic code to an amino acid chain.

One of the remaining unsolved problems of molecular biology is how the linear amino acid chain folds into a three dimensional protein. All information about how a protein structure will look like is present in its amino acid sequence [3].

A water soluble protein folds within fractions of a second and that makes it impossible to explore all different conformations that exist for the amino acids. The Levinthal paradox states that due to the large amount of degrees of freedom in the protein chain the molecule has an astronomical number of possible conformations as estimate is 10^{143}, [4]. If the protein would have to explore all conformations it would take longer time that the universe has existed. Thus, there exist folding pathways, yet the details of them remain unclear.

The key to a detailed understanding of protein function lies in its structure. A high resolution structure is needed to be able to study the protein function at a molecular level. The most common method to determine the structure is to use X-ray crystallography.

1.3 The cell membrane

Membranes are a key element for life since they act as a physical barrier to separate the interior of a cell from the outside world. They also serve to confine its different compartments, providing the means for a cell to be out from the equilibrium with the surroundings, allowing important processes occurring in the interior. In addition to be crucial for cell integrity, cell membranes also serve as a matrix and support for many types of proteins involved in important cell functions and therefore, they are essential for cell function.

Biological membranes consist of organized assemblies of lipids and proteins. The current knowledge of how they are structured is still scarce due to the difficulties associated to the experimental techniques required to investigate their properties. The modern view of biological membranes is basically based on the fluid mosaic model proposed in the seventies [5], in which lipids are arranged in bilayers, where proteins are embedded, and subject to a lateral freely diffusion. At that time, however, the involvement of lipids in membrane function was overemphasized, partly because membrane proteins

were not readily accessible and their function was poorly understood. With the growing knowledge of membrane proteins, the accepted vision has become more protein centered and the crucial role of lipids diluted. It is not until more recently that, with a better understanding of lipid-protein interactions, the role of lipids in protein function is again having an increasing attention [6, 7]. The current vision of membranes is more as a mosaic than a fluid: lipids organize a matrix where proteins are distributed in regions of biased composition with varying protein environment [8].

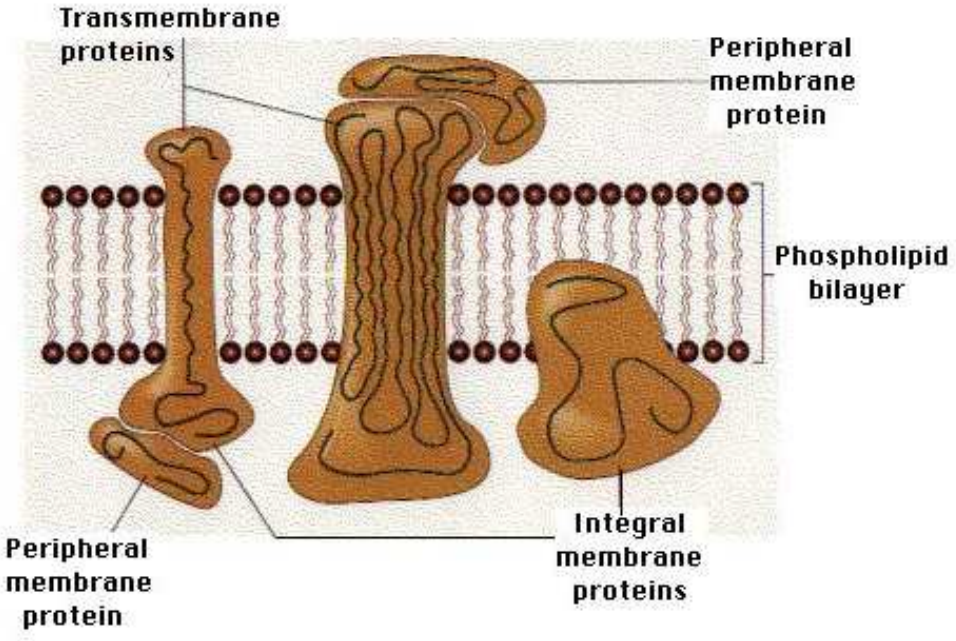

Figure 1.1: The cell membrane

Membranes consist of a complex mixture of lipids, where different proteins both, integral and peripheral are embedded (See Fig. 1.1). Protein content varies greatly among the different kinds of membranes, ranging typically between 15 to 75%, depending on the functions that they must carry out [9]. Furthermore, lipid composition changes from one membrane to another due to the enormous structural diversity found, that can be associated with the differential roles and properties of each membrane or region. The most widely found lipids consist of a structure of a fatty acid linked by an ester bond to an alcohol such as glycerol or cholesterol, or through amide bonds to a sphingoid base or to other amines. Most lipids have a highly polar head group and two hydrocarbon tails.

In a typical membrane, approximately half of the lipids are phospholipids, mainly phosphatidyl -cholines (PC), -ethanolamines (PE) and -serines (PS). Other major components following in importance are sphingolipids, glycolipids and cholesterol [10]. Interestingly, since the two sides of the membrane bilayer must deal with different surroundings, the two leaflets of a membrane typically exhibit an asymmetric composition. The lateral distribution of lipids is also nonhomogeneous in regard to lipid components, providing regions with different levels of molecular order that are known as rafts. These seem to be important for the function of membrane proteins [11]. Finally, the presence of specific components such as cholesterol has been found to trigger domain formation in lipid bilayer

membranes [12].

1.4 Computational biology and Bioinformatics

The amount of available genetic information continues to increase at a steady pace. Ever since GenBank, a sequence database, started 1982 it has doubled its size every 18 months. The latest release (v160) contains 77,248,690,945 bases. To put this enormous number into perspective, the number of bases that was sequenced from April to June 2007 was 42,880 each second. The DNA sequencing technology also gets faster and faster, e.g. a 454 gene sequencer can sequence 25 million bases with 99% accuracy in 4 hours, which makes it possible to sequence a bacterial genome during the course of a day [13].

There is clearly a need for methods that can make sense out of all this vast amount of genetic information and this is where bioinformatics and computational biology steps in. The difference in definition between the two terms is that computational biology refers to hypothesis driven investigations of biological problems using computers while bioinformatics is more technique driven and concerns development of algorithms and computational and statistical techniques. Many times the two terms are used interchangeably.

The unifying topic in this thesis is membrane proteins and Computational Biophysics technology. I will then continue with a brief overview about membrane proteins. Then, I will focus on metal transporting proteins.

1.4.1 Membrane proteins:

Inside the membrane there are membrane proteins with many diverse functions. They are responsible for selecting which substances are allowed to pass through the membrane, they are important for transmitting signals, cell-to-cell communication, and numerous other functions. More than half of all drugs are targeting membrane proteins [14, 15].

The amount of membrane proteins inside the lipid bilayer vary from organism to organism, from cell type to cell type. For instance, membranes proteins make up 75% of the mass of the membrane in *E.coli* while only 30% for human myelin sheath cells.

There are three main classes of membrane proteins:

1.4.2 Alpha (α)-Helical membrane proteins:

α-helical membrane proteins have their amino acids arranged in tightly packed helices in the transmembrane regions (Fig. 1.2A). The reason behind the α-helical arrangement is to have as large hydrophobic outer surface area as possible that can interact with the hydrophobic fatty acids of the lipids. At the same time, it also satisfies all its internal backbone hydrogen bonds.

α-helical membrane proteins typically make up 20-30% of the encoded proteins of an organism [16]. Most often they are located in the inner membrane but recently one example, a polysaccharide transporter, has been found that resides in the outer membrane of *E. coli* [17].

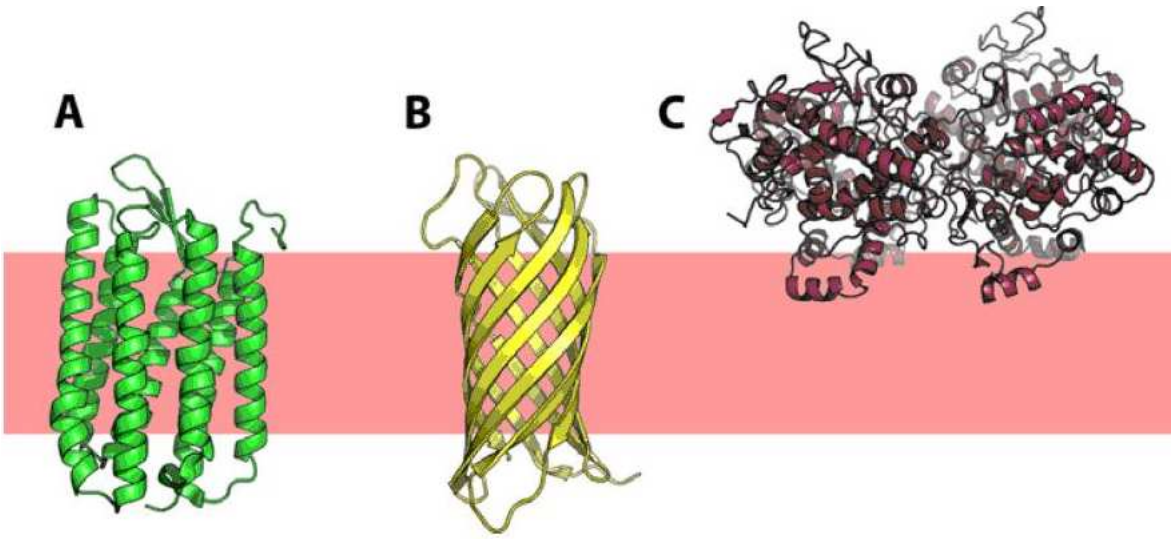

Figure 1.2: Example of membrane protein structures with the membrane depicted as a pink rectangle. A. An α-helical membrane protein with 7 transmembrane helices like bacteriorhodopsine studied in section 4.6.1.1. B. A β-barrel membrane protein with 8 transmembrane β- strands. C. A monotopic membrane protein.

1.4.3 Beta (β)-barrel membrane proteins:

The other possibility of spanning the membrane while satisfying internal backbone hydrogen bonds and maintaining a large hydrophobic surface area is the β-barrel fold (Fig. 1.2B). Since a β-strand always hydrogen bonds to another β-strand, these proteins consist of β-hairpins through the membrane. Thus, this class always has an even number of β-strands. Every second residue in the transmembrane region is facing the inside of the barrel whereas the other is facing the outside, towards the lipids. This alternating pattern is reflected in the amino acid sequence where one amino acid is non-polar and hydrophobic (and directed towards the lipids) and the next one more polar (and directed towards the interior of the barrel) etc. Since a β-strand is extended compared to the coiled α-helix, the transmembrane regions are shorter and typically consist of 8 to 15 amino acids [18]. The β- barrel membrane proteins can be found in the outer membrane of bacteria, mitochondria and chloroplasts.

Around 2-3% of a genome encodes β-barrel membrane proteins [19] but that is an uncertain number given the difficulties of identifying them. The main reason for this is that β-strands are more difficult to predict than α-helices since they consist of fewer residues and contain more long range interactions. There are also fewer structures of β-barrel membrane proteins than α-helical.

1.4.4 Monotopic and peripheral membrane proteins

Monotopic membrane proteins are only associated to one of the two leaflets of the membrane. Prostaglandin H2 synthase-1 is one example and is shown in Fig. 1.2C. [20]. The helices that are associated to the membrane are amphiphilic [21] and the protein is only removable from the membrane by detergents, organic solvents or denaturants that interfere with the hydrophobic interactions.

A peripheral membrane protein is more loosely attached to the membrane, mainly through elec-

trostatic interactions and hydrogen bonding with the polar headgroup of lipids or with hydrophilic domains of other integral membrane proteins. Peripheral membrane proteins can be dissociated from the membrane by treatment with solutions of high ionic strength or elevated pH.

1.5 Metals in Biological Systems

1.5.1 Metal homeostasis and trafficking

Organisms require essential heavy metals including Cu, Zn, Mn, Fe, Co, Ni and Mo to carry out biological functions. Heavy metals act as cofactors in enzyme reactions including group transfer, redox and hydrolysis [22]. Others like Na, K, and Ca are involved in physiological processes and/or maintaining structure as well as controlling the function of cell walls. It has been estimated that over a quarter of all known enzymes require a particular metal ion for function. These enzymes can be divided into two groups: metal activated enzymes and metalloenzymes. Metal-activated enzymes require the addition of metal ions to become stimulated [23]. Kinases are an example of this type of enzymes because of its use of the Mg-ATP complex as a phosphoryl group donating substrate. Metalloenzymes necessitate metal ions to function properly. The metal ion is firmly bound to the enzyme and is frequently recycled after protein degradation. Heme groups in hemoglobin or cytochromes bind a Fe^{2+} ion tightly. The Chlorophyll that sustains every ecosystem also needs a Mg^{2+} ion. The Cu-Zn superoxide dismutases (SOD) require metal ions to detoxify the cell from dangerous free radicals. These are some of the many examples of diverse roles that metal ions play. Metal ions are essential for life, yet they are toxic to the cells at high concentrations or in the free form [22, 24]. They can generate free radicals, which are highly reactive and oxidant molecules that can react with any biomolecule such as nucleic acids. This interaction with biomolecules may eventually lead to cell death. Cells have developed highly complex systems used for sensing, trafficking and transporting of metals. Many of these systems are not well understood yet [25, 26].

In biological systems, these metals are mostly bound to proteins. In these metalloproteins, they have catalytic and structural roles: 1) as constituents of enzyme active sites; 2) stabilizing enzyme tertiary or quaternary structure; 3) forming weak-bonds with substrates contributing to their orientation to support chemical reactions; and 4) stabilizing charged transition states [27]. In oxygenated states Cu, Fe and Mn have unpaired electrons that allow their participation in redox reactions in enzyme active sites [27]. For instance, Cu mediates the reduction of one superoxide anion to hydrogen peroxide and oxidation of a second superoxide anion to molecular oxygen in the active site of cytoplasmic superoxide dismutase [28]. Zn does not have any unpaired electrons in the Zn^{2+} state and it has been proposed to prevent the formation of harmful free radicals by competing with the redox active metals such as Fe and Cu in the enzyme active sites [29]. Zn^{2+} is also a cofactor of a number of enzymes including RNA polymerase, carbonic anhydrase and Cu/Zn superoxide dismutase [27]. Other heavy metals including Cd, Pb, Cr, Hg and As have no known physiological activity and are non-essential [30]. Elevated levels of both, essential and non-essential heavy metals, results in toxicity symptoms mostly associated with the formation of reactive oxygen species (ROS) that can initiate oxidative damage. Alternatively, they bind to sulphydryl and amino groups in proteins,

leading to activity inhibition or structure disruption [31]. Therefore, organisms have tightly controlled homeostatic mechanisms to maintain physiological concentrations of essential heavy metals in different cellular compartments and to minimize the damage from exposure to non-essential ones.

The main mechanisms of heavy metal homeostasis include transport, chelation, and detoxification by efflux or sequestration into organelles. Heavy metals are transported into the cells by various transmembrane metal carriers. It has been shown that although cellular Zn^{2+} or Cu^{2+} total concentrations are in the millimolar and micromolar range respectively, cytosolic free Zn^{2+} concentration is in the femtomolar range while free Cu^{+} is in the zeptomolar range, i.e. less than one free atom per cell [25, 32]. This indicates that the heavy metals are immediately complexed with molecules or peptides upon entry to the cell. Chelators buffer cytosolic metal concentrations and they include molecules such as phosphates, phytates, polyphenols and glutathiones, or small peptides such as phytochelatins and some proteins like metallothioneins [27, 33]. Some of these chelators are thought to be involved in metal transport into subcellular organelle. For instance, it has been shown that Cadmium ATPase from *Listeria monocytogenesis*, subject of chapter 4 of this work, transports Cd^{2+} complexes into the plasma. Chaperones are proteins that bind specific essential heavy metals and deliver them to particular target metalloproteins where they function as part of the enzymatic activity. Similarly, they traffic the metal to specific membrane transporters that efflux the metal to the extracellular space and the lumen of subcellular organelles [34]. Fig. 1.3 summarizes the interplay of Cu^{+} with different metallochaperones in yeast.

1.5.2 The human copper metallochaperone: Atox1 (or Hah1)

The redox potential of copper makes it a useful cofactor for many enzymes, but also requires copper to be sequestered at all times to avoid oxidative damage. In yeast, it has been demonstrated that almost no copper exists free in solution; instead all copper is bound to proteins or small molecules such as glutathione [35]. When copper enters the cell through the transporter Ctr1, it is ahought to immediately pass to another molecule with the help of a copper chaperone. (See Fig. 1.3). In humans, the cytosolic protein Atox1 (HAH1) has been shown to play a key role in the delivery of copper to specialized membrane proteins called Cu-ATPases (See section 3.4 for a complete description of these proteins). The deletion of the Atox1 gene in mice leads to both intracellular copper accumulation as well as a decrease in the activity of secreted copper-dependent enzymes [36], indicative of the diminished Cu-ATPase transport activity. In yeast, the Atox1 ortholog Atx1 has been shown to facilitate function of Ccc2, the P-type ATPase, which transports copper into the late Golgi compartment [37]. The evolutionary conservation of functional interactions between Atox1 and Cu-ATPases emphasizes the role of the chaperone in regulating the Cu-ATPase function.
Structurally, Atox1 is a 68-amino acid residues protein that has the $\beta\alpha\beta\beta\alpha\beta$ fold and a single MxCxxC copper binding motif common with the N-terminal Metal Binding Domains (MBDs) of Cu-ATPases [38]. Similarly to the N-terminal MBDs, Atox1 binds copper with a linear, two-coordinate geometry but other data, including work done in the laboratory [39], are highly suggestive of a transfer mechanism where copper moves from Atox1 to a metal-binding domain of Cu-ATPase through a three-coordinate intermediate. The crystal structure of Atox1, showing the molecule as a homo-dimer

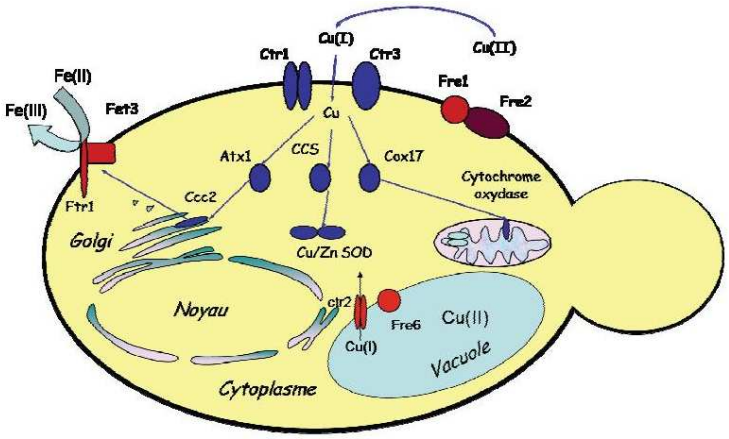

Figure 1.3: Copper trafficking in a yeast cell *S. cerevisiae*. Under Cu-starvation conditions, Cu is transported into yeast cells with high affinity, following reduction from Cu^{2+} to Cu^{+} by the Fre1 and Fre2 plasma membrane Cu(II)/Fe(III) ion reductases. Then, the high-affinity Cu transporters Ctr1 and Ctr3 mediate the passage of Cu across the plasma membrane Once inside the cell, CCS ensures the distribution of copper to the Cu/Zn Super oxyde dismutase (SOD), Cox17 is a key-element for the incorporation of copper into mitochondriae and Atx1 supplies copper to to the P-type ATPase: Ccc2 in the Golgi apparatus. Once Cu^{+} reaches Ccc2, it is transported to the lumen of the Golgi/endosome compartment. Here, four Cu atoms are assembled with Fet3. The Fet3-Ftr1 high-affinity Fe-transport complex assembles at the plasma membrane. Ctr2 is a vacuolar copper conveyor making it possible to release the ion into the cytosol.

situated around a single copper molecule, provided a glimpse of how such copper-transfer intermediate may look [38]. In solution, the intermediate is transient and cannot be visualized [40].

1.6 What this Thesis is About ?

Both chemistry and most other experimental sciences usually rely on a top-down approach. That is, measurements are gradually refined to be able to observe smaller structures and faster processes until technical limits are reached. If we had access to very, very powerful computers one could try to reverse this algorithm and do bottom-up modelling instead. This is the basic idea behind the approach I have used for the research summarized in this thesis; starting from simple (almost trivial) pairwise interactions of atoms, computers can be used to simulate what happens in complex biological molecules on longer timescales. This way, it is actually possible to see the atomic motions on a level usually not accessible to experiments. The knowledge gained can then be used to return to the drawing table and formulate better models for the phenomena observed, to be able to understand and perhaps even manipulate the systems, e.g. transport drug molecules or metals through cellular membranes. Further, when the simulations reach time and length scale where it is also possible to perform experimenst (and agree with these), the chemistry and physics of the molecules can be traced all the way from individual atoms up "real-world" macroscopic systems. (Of course, that is the nice, ideal picture in practice it is hours, days and months of programming, debugging, running dynamics and waiting for those long runs to finish. But at least you never have to do any wet laboratory work !)

More precisely, and from the title of this thesis, I've been interested in structure prediction of heavy metal ATPases and dynamics of their Metal Binding Domains using an *in silico* approach. After this introduction on membrane proteins and metals in biological systems, the thesis is divided in 4 main chapters:

- Theory and Methods: I will present computational techniques used to predict the three-dimensional structure of proteins when it is unknown experimentally. Then, I will present molecular mechanics force fields and molecular dynamics simulation algorithms including adaptive molecular dynamics that I have used to study the function of these proteins.

- Metal-Binding Membrane Proteins: In this chapter, I will detail the current knowledge on heavy metal transporting ATPases focussing on Cd^{2+} and Cu^{+} ATPases. I will also show how we could model the structure of an ion channel coupled to a 7-transmembrane-helix receptor and open the way for the study of signal transduction in this system.

- Structure and dynamics of The TransMembrane region of a Cd^{2+} ATPase: I have applied computational techniques to yield several possible models of the TM region of a Cd^{2+} ATPase also studied experimentally in the laboratory.

- Dynamics and stability of the Metal Binding Domains of the Menkes ATPase: In this chapter, I've used the known 3D structure of the N-terminal soluble part of a Cu^{+} ATPase as a starting point for molecular dynamics simulations of the metal binding domains of this protein in the presence and absence of metal.

To summarize in a few words, the purpose of my work has been to try to relate the structure and function of two particular heavy metal ATPases. When the structure of these ATPases was unknown structure prediction methods have been used. Molecular dynamics simulations have been used to approach the function of these proteins. The final goal is to understand how the metal is passed from soluble chaperones which bind the metal as soon as it enters the cell to their partner ATPases whose role is to distribute this metal where it's needed or eliminate it. My work may also be used to predict the effect of or propose amino acid mutations that impair the normal behavior of these proteisn and yield diseases like the Menkes disease.

Chapter 2

Theory and Methods

When attempting to study biological and chemical mechanisms in proteins using computational techniques, the first requirement is the structure of the macromolecule involved. When this structure is not available from databases, we can either try to predict it *ab initio*, that is without any knowledge about other existing structures, or use homology modeling when the structure of a similar protein is known.

I will start describing a new method that we have developed to build structural models of the transmembrane part of proteins when no homology is present. Then I will present homology modelling and the program MODELLER that we have used to build models of an ABC transporter. Finally, the structure of the protein being known, I will present the techniques of molecular mechanics and dynamics and the program CHARMM that I have used to study the dynamics of ATPases.

2.1 Structure prediction using NOE-like restraints with X-PLOR

2.1.1 Introduction

X-PLOR [41] is a program used to determine and refine solution NMR structures based on interproton distance estimates (from Nuclear Overhauser Effects or NOE), coupling constants measurements, and other information, such as known hydrogen-bonding patterns. What such measures can we theoretically derive or predict when building models of the transmembrane (TM) part of an α-helical membrane protein ?

- Each TM helix can be considered as a regular α-helix. We then know that the distance between H atom of residue i in the helix and O atom of residue $i+4$ is close to 1.5 Å .
- We also know that in such a regular helix, the ϕ and ψ dihedral angles have values close to -60 and -50 degrees, respectively.
- TM helices form a bundle in which the center of 1 helix is close to the center of 3 or 4 neighbour helices; the distance between centers lies around 10 Å .
- In Metal transporting TM proteins, residues in 3 or 4 TM helices are known to bind the transported metal; this knowledge allows us to set up some new distance restraints between metal binding residues.

An overview of distance-based NMR structure determination is shown in Fig. 2.1. There are several alternative methods possible in X-PLOR: full-structure distance geometry, substructure distance geometry, and *ab initio* simulated annealing (SA) starting from template structures or random coordinates. The choice of protocol depends on the desired efficiency and sampling of conformational space. In our case of theoretical distance restraints substructure embedding and regularization followed by SA refinement have been used. Simulated annealing is so powerful that it can convert a random array of atoms into a well defined structure through distance restraints. Structures obtained by this protocol have to be regularized and refined.

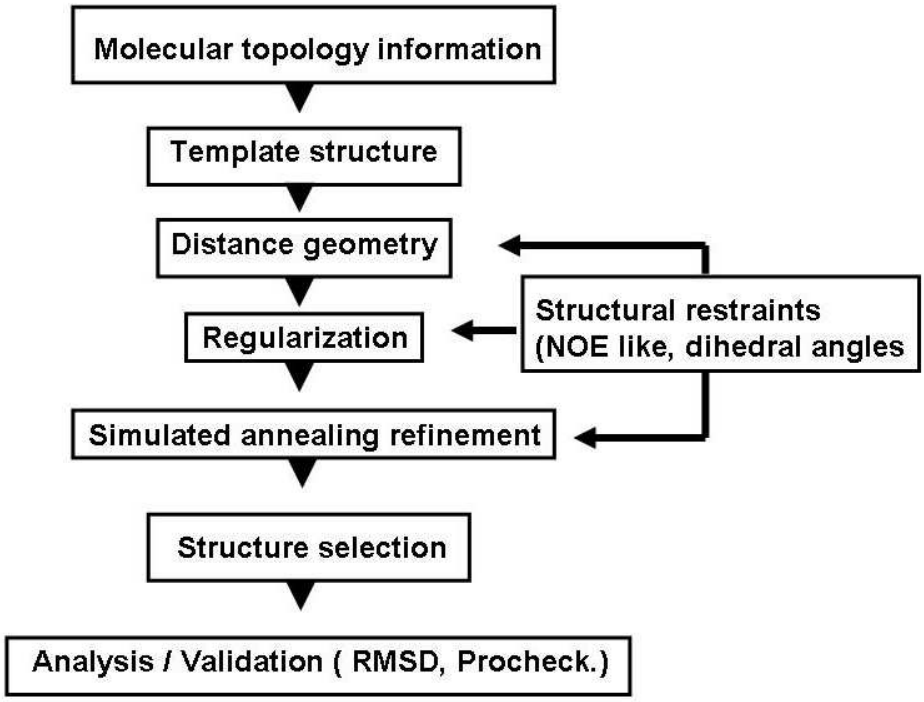

Figure 2.1: Overview of structure calculations

2.1.2 Distance Geometry

Many types of structural information (distances, J-coupling data, chemical cross-linking, neutron scattering, predicted secondary structures, etc.) can be conveniently expressed as intra- or intermolecular distances. In our case, we define NOE-like distances which are distance restraints imposed on our models and similar to NOE results. In the absence of J-coupling data, we define dihedral angle restraints. The distance geometry formalism allows these distance restraints to be assembled and three-dimensional structures consistent with them to be calculated. The distance geometry routines in X-PLOR begin by translating the bond lengths, bond angles, dihedral angles, improper angles, and van der Waals radii in the current molecular structure into a (sparse) matrix of distance bounds between the bonded atoms, atoms that are bonded to a common third atom, or atoms that are connected to each other through three bonds, as appropriate, using the equations in Crippen and Havel [42].

2.1.3 NOE-like and dihedral angle restraints

Experimental constraints can be added by the "NOE assign" statements and the "restraints dihedral" statements. These lists of restraints are automatically read, translated into distance constraints, and entered into the bounds matrix. They have the following forms:

NOE-like distance restraints:

The total distance restraint energy E_{NOE} is a sum over all distance restraints:

$$E_{NOE} = \sum_{restraints} e_{NOE}$$

where:

$$e_{NOE} = \min(1000, S)\Delta^2 \tag{2.1}$$

and Δ is defined as

$$\Delta = \begin{cases} R - (d + d_{plus}) & \text{if } d + d_{plus} < R \\ 0 & \text{if } d - d_{minus} < R < d + d_{plus} \\ d - d_{minus} - R & \text{if } R < d - d_{minus} \end{cases}$$

S is the scale factor, and d, d_{plus}, and d_{minus} are the average target distance and ranges and R is the distance between the two atoms. This defines the biharmonic function with a value of 0 for R larger than $d - d_{minus}$ and smaller than $d + d_{plus}$.

Dihedral Angle Restraints:

The functional form of the effective dihedral restraint energy E_{CDIH} is given by

$$E_{CDIH} = \sum well(modulo_{2\pi}(\phi - \phi_o), \Delta\phi)^2 \tag{2.2}$$

where the sum extends over all restrained dihedral angles and the square-well potential $well(a, b)$ is given by

$$well(a, b) = \begin{cases} a - b & \text{if } a > b \\ 0 & \text{if } -b < a < b \\ a + b & \text{if } a < -b \end{cases}$$

Plane Restraints:

Planar restraints have also been added in some simulations to maintain atoms in a common plane mimicking the phospholipid membrane. Restraints on an individual atom are based on its distance from a plane defined by the normal vector $\vec{z}$. An atom for which plane restraints are defined experiences restraints only in the direction of $\vec{z}$. Plane restraints are defined from the vector difference between present ($\vec{r}$) and reference ($\vec{r}_i^{ref}$) coordinates by:

$$E_{Plane} = \sum \left[\frac{\vec{z}}{|\vec{z}|} . \left(\vec{r}_i - \vec{r}_i^{ref} \right) \right]^2 \tag{2.3}$$

2.1.4 Structure determination

Structure determination of extended polypeptides or DNA/RNA double strands is usually underdetermined. In particular, the overall shape or bend of the molecule is a free parameter. Thus, neither distance geometry nor *ab initio* simulated annealing will produce unique structures. The problem can be avoided by including additional restraints

The first step of structure determination using X-PLOR involves providing the program with the information it needs about the molecular structure, NOE-like distance bounds, dihedral angle restraints. The molecular information of the macromolecule has to be generated using the all-hydrogen force fields "topallhdg.pro", "parallhdg.pro" for proteins.

Template Structure

The next step involves generation of a template coordinate set. The template coordinate set can be any conformation of the macromolecule with good local geometry and no nonbonded contacts. It can be generated by using most molecular modeling graphics programs or, preferably, by using the X-PLOR protocol described below. The purpose of the template coordinate set is to provide distance geometry information about the local geometry of the macromolecule.

The protocol we used initially places the atoms of the macromolecule along the x-axis, with y and z set to random numbers. The coordinates are then regularized using simulated annealing (See Fig. 2.2).

Generally, when too many covalent links are present, the structure may get entangled in a knot which will result in poor local geometry. In general, some experimentation may be required to find out if certain covalent links have to be removed; the goal is to obtain an energy below 1000 kcal/mole for the final step of minimization. We have taken advantage of the adaptive dynamics program (AMD) (See section 2.4) which allows us to easily manipulate structures and entangle the knots by setting the van der Waals force to zero in the process.

Substructure embedding

A family of embedded substructures is produced using distance geometry. The substructures are regularized after embedding using minimization against the DG energy term. (This Distance Geometry term is still another harmonic potential meant to maintain a given distance between a lower and an upper bound). Covalent links should be treated in the same fashion as in the prior template generation. The removed covalent links are reintroduced as distance restraints.

SA-Regularization of DG-Structures

Embedded distance geometry (substructure) coordinates require extensive regularization. The next protocol uses template fitting followed by simulated annealing to regularize the coordinates. The protocol is close in spirit to the one published by Nilges et al [43]. The starting coordinates have to be defined for at least three atoms in each residue. Covalent links are now treated as "real" bonds in this protocol, in contrast to template generation.

Simulated Annealing Refinement

Structures are finally refined using a protocol of the slow-cooling type reminiscent of the protocol used in crystallographic refinement with softening of the van der Waals repulsions. This enables atoms to move through each other.

In both SA regularization and refinement, the simulated annealing starting temperature and the

duration of the cooling stage are adjustable parameters.

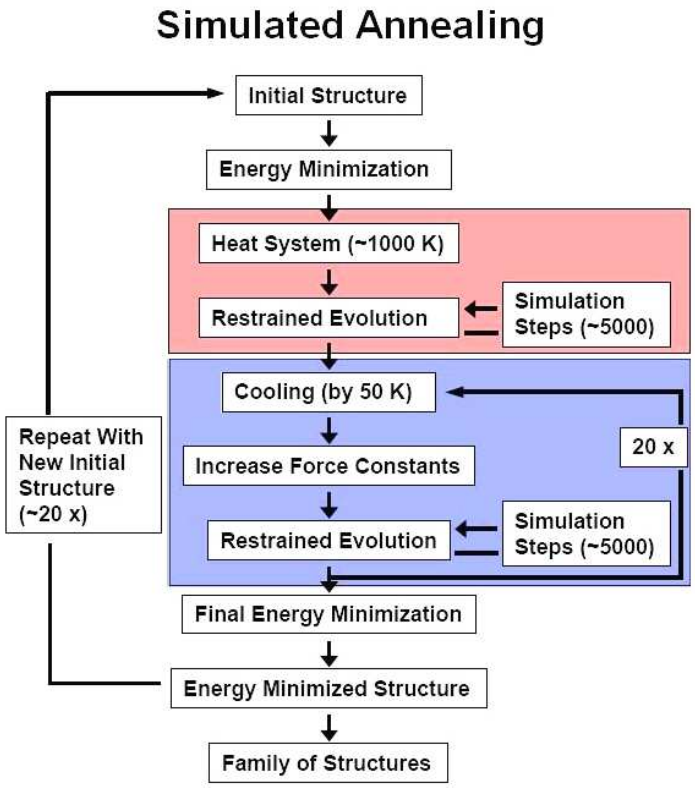

Figure 2.2: Flowchart for simulated annealing

2.2 Homology modelling

We have also used a more standard method for structure prediction when a homology between the protein of interest and other proteins of known 3D structure exists:

2.2.1 Introduction to Homology modelling

During evolution, sequence changes much faster than structure. It is possible to identify the 3D-structure by looking at a molecule with some sequence identity. Fig. 2.3 shows how much sequence identity is needed with a certain number of aligned residues to reach the safe homology modeling zone. For a sequence of 100 residues, for example, a sequence identity of 40% is sufficient for structure prediction. When the sequence identity falls in the safe homology modeling zone, we can assume that the 3D-structure of both sequences is the same.

2.2.2 Classical steps in Homology modeling

The known structure is called the template, the unknown structure is called the target. Homology modeling of the target structure can be done in several steps:

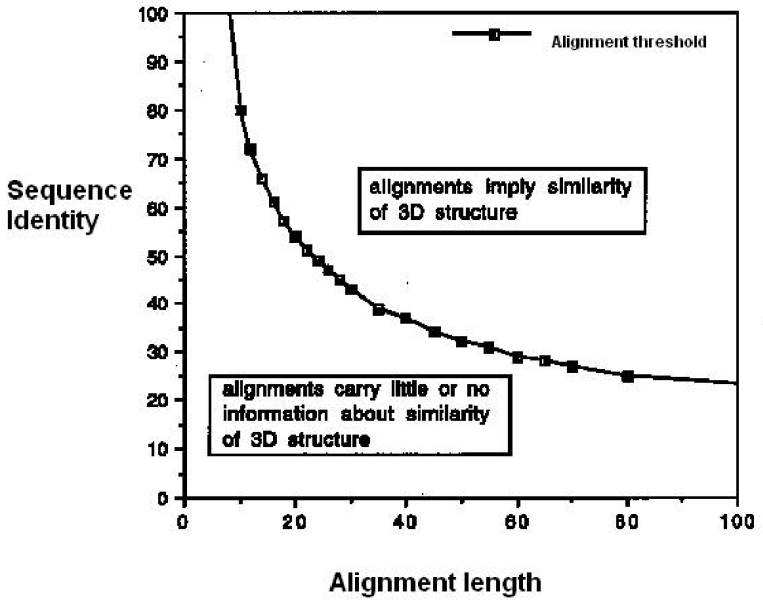

Figure 2.3: Threshold for homology modelling: we can assume that two proteins share the same 3D-structure when their sequence identity falls in the safe homology zone, the upper part of the picture (From work by Sander and Schneider [44]).

2.2.2.1 Template recognition and initial alignment

In this step you compare the sequence of the unknown structure with all the known structures stored in the Protein Data Bank (PDB). A search with BLAST against this database will give a list of known protein structures that match the sequence. If BLAST cannot find a template a more sophisticated technique might be necessary to identify the structure of a molecule. BLAST uses a residue exchange matrix to define a score for every hit. Residues that are easily exchanged (for example Ile to Leu) get a better score than residues that have different properties (for example Glu to Trp). Conserved residues with a specific function get the best score (for example Cys to Cys). Every hit is scored using this matrix and BLAST will provide a list of possible templates for the uknown structure. To make the best initial alignment, BLAST uses an alignment-matrix based on the residue exchange matrix and adds extra penalties for opening and extension of a gap between residues. In practice the target-sequence is sent to a BLAST server, which searches the PDB to obtain a list of possible templates and their alignments. Subsequently the best hit has to be chosen, which is not necessarily the first one. One has to keep in mind the resolution, missing parts, different states of action and possible ligands of the molecules in doing so.

2.2.2.2 Alignment correction

It is possible that the alignment has to be corrected. A change of Ala to Glu is possible but unlikely to happen in a hydrophobic core, so this Ala and Glu cannot be aligned. Using a multiple sequence alignment program (ClustalW) the residues and properties that have to be conserved can be found.

By looking at the template structure it will become clear which residues are in the core and are less likely to be changed than the residues at the outside. Insertions and deletions can be made in widely divergent parts of the molecule and a multiple sequence alignment can be helpful to find these places. Gaps have to be shifted around until they are as small as possible. In Fig. 2.4 is shown that after a deletion of 3 residues a big gap occurs in the red structure, which was the best alignment. After shifting several residues, the gap is much smaller (blue structure) and more likely to be correct. Correction of the alignment is typically done by hand.

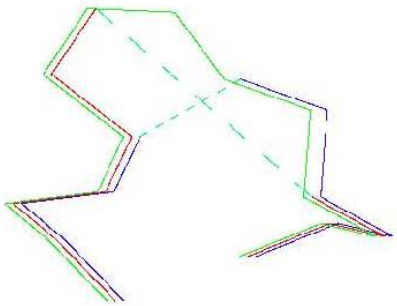

Figure 2.4: Template structure (green) with the best aligned target (red) with a large gap, and the target after shifting several residues (blue). The gap is much smaller now.

2.2.2.3 Backbone generation

When the alignment is correct, the backbone of the target can be created. The coordinates of the template-backbone are copied to the target. When the residues are identical, the side-chain coordinates are also copied. Because a PDB-file can always contain some errors, it can be useful to make use of multiple templates.

2.2.2.4 Loop modeling

Often the alignment will contain gaps as a result of deletions and insertions. When the target sequence contains a gap, one can simply delete the corresponding residues in the template. This creates a hole in the model, this has already been discussed in previous step. When there is an insertion in the target, the template will contain a gap and there are no backbone coordinates known for these residues in the model. The backbone from the template has to be cut to insert these residues. Such large changes cannot be modeled in secondary structure elements and therefore have to be placed in loops and strands. Surface loops are, however, flexible and difficult to predict. One way to handle loops is to take some residues before and after the insertion as "anchor" residues and search the PDB for loops with the same anchor-residues. The best loop is simply copied in the model. In the MODELLER program, that we have used (See section 2.2.3), loop modelling is treated with special care.

2.2.2.5 Side-chain modeling

The next step is to add side-chains to the backbone of the model. Conserved residues were already copied completely. The torsion angle between C-α and C-β of the other residues can also be copied

to the model because these rotamers tend to be conserved in similar proteins. It is also possible to predict the rotamer because many backbone configurations strongly prefer a specific rotamer as shown in Fig. 2.5, in the case of a tyrosine residue. There are libraries based upon the backbone of the residues flanking the residue of interest. By using these libraries the best rotamer can be predicted.

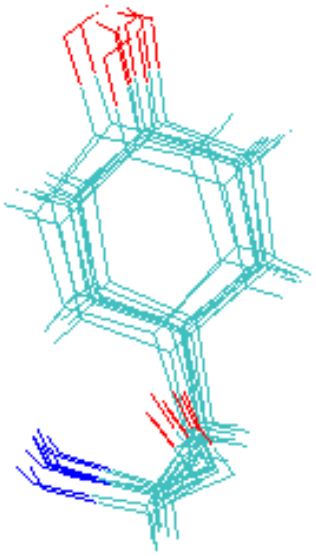

Figure 2.5: Prefered rotamers of tyrosin exemplified with different positions of TYR52 in the 10 NMR models of the first Metal Binding Domain of the Menkes ATPase (PDB code 1KVJ).

2.2.2.6 Model optimization

The model has to be optimized because the changed side-chains can effect the backbone, and a changed backbone will have effect on the predicted rotamers. Optimization can be done by performing refinements using Molecular Dynamics simulations of the model. The model is placed in a force-field and the movements of the molecules are followed in time, this mimics the folding of the protein. The big errors like bumps will be removed but new smaller errors can be introduced. The calculated energy should be as low as possible.

2.2.2.7 Model validation

Every model contains errors. The model with the lowest forcefield energy might still be folded completely wrong. That is why the model should be checked for bumps and if the bond angles, torsion angles and bond lenghts are within normal ranges. Other properties, like the distribution of polar/apolar residues, can be compared with real structures. This can be done by using Procheck, for example. The output can help in the identification of errors in the model. When an error occurs far away from the active site, it does not have to be bad. But when an error occurs in the active site, one should reconsider the template and/or alignment.

2.2.3 The program MODELLER

MODELLER [45] is a computer program that models three-dimensional structures of proteins and their assemblies by satisfaction of spatial restraints. MODELLER is most frequently used for homology or comparative protein structure modeling: The user provides an alignment of a sequence to be modeled with known related structures and MODELLER will automatically calculate a model with all non-hydrogen atoms. More generally, the inputs to the program are restraints on the spatial structure of the amino acid sequence(s) and ligands to be modeled. The output is a 3D structure that satisfies these restraints as well as possible. Restraints can in principle be derived from a number of different sources. These include related protein structures (comparative modeling), NMR experiments (NMR refinement), rules of secondary structure packing (combinatorial modeling)... The restraints can operate on distances, angles, dihedral angles, pairs of dihedral angles and some other spatial features defined by atoms or pseudo atoms. Presently, MODELLER automatically derives the restraints only from the known related structures and their alignment with the target sequence. A 3D model is obtained by optimization of a molecular probability density function (pdf). The molecular pdf for comparative modeling is optimized with the variable target function procedure in Cartesian space that employs methods of conjugate gradients and molecular dynamics with simulated annealing. MODELLER can also perform multiple comparison of protein sequences and/or structures, clustering of proteins, and searching of sequence databases.

2.3 Classical Mechanics and Dynamics

The 3D structure of the molecules being known, we are now faced with the significant complexity of the structures and interactions that we want to simulate for a realistic model. Two possible strategies are generally used: we can either increase the processing power (e.g. use costly parallel supercomputers), or use simplified representations of the geometry or of the dynamics of the involved molecules. Frequently, these simplification methods involve representations as mechanical models or in reduced coordinates (e.g. modelling the molecule as an articulated body), where subsets of atoms are replaced by idealized structures [46] [47], or performing normal-mode or principal components analysis in order to determine the essential dynamics of the system. Because they contain fewer degrees of freedom, these simplified representations allow us to accelerate the computation of the molecular dynamics, and facilitate the study of the molecular interactions. In the following, I will describe how a molecular mechanics model is built with atoms represented as charged spheres and covalent bonds by springs. This mechanical model can consequently be studied using the classical equations of motion of mechanics.

2.3.1 Newton's Second Law

Molecular Dynamics simulations are based on Newton's second law, the equation of motion [48],[49]:

$$\vec{F_i} = m_i.\vec{a}_i = m_i.\frac{d\vec{v}_i}{dt} \tag{2.4}$$

It describes the motion of a particle of mass m_i along the coordinate x_i with F_i being the force on m_i in that direction. This is used to calculate the motion of a finite number of atoms or molecules, respectively, under the influence of a force field that describes the interactions inside the system with

a potential energy function, $V(\vec{x})$, where $\vec{x}$ corresponds to the coordinates of all atoms in the system. The relationship of the potential energy function and Newton's second law is given by

$$\vec{F}(x_i) = -\nabla_i V(x_i), \tag{2.5}$$

with $\vec{F}(x_i)$ being the force acting on a particle due to a potential, $V(\vec{x})$. Combining these two equaions gives

$$\frac{dV(\vec{x})}{dx_i} = -m_i.\frac{d^2x_i}{dt^2}, \tag{2.6}$$

which relates the derivative of the potential energy to the changes of the atomic coordinates in time. As the potential energy is a complex multidimensional function this equation can only be solved numerically with some approximations.

With the acceleration being $a = -\frac{1}{m}.\frac{dV}{dx}$ we can then calculate the changes of the system in time by just knowing (i) the potential energy $V(\vec{x})$, (ii) initial coordinates x_{i0} and (iii) an initial distribution of velocities, v_{i0}. Thus, this method is deterministic, meaning we can predict the state of the system at any point of time in the future or the past.

The initial distribution of velocities is usually randomly chosen from a Gaussian or Maxwell-Boltzmann distribution [49], which gives the probability of atom i having the velocity in the direction of x at the temperature T by:

$$p(v_{i,x}) = \left(\frac{m_i}{2\pi k_b T}\right)^{\frac{1}{2}} .\exp\left(-\frac{1}{2}\frac{m_i v_{i,x}^2}{k_b T}\right).$$

Velocities are then corrected so that the overall momentum of the system equals a zero vector:

$$P = \sum_{n=1}^{N} m_i.\vec{v_i} = \vec{0}.$$

2.3.2 Integration Algorithms

The solution of the equation of motion given above is a rather simple one which is only sufficiently good over a very short period of time, in which the velocities and accelerations can be regarded as constant. So algorithms were introduced repeatedly performing small time steps, thus propagating the system's properties (positions, velocities and accelerations) in time. Time steps are typically chosen in the range of 1 fs [49]. It is necessary to use such a small time step, as many molecular processes occur in such small periods of time that they cannot be resolved with larger time steps. A time series of coordinate sets calculated this way is referred to as a trajectory and a single coordinate set as a frame.

2.3.2.1 Verlet Algorithm

All algorithms assume that the system's properties can be approximated by a Taylor series expansion [50]:

$$\begin{aligned}
\vec{x}\,(t+\delta t) &= \vec{x}\,(t)+\delta t.\vec{v}\,(t)+\frac{1}{2}\delta t^2.\vec{a}\,(t)+....\\
\vec{v}\,(t+\delta t) &= \vec{v}\,(t)+\delta t.\vec{a}\,(t)+\frac{1}{2}\delta t^2.\vec{b}\,(t)+....\\
\vec{a}\,(t+\delta t) &= \vec{a}\,(t)+\delta t.\vec{b}\,(t)+\frac{1}{2}\delta t^2.\vec{c}\,(t)+....
\end{aligned}$$

with $\vec{x}$,$\vec{v}$ and a being the positions, the velocities and the accelerations of the system.The series expansion is usually truncated after the quadratic term. Probably the most widely used algorithm for integrating the equations of motion in MD simulations is the Verlet algorithm (1967)[48],[49]. It can be derived by simply summing the Taylor expressions for the coordinates at the time $(t+\delta t)$ and $(t-\delta t)$:

$$\begin{aligned}
\vec{x}\,(t+\delta t) &= \vec{x}\,(t)+\delta t.\vec{v}\,(t)+\frac{1}{2}\delta t^2.\vec{a}\,(t)+....\\
\vec{x}\,(t-\delta t) &= \vec{x}\,(t)-\delta t.\vec{v}\,(t)+\frac{1}{2}\delta t^2.\vec{a}\,(t)-....\\
\Rightarrow \vec{x}\,(t+\delta t) &= 2\vec{x}\,(t)-\vec{x}\,(t-\delta t)+\delta t^2.\vec{a}\,(t)\,.
\end{aligned}$$

Thus, it uses the position $\vec{x}\,(t)$ and acceleration $\vec{a}\,(t)$ at time t and the positions from the previous step $\vec{x}\,(t-\delta t)$ to calculate new positions $\vec{x}\,(t+\delta t)$. In this algorithm velocities are not explicitly calculated but can be obtained in several ways. One is to calculate mean velocities between the positions $\vec{x}\,(t+\delta t)$ and $\vec{x}\,(t-\delta t)$.

$$\vec{v}\,(t)=\frac{1}{2\delta t}.[\vec{x}\,(t+\delta t)-\vec{x}\,(t-\delta t)]$$

The advantages of this algorithm are that it is straightforward and has modest storage requirements, comprising only two sets of positions [$\vec{x}\,(t)$ and $\vec{x}\,(t-\delta t)$] and the accelerations $\vec{a}(t)$. The disadvantage, however, is its moderate precision, because the positions are obtained by adding a small term $[\delta t^2\vec{a}(t)]$ to the difference of two much larger terms $[2\vec{x}(t)-\vec{x}\,(t-\delta t)]$. This results in rounding errors due to numerical limitations of the computer.

Furthermore, this is obviously not a self-starting algorithm. New positions $\vec{x}\,(t+\delta t)$ are obtained from the current positions $\vec{x}(t]$ and the positions at the previous step $\vec{x}\,(t-\delta t)$. So at $t=0$ there are no positions for $(t-\delta t)$ and therefore it is necessary to provide another way to calculate them. One way is to use the Taylor expansion truncated after the first term:

$$\begin{aligned}
\vec{x}\,(t-\delta t) &= \vec{x}\,(t)-\delta t.\vec{v}\,(t)+....\\
\Rightarrow \vec{x}\,(-\delta t) &= \vec{x}\,(0)-\delta t.\vec{v}\,(0)
\end{aligned}$$

2.3.2.2 Leap-Frog Algorithm

There are several variations of the Verlet algorithm trying to avoid its disadvantages. One example is the leap-frog algorithm [49]. It uses the following equations:

$$\begin{aligned}\vec{v}\left(t+\frac{1}{2}\delta t\right) &= \vec{v}\left(t-\frac{1}{2}\delta t\right)+\delta t.\vec{a}(t)\\ \vec{x}\left(t+\delta t\right) &= \vec{x}\left(t\right)+\delta t.\vec{v}\left(t+\frac{1}{2}\delta t\right),\end{aligned}$$

where $\vec{a}(t)$ is obtained using

$$\vec{a}(t)=-\frac{1}{m}.\frac{dV}{d\vec{x}}.$$

First, the velocities $\vec{v}\left(t+\frac{1}{2}\delta t\right)$ are calculated from the velocities at $t-\delta t$ and the accelerations $\vec{a}(t)$. Then the positions $\vec{x}\left(t+\delta t\right)$ are deduced from the velocities just calculated and the positions at time t. In this way the velocities first 'leap-frog' over the positions and then the positions leap over the velocities. The leap-frog algorithm's advantages over the Verlet algorithm are the inclusion of the explicit velocities and the lack of the need to calculate the differences between large numbers. An obvious disadvantage, however, is that the positions and velocities are not synchronized. This means it is not possible to calculate the contribution of the kinetic energy (from the velocities) and the potential energy (from the positions) to the total energy simultaneously.

2.3.2.3 Langevin Dynamics (LD) Simulation

The Langevin equation is a stochastic differential equation in which two force terms have been added to Newton's second law to approximate the effects of neglected degrees of freedom. One term represents a frictional force, the other a random force $\vec{R}$. For example, the effects of solvent molecules not explicitly present in the system being simulated would be approximated in terms of a frictional drag on the solute as well as random kicks associated with the thermal motions of the solvent molecules. Since friction opposes motion, the first additional force is proportional to the particle's velocity and oppositely directed. Langevin's equation for the motion of atom i is:

$$\vec{F_i}-\gamma_i\vec{v_i}+\vec{R_i}(t)=m_i\vec{a}_i,$$

where $\vec{F_i}$ is still the sum of all forces exerted on atom i by other atoms explicitly present in the system. This equation is often expressed in terms of the 'collision frequency' $\zeta=\gamma/m$.

The friction coefficient is related to the fluctuations of the random force by the fluctuation-dissipation theorem:

$$\langle\vec{R_i}(t)\rangle=0,$$

$$\int\langle\vec{R_i}(0)\cdot\vec{R_i}(t)\rangle dt=6k_BT\gamma_i.$$

In simulations it is often assumed that the random force is completely uncorrelated at different times. That is, the above equation takes the form:

$$\langle\vec{R_i}(t)\cdot\vec{R_i}(t')\rangle=6k_BT\gamma_i\delta(t-t').$$

The temperature of the system being simulated is maintained via this relationship between $\vec{R}(t)$ and γ.

The jostling of a solute by solvent can expedite barrier crossing, and hence Langevin dynamics can search conformations better than Newtonian molecular dynamics ($\gamma = 0$).

2.4 Adaptive Molecular Dynamics

2.4.1 Introduction

In the previous simple mechanical model, the only way to reduce the computational cost of a calculation knowing that some part of the system is "less" important in the mechanism under study is to fix the corresponding atoms. Thus, we must have some prior knowledge about the function of the system and we completely neglect the dynamics of this "less"important part and its possible interactions with the reaction site. In other words, there was no method that automatically determines which parts of the molecule must be precisely simulated, and which parts can be simplified without affecting the study of the molecular interaction. Researchers in Stephane Redon's team (Nano-D, at INRIA Grenoble) have recently introduced adaptive torsion-angle quasi-statics and Adaptive Molecular Dynamics (AMD), a general technique to rigorously and automatically determine the most important regions in a simulation of molecules represented as articulated bodies. At each time step, the adaptive algorithm determines the set of joints that should be simulated in order to best approximate the motion that would be obtained if all degrees of freedom were simulated, based on the current state of the simulation and user-defined precision or time constraints. They built on previous research on adaptive articulated-body simulation [51] and proposed novel data structures and algorithms for adaptive update of molecular forces and energies.

2.4.2 Divide and Conquer Algorithm

The starting point for the study of the dynamics of an articulated body is a " Divide-And-Conquer Algorithm " proposed by Roy Featherstone [52]. Featherstone recursively defines an articulated body by assembling two (rigid or articulated) bodies together. A complete articulated body is thus represented by a binary tree: the root node describes the whole articulated body, while each leaf node is a rigid body with a set of handles, i.e. locations attached to some other rigid bodies. Let C be an articulated body with m handles, Featherstone defines the articulated- body equation:

$$\begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_m \end{bmatrix} = \begin{bmatrix} \Phi_1 & \Phi_{12} & \cdots & \Phi_{1m} \\ \Phi_{21} & \Phi_2 & \cdots & \Phi_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ \Phi_{m1} & \Phi_{m2} & \cdots & \Phi_{mn} \end{bmatrix} \begin{bmatrix} f_1 \\ f_2 \\ \vdots \\ f_m \end{bmatrix} + \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_m \end{bmatrix} \tag{2.7}$$

where a_i is the spatial acceleration of handle i, f_i is the spatial force applied to handle i, b_i the bias acceleration of handle i, Φ_i is the inverse articulated body inertia of handle i and Φ_{ij} the cross-coupling

inverse inertia between handles i and j. This equation is the equivalent of the classical equation of motion:

$$a_i = \frac{1}{m_i}.f_i$$

We thus consider the molecules as articulated bodies: every rigid body is one atom or a group of atoms. Joints (handles) between rigid bodies are covalent bonds around which rotation is possible. The dynamics is calculated in the dihedral angle space constituting the system like all the ϕ, ψ and χ angles in proteins. If at a given time, parts of the system are considered as rigid, they form a subtree of the complete assembly tree where forces need not be recalculated. The rest is constituted of active joints forming an active region (See Fig. 2.6).

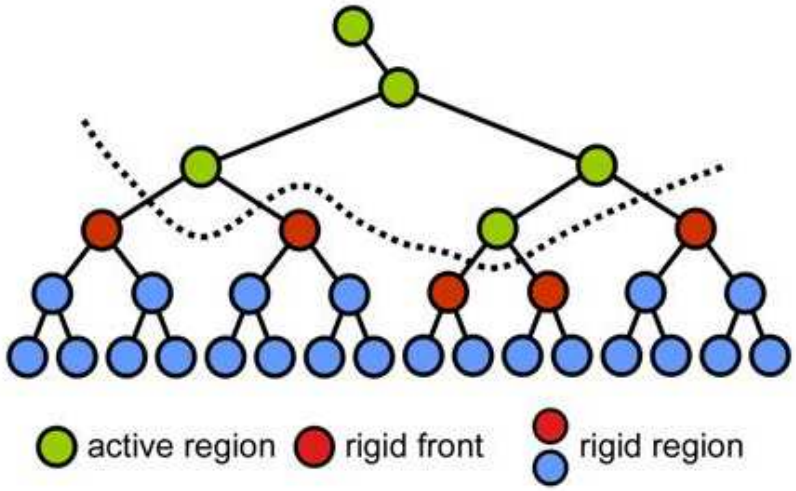

Figure 2.6: Active and rigid regions. The figure corresponds to 5 active joints

The usefulness of the tree representation will now be made clear: The Featherstone algorithm is accomplished in two stages: i) 1 stage from bottom to top, from the leaves to the top where articulated body coefficients b_i and Φ_{ij} for each composite object C is calculated from ithe coefficients of its sons A and B. ii) 1 stage from top to bottom to yield joint accelerations $\ddot{q}_i$ and forces. The joint accelerations $\ddot{q}_i$ are the second derivatives of the motion variables such as $\ddot{\phi}_i$ if we are interested to the movement around dihedral angle ϕ_i Now, a theorem states that the sum of the squares of the accelerations :

$$\mathcal{A} = \sum_{\mathrm{i}} \ddot{\mathrm{q}}_{\mathrm{i}}^2$$

of the joints in one node of the tree can be calculated without knowing the values of the individual $\ddot{q}_i$ of its children nodes. This means that the algorithm can decide by its own to partition the tree in active and rigid regions, the latter being those where the metrics A is the lowest and thus the dynamics may be considered as "less" important for the mechanism under study.

I participated in this work which was published in Bioinformatics in 2007 [53]. I have used AMD many times during my thesis both for the interactive visualization of molecules and in attempts to study biological mechanisms (See section 3.1.5). It is indeed very easy with AMD to apply external

forces (f_i) on the system with the computer mouse and let the system relax to a new equilibrium positiuon.

2.5 Statistical Mechanics

2.5.1 Introduction to Statistical Mechanics

MD simulations provide information at the microscopic level. Statistical mechanics is then required to convert this microscopic information to macroscopic observables such as pressure, energy, heat capacities, etc. Statistical mechanics relates these macroscopic observables to the distribution of molecular positions and motions. Therefore, time independent statistical averages are introduced. For a better understanding some definitions are reviewed here [54]:

2.5.2 Definitions

The mechanical or microscopic state of a system is defined by the atomic positions x_i and the momenta $p_i = m_i v_i$. They can be considered as a multidimensional space with 6N coordinates, for which they both contribute 3N coordinates. This space is called phase space.

The thermodynamic or macroscopic state of a system is defined by a set of parameters that completely describes all thermodynamic properties of the system. An example would be the temperature T, the pressure P, and the number of particles N. All other properties can be derived from the fundamental thermodynamic equations.

An ensemble is the collection of all possible systems which have different microscopic states but have the same macroscopic or thermodynamic state. Ensembles can be defined by fixed thermodynamic properties as already stated before. Examples for ensembles with different characteristics are: NVE, NVT, NPT, μVT, (E = total energy, P = pressure, V = volume, μ = chemical potential)

2.5.3 Ensemble Averages and Time Averages

In an experiment one examines a macroscopic sample with an enormously high number of atoms or molecules respectively. So the measured thermodynamic properties reflect an extremely large number of different conformations of the system, representing a subset of the ensemble. We have to say subset, because an ensemble is the complete collection of microscopic systems and a macroscopic sample can only consist of a finite number of systems. A sufficiently big sample, however, can be seen as good approximation to an ensemble. That is why statistical mechanics defines averages corresponding to experimentally measured thermodynamic properties as ensemble averages [54]. The ensemble average is given by:

$$\langle A \rangle_{\text{ensemble}} = \iint d\vec{p}^N d\vec{x}^N A\left(\vec{p}^N, \vec{x}^N\right) \rho\left(\vec{p}^N, \vec{x}^N\right), \tag{2.8}$$

where $\langle A \rangle$ is the measured observable, which is stated as a function of the momenta pi and the positions $\vec{x}_i$. Quantity $\rho\left(\vec{p}^N, \vec{x}^N\right)$ is the probability density for the ensemble and the integration is performed over all momenta and positions of the system $d\vec{p}^N$, $d\vec{x}^N$. So, the ensemble average is the average value of an observable weighted by its probability. This integral is extremely difficult to calculate as it involves calculating all possible states of the system. In an MD simulation an extremely

large number of conformations is generated sequentially in time. To calculate an ensemble average the simulation has to cover all possible conformations corresponding to the ensemble, at which the simulation takes place. The time average is given by

$$\langle A \rangle_{\text{time}} = \lim_{\tau \to \infty} \frac{1}{\tau} \int_{t=0}^{\tau} A\left(\vec{p}^N \vec{x}^N\right) dt, \tag{2.9}$$

where τ is the simulation time.

This expression is well approximated by an average over a simulation performed over a sufficiently long period of time and so representing a sufficient amount of phase space:

$$\langle A \rangle_{\text{time}} = \lim_{\tau \to \infty} \frac{1}{\tau} \int_{t=0}^{\tau} A\left(\vec{p}^N, \vec{x}^N\right) dt = \frac{1}{M} \sum_{i=1}^{M} A\left(\vec{p}_M^N, \vec{x}_M^N\right) \tag{2.10}$$

where M is the number of frames and $A\left(\vec{p}_M^N, \vec{x}_M^N\right)$ the values of the observable A in frame M. The idea is based on the Ergodic Hypothesis [54], one of the most fundamental axioms of statistical mechanics, which states that the time average equals the ensemble average:

$$\langle A \rangle_{\text{ensemble}} = \langle A \rangle_{\text{time}} \tag{2.11}$$

The idea, as already indicated above, is to simulate the system for a relatively long time, so that it will pass through an extremely high number of conformations, which can then be referred to as a representative subset of an ensemble.

2.6 Force Fields

2.6.1 General Features of Molecular Mechanics Force Fields

Force fields enable the potential energy of a molecular system V to be calculated rapidly. The energy is a function of the atomic positions of all the atoms in the system which are usually expressed in term of Cartesian coordinates. Unlike quantum mechanical methods that deal with the electrons in a system, force field techniques calculate the energy of a system only as a function of the nuclear positions. This is legitimated by the Born-Oppenheimer-Approximation [55]. Thus, Molecular Mechanics is invariably used to perform calculations on systems containing a significant number of atoms which would bring enormous quantum mechanical calculations with them.

A typical force field represents each atom in the system as a single point and energies as a sum of two-, three-, and four-atom interactions such as bond stretching and angle bending. Although, simple functions (e.g. Hooke's Law) are used to describe these interactions, the force field can work quite well. The potential energy of a certain interaction is described by an equation which involves the positions of the particles and some parameters (e.g force constants or reference values) which have been determined experimentally or by quantum mechanical calculations.

Several types of force fields exist. Two of those may use an identical functional form yet have very different parameters and thus bring about different energies for the same system. Moreover, force fields with the same functional form but different parameters, and force fields with different functional forms, may give close results. A force field should be considered as a single entity; it

doesn't need to be correct to divide the energy into its individual components or even to take some of the parameters from one force field and mix them with parameters from another one.

An important point that one shouldn't forget is that no 'correct' form for a force field exists. If one functional form performs better than another, that form will be favored. Most of the force fields commonly used do have a very similar form but it should always be kept in mind that there may be better functional forms, particularly when developing a force field for new classes of molecules. Molecular mechanics force fields are often a compromise between accuracy and computational efficiency; the most accurate ones may often be unsatisfactory for efficient computation. As the performance of computers increases, it becomes possible to incorporate more sophiticated models.

A concept that is common to most force fields is that of an atom type. For a quantum mechanics calculation it is usually necessary to specify the charge of the nuclei, together with the geometry of the system and the overall charge and spin multiplicity. For a force field calculation, however, the overall charge and spin multiplicity are not explicitly required, but it is usually necessary to assign an atom type to each atom in the system. This contains information about its hybridization state and sometimes the local environment. For example, it is necessary to distinguish between carbon atoms which adopt a tetrahedral geometry (sp^3-hybridized), which are trigonal (sp^2- hybrdized) and carbons which are linear (sp-hybridized). The corresponding parameters are expressed in terms of these atom types, so that the reference angle Θ_0 for a tetrahedral carbon atom would be about 109.5 degree and that for a trigonal carbon near 120. For example, the MM2 [56], MM3 [57] and MM4 [58] force fields of Allinger and co-workers, that are widely used for calculations on "small" molecules, distinguish the following types of carbon atoms: sp^3, sp^2, sp, carbonyl, cyclopropane, radical, cyclopropene and carbonium ion. The value of the potential energy V is calculated as a sum of internal or bonded terms, which describe the bonds, angles and bond rotations in a molecule, and a sum of external or non-bonded terms, which account for interactions between non-bonded atoms or atoms separated by three or more covalent bonds. So it is:

$$V(\vec{x}) = V_{\text{bonded}}(\vec{x}) + V_{\text{nonbonded}}(\vec{x}), \tag{2.12}$$

Let us now discuss the individual contributions to a molecular mechanics force field, giving a selection of the various functional forms that are in common use. We shall then have a look at the CHARMM force field which is used for our calculations.

2.6.2 Bonded terms

2.6.2.1 Bond Streching

The potential energy curve (Morse potential) for a typical bond has the shape shown in Fig. 2.7 (Solid line). This potential has the form:

$$V_B(t) = D_e\{1 - exp[-a(l - l_0)]\}^2 \tag{2.13}$$

D_e is the depth of the potential energy minimum and $a = \omega.\sqrt{\frac{\mu}{2D_e}}$, where μ is the reduced mass and ω is the frequency of the bond vibration. The frequency ω is related to the streching constant of the bond k_l, by $\omega = \sqrt{\frac{k_l}{\mu}}$, where k_l determines the strength of the bond. The length l_0 is the

reference bond length. It is the value that the bond adopts when all other terms in the force field are zero. Both l_0 and k_l are specific for each pair of bound atom. Values of k_l are often evaluated from experimental data such as infrared stretching frequencies or from quantum mechanical calculations. Values of l_0 can be inferred from high resolution crystal structures or microwave spectroscopy data. The Morse potential is not usually used in molecular mechanics force fields. It is computationally demanding the curve describes a wide range of behavior, from the strong equilibrium to dissociation. Normally, this is not necessary for Molecular Mechanics calculations where we are more interested in slight deviations of bonds from their equilibrium values. Consequently, simpler expressions are often used. The most elementary approach is to use a Hooke's Law formula in which the energy varies with the square of the displacement from the reference bond length l_0:

$$V_B\left(l\right) = \frac{1}{2} k_l . \left(l - l_0\right)^2 \tag{2.14}$$

The Hooke's Law functional form is a reasonable approximation to the shape of the potential energy curve at the bottom of the potential well, at distances that correspond to bonding in ground-state molecules. It is less accurate away from equilibrium (Fig. 2.7 dotted line).

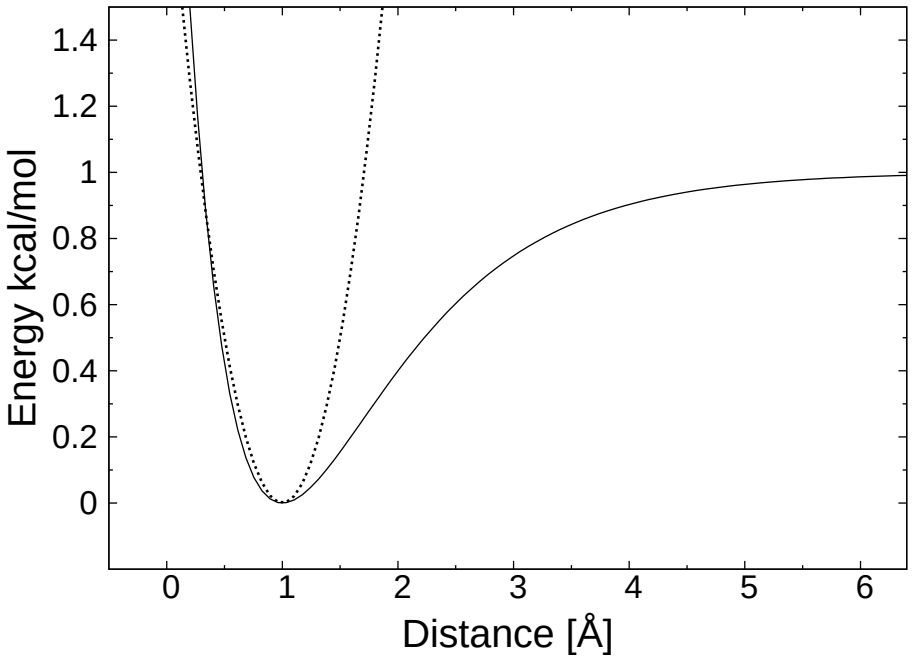

Figure 2.7: Comparison of the simple harmonic potential (Hooke's Law) with the Morse curve: The Morse potential follows equation 2.13 with D_e=1.0 kcal/mol, a = 1.0 and l_0 = 1.0 Å and is shown as solid line while the harmonic potential $h(r) = 2.0(r - 1.0)^2$ is represented as dotted line.

To approximate the Morse curve more accurately, cubic and higher terms can be included which give a better model closer to the equilibrium structure than the quadratic form but also create a potential passing through maxima further away from l_0. This can lead to a catastrophic lengthening of bonds. Researchers in the team have used such a cubic potential to better approximate the energy of the S-Cu and S-Hg bonds [59] added to the CHARMM force field. A plot of energy versus r_1 =

S-Cu(I) distance is shown in Fig. 2.8. A comparison is done with a quantum chemical calculation (*ab initio*) run with the program Gaussian. There is a good agreement between the *ab initio* and CHARMM energies for values of r extending plus or minus 0.3 Å away from the minimum. The discrepancy between the *ab initio* and CHARMM energies for large r is due to the fact that either the corrective anharmonic term in r^3 is too strong (and negative : $(r_0 - r) \leq 0$ for large r) or higher order anharmonic terms are missing in the potential.

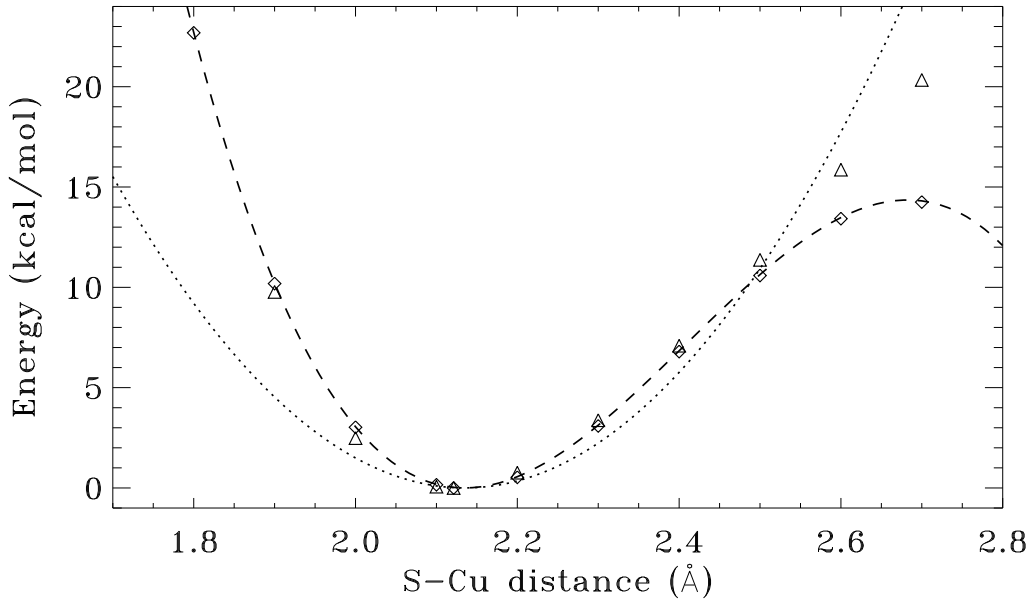

Figure 2.8: Plot of energy versus S-Cu(I) distance. The points corresponding to the *ab initio* and CHARMM calculations, evenly spaced every 0.1 Å , are shown as triangles and diamonds respectively. An additional point at r=2.121 Å corresponds to the minimum of the *ab initio* energy. The dashed curve close to the CHARMM points represents the fitted potential in $r^2 + r^3$ and a dotted curve represents the contribution of the harmonic term only to this potential.

2.6.2.2 Angle Bending

The deviation of valence angles from their reference values (Fig. 2.9) is also frequently described using a harmonic potential:

$$V_A(\theta) = \frac{1}{2} k_\theta (\theta - \theta_0)^2 \tag{2.15}$$

The force constant k_θ and the reference value θ_0 depend on the chemical type of atoms constituting the angle. Rather less energy is required to distort an angle away from equilibrium than to stretch or compress a bond, and thus force constants are smaller here. As with the bond-stretching terms, the accuracy of the force field can be improved by the incorporation of higher-order terms. These two terms, the bond-stretching and angle-bending, describe the deviation from an ideal geometry; effectively, they are penalty functions and the sum of them should be close to zero in a perfectly

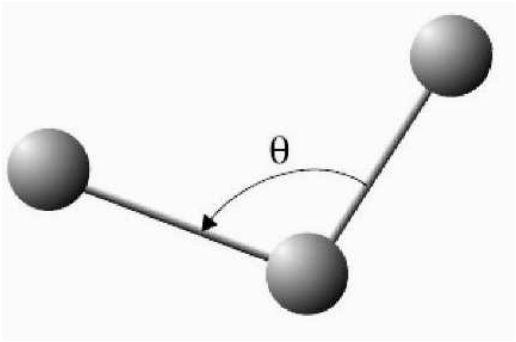

Figure 2.9: Bond Angle θ

optimized structure. Most of the variation in structure and relative energies is due to torsional and non-bonded contributions.

2.6.2.3 Torsional Terms

This term represents the torsion angle potential function which models the presence of steric barriers between atoms separated by three covalent bonds. The existence of barriers to rotation about chemical bonds is fundamental to understanding the structural properties of molecules and conformational analysis. The three minimum energy staggered conformations and the three maximum-energy eclipsed structures of ethane are a classic example of the way in which the energy changes with a bond rotation [60].

Quantum mechanical calculations suggest that the rotation-barrier arises from antibonding interactions between the H-atoms on opposite ends of the molecule; they are minimized when the conformation is staggered. Not all molecular mechanics force fields use torsional potentials; it may be possible to rely on non-bonded interactions between the atoms at the end of each torsion angle. However, most force fields for organic molecules do use explicit torsional potentials with a contribution from each bonded quartet of atoms A-B-C-D in the system (Fig. 2.10).

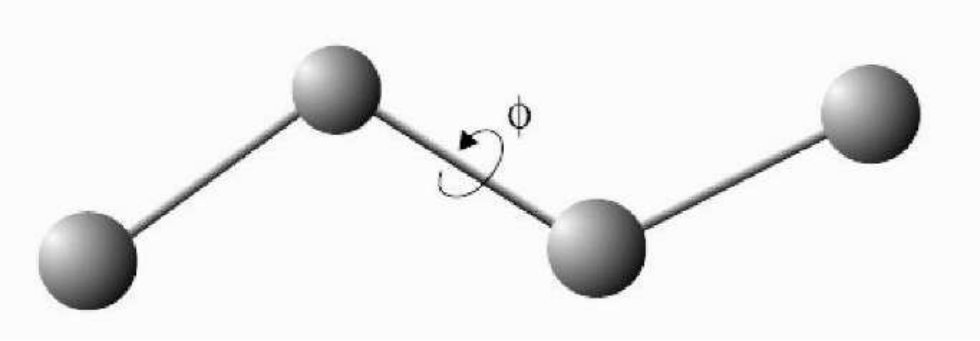

Figure 2.10: A torsion angle (dihedral angle) A-B-C-D is defined as the angle Phi between the planes (ABC) and (BCD). A torsion angle can vary through 360 degrees.

Thus, there would be nine individual torsional terms for ethane. Torsional potentials are almost

always expressed as a cosine series expansion. One functional form is:

$$V_T(\phi) = \frac{1}{2}k_\phi.[1 + \cos(n\phi - \delta)] \tag{2.16}$$

where the quantity k_ϕ is often referred to as the barrier height. The value of k_ϕ gives a qualitative indication of the relative barriers to rotation; for example k_ϕ for an amide bond (A-C=N-D) will be larger than for a bond between two sp^3 carbon atoms (A-C-C-D). n is the periodicity; its value gives the number of minimum points in the function as the bond is rotated through 360°C. δ (the phase factor) sets the minimum energy angle.

2.6.2.4 Improper Torsions / Out-of-Plane Bending

Several chemical groups involve arrangements of 4 or more atoms in a plane like the peptide bond in proteins. For these groups it is sometimes advantageous to have an additional term enforcing planarity.

The simplest way to achieve the desired geometry is to use an out-of-plane bending term. One approach is to treat the four atoms as an improper torsion angle i.e., a torsion angle in which the four atoms are not bonded in the sequence A-B-C-D. (Fig. 2.11).

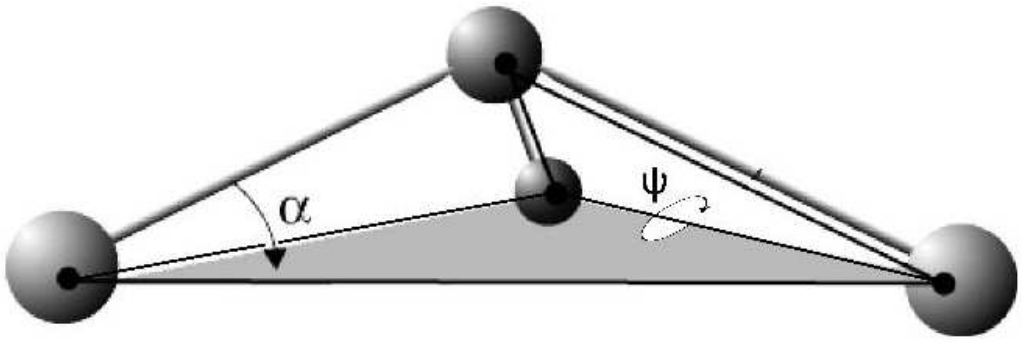

Figure 2.11: The improper dihedral term is designed to maintain planarity about certain atoms. The potential is described by a harmonic function. α is the angle between the plane formed by the central atom and two peripheral atoms and the plane formed by the peripheral atoms only.

A value of 0° corresponds to all four atoms being planar. The deviation of the out-of-plane coordinate can be modeled using a harmonic potential of the form:

$$V(\alpha) = \frac{1}{2}k_\alpha.\alpha^2 \tag{2.17}$$

We are now going to have a look at the non-bonded terms which consist at least of the electrostatic and the van der Waals interactions in the system.

2.6.3 NonBonded terms

2.6.3.1 Electrostatic Interactions

Interactions between atoms due to their permanent dipole moments are described approximately by treating the charged portions as point charges. Then we use the Coulomb potential for point charges to estimate the forces between the charged atoms. The Coulomb potential is an effective pair potential that describes the interaction between two point charges. It acts along the line connecting the two charges. It is given, in SI units, by the equation:

$$V_E(i,j) = \frac{q_i q_j}{4\pi\epsilon_0 r_{ij}}. \tag{2.18}$$

or, in more useful AKMA units

$$V_E(i,j) = 332\frac{q_i q_j}{r_{ij}}. \tag{2.19}$$

r_{ij} is the distance between q_i and q_j, the electric charge in coulombs carried by charge i and j respectively, and ϵ_0 is the electrical permittivity of space. Alternative approaches to the calculation of electrostatic interactions, e.g. the central multipole expansion which is based on the electric moments, may provide more exact solutions to the electrostatic interactions [61].

2.6.3.2 Van der Waals Interactions

The van der Waals interaction between two atoms arises from a balance between repulsive and attractive forces. The repulsive force arises at short distances where the electron-electron interaction is strong (Pauli repulsion). The attractive force, also referred to as dispersion force, arises from fluctuations in the charge distribution in the electron clouds. The fluctuation in the electron distribution on one atom gives rise to an instantaneous dipole which, in turn, induces a dipole in a second atom giving rise to an attractive interaction. Each of these two effects is equal to zero at infinite atomic separation and becomes significant as the distance decreases. The attractive interaction is longer range than the repulsion but as the distance becomes short, the repulsive interaction becomes dominant. This gives rise to a minimum in the energy (see Fig. 2.12).

The best known of the van der Waals potential functions is the Lennard-Jones 12-6 function, which takes the following form for the interaction between two atoms:

$$V_{\text{vdw}}(r) = 4\epsilon.\left[\left(\frac{\sigma}{r}\right)^{12} - \left(\frac{\sigma}{r}\right)^{6}\right] \tag{2.20}$$

The Lennard-Jones 12-6 potential contains just two adjustable parameters: the collision diameter σ (the separation for which the energy is zero) and the well depth ϵ. These parameters are graphically illustrated in Fig. 2.12. The need for a function that can be rapidly evaluated is a consequence of the large number of van der Waals interactions that must be determined in many of the systems that we would like to model. The 12-6 potential is widely used, particularly for calculations on large systems, as r^{-12} can be quickly calculated by squaring the r^{-6} term. The r^{-6} term can also be calculated from the square of the distance without having to perform a computationally expensive square root calculation. The 6-12 term between hydrogen-bonding atoms is replaced by an explicit hydrogen bonding term in some force fields, which is often described using a Lennard-Jones 10-12

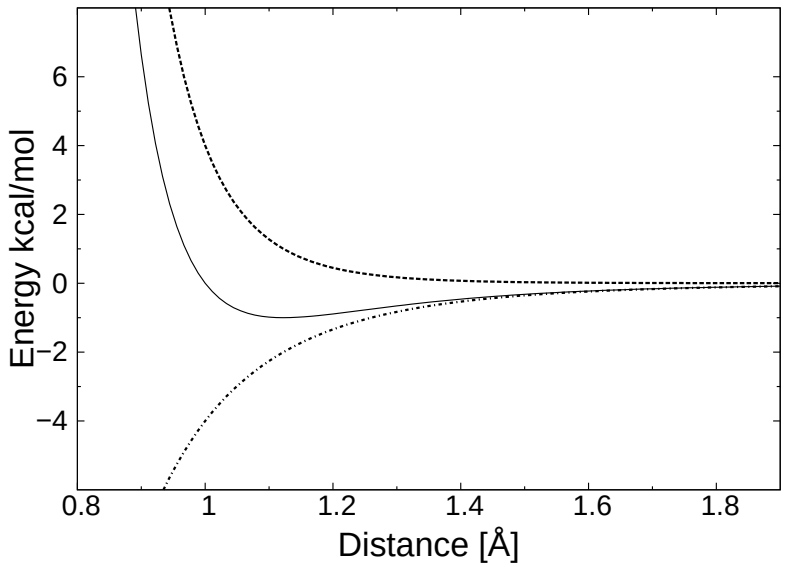

Figure 2.12: The Lennard-Jones potential following equation 2.20 with ϵ=1.0 kcal/mol and σ = 1.0 Å is shown as solid line while the van der Waals attraction and Pauli repulsion are represented as dash-dotted and dotted lines, respectively.

potential. The most time consuming part of a molecular dynamics simulation is the calculation of the non-bonded terms in the potential energy function. In principle, these energy terms should be evaluated between every pair of atoms; in this case, the number increases as the square of the number of atoms for a pair-wise model. To speed up the computation two approaches are applied. In the first approach the interactions between two atoms separated by a distance greater than a pre-defined distance, the "cutoff" distance, are ignored. The interactions are simply set to zero for interatomic distances greater than the cutoff distance (Truncation-Method) or the entire potential energy surface is modified such that at the cutoff distance the interaction potential is zero (Shift-Method). The other way is to reduce the number of interaction sites. The simplest way to do this is to subsume some or all of the atoms (usually just the hydrogen atoms) into the atoms to which they are bonded (United-Atom-Method). Considerable computational savings are possible; for example, if butane is modeled as a four- site model rather than one with twelve atoms, the van der Waals interaction between all the atoms involves the calculation of six terms rather than 78.

2.7 The CHARMM program and force field

2.7.1 The CHARMM forcefield

The CHARMM (Chemistry at HARvard Molecular Mechanics) program is a general purpose molecular mechanics simulation program. Besides energy minimization, dynamics simulation, vibrational analysis and thermodynamic calculations which are all performed with the use of the empirical potential energy function (the force field), there are interfaces to several quantum mechanical programs allowing mixed QM and MM calculations. The program can treat systems ranging in size from small

individual organic molecules to large proteins and DNA molecules either isolated, in solutions or in crystalline solids. In this section we explicitly describe the core of the CHARMM program, the force field, and some important files which the program works with. We start with an introduction into characteristics of force fields before dealing with the single portions of the function.

The CHARMM Force Field is a simple Molecular Mechanics force field which includes energy terms previously described:

$$\begin{aligned} U_p = & \sum_{\text{bonds}} k_r(r - r_0)^2 + \sum_{\text{angles}} k_\theta(\theta - \theta_0)^2 + \sum_{\text{dihedrals}} k_\phi(1 + \cos(n\phi - \delta)) \\ & + \sum_{\text{impropers}} k_\omega(\omega - \omega_0)^2 + \sum_{\text{UB}} k_\xi(\xi - \xi_0)^2 \\ & + \sum_{\text{nonbonded}} \sum \left\{ \left[\left(\frac{A_{ij}}{r_{ij}} \right)^{12} - \left(\frac{B_{ij}}{r_{ij}} \right)^{6} \right] + \frac{1}{4\pi\epsilon_0\epsilon_r} \frac{q_i q_j}{r_{ij}} \right\} \end{aligned} \tag{2.21}$$

2.7.2 Data Structures

Data Structures include information about the molecule, its composition, its chemical connectivity, certain atomic properties and parameters for the energy function and more. This information for a particular class of molecules, e.g. proteins or nucleic acids, is contained in the topology file and the parameter file. For a specific molecule, the necessary data is stored in the Protein Structure File and the Coordinate File, respectively.

2.7.2.1 Residue Topology File (RTF)

The RTF contains local information about atoms, bonds, angles etc. for each possible type of monomer unit (residue) that can be used in building a particular type of macromolecule. For each residue the covalent structure is defined, i.e., how the atoms are connected to one another to form amino acids, DNA bases or lipid molecules.

2.7.2.2 Parameter File (PARAM)

The parameter File is associated with the RTF file as it contains all the necessary parameters for calculating the energy of the molecule(s). These include the equilibrium bond lengths and angles for bond stretching, angle bending and dihedral angle terms in the potential energy function as well as the force constants and the Lennard-Jones 12-6 potential parameters. As already mentioned, these parameters are associated with particular atom types.

For our particular study of metalloproteins, new parameters had been added to the CHARMM potential following previous work by researchers in the laboratory [59]. Some of the new CHARMM parameters for metals Cu(I) and Hg(II) are listed in Table 2.1. 4 atom types have been added : CU for copper(I) and HG for mercury (II), SCU for a charged sulfur atom bonded to copper(I) and SHG for a charged sulfur atom bonded to mercury (II).

Nonbonded parameters for the 4 new atom types are given in Table 2.2

Bond distance	Force constant (kcal/mol/Å^2)	Equilibrium value (Å)
SCU CU	82	2.135
SHG HG	97	2.319
Bond angle	Force constant (kcal/mol/rad^2)	Equilibrium value (degrees)
SCU CU SCU	18	182
SHG HG SHG	26	178.5

Table 2.1: New CHARMM bonded parameters for metals

Atom	Charge	R (Å) ($= \sigma/2^{5/6}$)	ϵ (kcal/mol)
HG	0.31	1.60	1.0
SHG	-0.27		
CU	0.15	1.40	0.19
SCU	-0.60		

Table 2.2: Charges and van der Waals parameters for a new metal force field in CHARMM.

2.7.2.3 Protein Structure File (PSF)

The PSF is the most fundamental data structure used in CHARMM. It is generated for a specific molecule or molecules and contains the detailed composition and connectivity of the molecule(s). It describes how molecules are divided into residues and molecular entities (segments), which can range from a single macromolecular chain to multiple chains solvated by explicit water molecules.

2.7.2.4 Coordinate File (CRD)

The CRD file contains the Cartesian coordinates of all atoms in the system. Those are most often obtained from X-ray crystal structures or from NMR structures. Missing coordinates can be built within the CHARMM program.

2.7.3 Energy Minimization in CHARMM

2.7.3.1 Description of Minimization

Minimization reduces the energy of the system to the lowest possible point, simulating the system's energy as absolute zero. There are different methods used for calculating the lowest-energy conformation of a given structure.

2.7.3.2 Minimization methods

The simplest minimization algorithm is steepest descent (SD). In each step of this iterative procedure, the coordinates are adjusted in the negative direction of the gradient. It has one adjustable parameter, the step size, which determines how far to shift the coordinates at each step. The step size is adjusted

depending on whether a step results in a lower energy. i.e., if the energy drops, the step size is increased by 20% to accelerate the convergence. If the energy rises, overshooting a minimum, the step size is halved. Steepest descent does not converge in general, but it will rapidly improve a very poor conformation.

A second method is the conjugate gradient technique (CONJ) which has better convergence characteristics [62]. The method is iterative and makes use of the previous history of minimization steps as well as the current gradient to determine the next step. It can be shown that the method converges to the minimum energy in N steps for a quadratic energy surface where N is the number of degrees of freedom in the energy. Since several terms in the potential are quadratic, it requires less evaluations of the energy and gradient to achieve the same reduction in energy in comparison to steepest descent. Its main drawback is that with very poor conformations, it is more likely to generate numerical overflows than steepest descent. The algorithm used in CHARMM has a slightly better interpolation scheme and automatic step size selection.

A third method involves solving the Newton-Raphson (NRAP) minimization equations iteratively. This procedure requires the computation of the derivative of the gradient which is a matrix of size $O(n^2)$. The procedure here is to find a point where the gradient will be zero (hopefully a minimum in energy) assuming that the potential is quadratic. The Newton-Raphson equations can be solved by a number of means, but the method adopted for CHARMM involves diagonalizing the second derivative matrix and then finding the optimum step size along each eigenvector. The advantages of this method are that the geometry cannot remain at a saddle point, as sometimes occurs with the previous procedures, and that the procedure converges rapidly when the potential is nearly quadratic (or cubic). The major disadvantage is that this procedure can require excessive storage requirements, $O(n^2)$, and computation time, $O(n^3)$, for large molecules. Thus we are currently restricted to systems with about 200 atoms or less for this method. IMAGES and SHAKE are currently unavailable with this method.

Finally, the most generally used method is an adopted basis Newton-Raphson method (ABNR). This routine performs energy minimization using a Newton-Raphson algorithm applied to a subspace of the coordinate vector spanned by the displacement coordinates of the last (MINDim) positions. The second derivative matrix is constructed numerically from the change in the gradient vectors, and is inverted by an eigenvector analysis allowing the routine to recognize and avoid saddle points in the energy surface. At each step the residual gradient vector is calculated and used to add a steepest descent step onto the Newton-Raphson step, incorporating new direction into the basis set. This method is the best for most circumstances.

2.8 Solvation

2.8.1 Explicit solvation

Solvating consists in surrounding the solute, generally a protein or DNA strand, with a solvent. This is very useful in biomolecular simulations since biochemical reactions generally take place in a viscous environment. To deal with the problem of a theoretically "infinite" bulk solvent compared to the size of the macromolecule, one usually employs periodicity and so-called "Periodic boundary conditions".

2.8.1.1 Periodic boundaries

Periodic boundary conditions (PBC) are a set of boundary conditions that are often used to simulate a large system by modelling a small part that is far from its edge. Periodic boundary conditions resemble the topologies of some video games; a unit cell or simulation box of a geometry suitable for perfect three-dimensional tiling is defined, and when an object passes through one face of the unit cell, it reappears on the opposite face with the same velocity. The simulation is of an infinite perfect tiling of the system. The tiled copies of the unit cell are called images, of which there are infinitely many. During the simulation, only the properties of the unit cell need be recorded and propagated. The minimum-image convention shown in Fig. 2.13 is a common form of PBC particle bookkeeping in which each individual particle in the simulation interacts with the closest image of the remaining particles in the system.

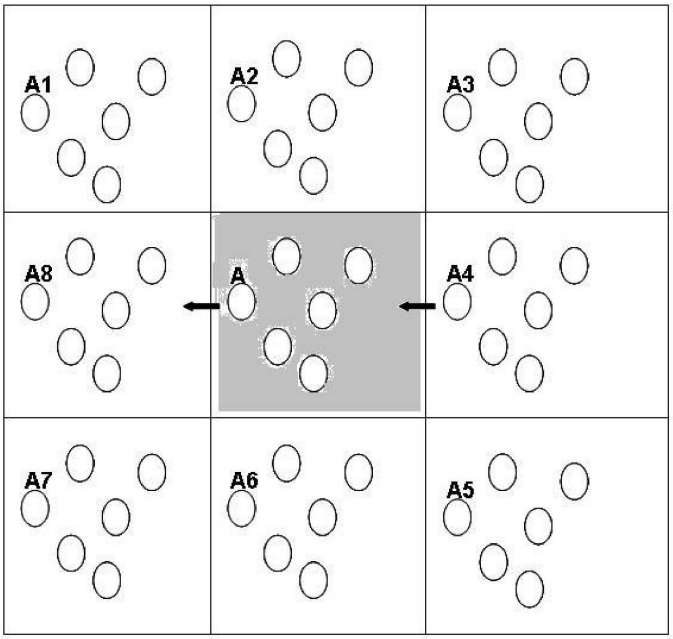

Figure 2.13: Two-dimensional view of the minimum-image convention for Periodic boundary conditions. Each particle in the simulation cell (center) interacts with images in the 8 neighbouring cells (26 in 3-D). When a particle A lives the central simulation cell, for instance by the left, it automatically "reenters" by the right side through its periodic copy A4.

In CHARMM, to solvate a protein prior to a simulation using PBC, we generally embed it inside a parallelepipedic box of TIP3P waters. Then we delete all waters with O atom closer than 2.4 Å from any heavy atom of the protein. Then, we minimize the energy of the system and run equilibration MD phases where protein atoms are harmonically restrained to their original positions with decreasing force constants. Alternatively, for a simulation inside a spherical droplet of water, the system is immersed inside a sphere of pre-equilibrated water molecules and equilibrated after removing overlapping waters. The immersion procedure can be repeated several times so as to prevent solvent holes.

In CHARMM, there is no specific command for solvation, but all the preceding stages can be done through a series of commands and scripts.

2.8.2 Implicit solvation

Implicit solvation (sometimes known as continuum solvation) is a method of representing solvent as a continuous medium instead of individual "explicit" solvent molecules most often used in molecular dynamics simulations and in other applications of molecular mechanics (see above). The method is often applied to estimate free energy of solute-solvent interactions in structural and chemical processes, such as folding or conformational transitions of proteins, DNA, RNA, and polysaccharides, association of biological macromolecules with ligands, or transport of drugs across biological membranes.

The implicit solvation model is justified in liquids, where the potential of mean force can be applied to approximate the averaged behavior of many highly dynamic solvent molecules. However, the interiors of biological membranes or proteins can also be considered as media with specific solvation or dielectric properties. These media are continuous but not necessarily uniform, since their properties can be described by different analytical functions, such as "polarity profiles" of lipid bilayers [63]. There are two basic types of implicit solvent methods: models based on accessible surface areas (ASA) that were historically the first, and more recent continuum electrostatics models, although various modifications and combinations of the different methods are possible.

The accessible surface area (ASA) method is based on experimental linear relations between Gibbs free energy of transfer and the surface area of a solute molecule [64]. This method operates directly with free energy of solvation, unlike molecular mechanics or electrostatic methods that include only the enthalpic component of free energy. The continuum representation of solvent also significantly improves the computational speed and reduces errors in statistical averaging that arise from incomplete sampling of solvent conformations [65] so that the energy landscapes obtained with implicit and explicit solvent are different [66]. Although the implicit solvent model is useful for simulations of biomolecules, this is an approximate method with certain limitations and problems related to parameterization and treatment of ionization effects.

2.8.2.1 Accessible surface area based method

The free energy of solvation of a solute molecule in the simplest ASA-based method is given by:

$$\Delta G_{\text{solv}} = \sum_i \sigma_i \text{ASA}_i \tag{2.22}$$

where ASA_i is the accessible surface area of atom i, and σ_i is the solvation parameter of atom i, i.e. a contribution to the free energy of solvation of the particular atom i per surface unit area. The required solvation parameters for different types of atoms (C, N, O, S, etc.) are usually determined by least squares fit of the calculated and experimental transfer free energies for a series of organic compounds. The experimental energies are determined from partition coefficients of these compounds between different solutions or media using standard mole concentrations of the solutes [67].

It is noteworthy that solvation energy is free energy required to transfer a solute molecule from a solvent to "vacuum" (gas phase). This solvation energy can supplement the intramolecular energy in vacuum calculated in molecular mechanics. Therefore, the required atomic solvation parameters were initially derived from water-gas partition data [68]. However, the dielectric properties of proteins and lipid bilayers are much more similar to those of nonpolar solvents than to vacuum. Newer parameters

have therefore been derived from water-octanol partition coefficients [69] or other similar data. Such parameters actually describe transfer energy between two condensed media or the difference of two solvation energies.

2.8.2.2 Poisson-Boltzmann

Although this equation has solid theoretical justification, it is computationally expensive to calculate without approximations. The Poisson-Boltzmann equation (PB) describes the electrostatic environment of a solute in a solvent containing ions. It can be written in cgs units as:

$$\vec{\nabla}.\left[\epsilon(\vec{r})\vec{\nabla}\Psi(\vec{r})\right] = -4\pi\rho(\vec{r}) - 4\pi\sum_i c_i z_i q\lambda(\vec{r}).exp\left[\frac{-z_i q\Psi(\vec{r})}{k_B T}\right] \tag{2.23}$$

where $\epsilon(\vec{r})$ represents the position-dependent dielectric, $\Psi(\vec{r})$ represents the electrostatic potential, $\rho(\vec{r})$ represents the charge density of the solute, c_i represents the concentration of the ion i at a distance of infinity from the solute, z_i is the valence of the ion, q is the charge of a proton, k_B is the Boltzmann constant, T is the temperature, and $\lambda(\vec{r})$ is a factor for the position-dependent accessibility of position r to the ions in solution (often set to uniformly 1). If the potential is not large, the equation can be linearized to be solved more efficiently [70].

A number of numerical Poisson-Boltzmann equation solvers of varying generality and efficiency have been developed [71], including one application with a specialized computer hardware platform [72]. However, performance from PB solvers does not yet equal that from the more commonly used generalized Born approximation [73].

2.8.2.3 Generalized Born

The Generalized Born (GB) model is an approximation to the exact (linearized) Poisson-Boltzmann equation. It is based on modeling the protein as a sphere whose internal dielectric constant differs from the external solvent. The model has the following functional form:

$$G_s = \frac{1}{8\pi}\left(\frac{1}{\epsilon_0} - \frac{1}{\epsilon}\right)\sum_{i,j}^{N}\frac{q_i q_j}{f_{GB}}$$

where

$$f_{GB} = \sqrt{r_{ij}^2 + a_{ij}^2 e^{-D}}$$

and

$$D = \left(\frac{r_{ij}}{2a_{ij}}\right)^2,\ a_{ij} = \sqrt{a_i a_j}$$

where ϵ_0 is the permittivity of free space, ϵ is the dielectric constant of the solvent being modeled, q_i is the electrostatic charge on particle i, r_{ij} is the distance between particles i and j, and a_i is a quantity (with the dimension of length) known as the effective Born radius [74]. The effective Born radius of an atom characterizes its degree of burial inside the solute; qualitatively it can be thought of as the distance from the atom to the molecular surface. Accurate estimation of the effective Born radii is critical for the GB model [75].

2.8.2.4 GBSA

GBSA is simply a Generalized Born model augmented with the hydrophobic solvent accessible surface area SA term. It is among the most commonly used implicit solvent model combinations. The use of this model in the context of molecular mechanics is known as MM/GBSA. Although this formulation has been shown to successfully identify the native states of short peptides with well-defined tertiary structure [76], the conformational ensembles produced by GBSA models in other studies differ significantly from those produced by explicit solvent and do not identify the protein's native state [66]. In particular, salt bridges are overstabilized, possibly due to insufficient electrostatic screening, and a higher-than-native alpha helix population was observed. Variants of the GB model have also been developed to approximate the electrostatic environment of membranes, which have had some success in folding the transmembrane helices of integral membrane proteins [77].

2.8.2.5 Ad-hoc fast solvation models

Another possibility is to use ad-hoc quick strategies to estimate solvation free energy. A first generation of fast implicit solvents is based on the calculation of a per-atom solvent accessible surface area. For each of group of atom types, a different parameter scales its contribution to solvation ("ASA-based model" described above). [78]

Another strategy is implemented for the CHARMM19 force-field and is called EEF1 [79]. EEF1 is based on a Gaussian-shaped solvent exclusion. The solvation free energy is

$$\Delta G_i^{solv} = \Delta G_i^{ref} - \sum_j \int_{Vj} f_i(r) dr \tag{2.24}$$

The reference solvation free energy of i corresponds to a suitably chosen small molecule in which group i is essentially fully solvent-exposed. The integral is over the volume V_j of group j and the summation is over all groups j around i. EEF1 additionally utilizes a distance-dependent (non-constant) dielectric, and ionic side-chains of proteins are simply neutralized. It is only 50% slower than a vacuum simulation. This model was later augmented with the hydrophobic effect and called Charmm19/SASA. [80]

2.8.3 Hybrid implicit/explicit solvation models

It is possible to include a layer or sphere of water molecules around the solute, and model the bulk with an implicit solvent. Such an approach is proposed by M.S. Lee and coworkers [81]. The bulk solvent is modeled with a Generalized Born approach and the multi-grid method used for Coulombic pairwise particle interactions. It is reported to be faster than a full explicit solvent simulation with the particle mesh Ewald (PME) method of electrostatic calculation.

2.8.4 Effects not accounted for

2.8.4.1 The hydrophobic effect

Models like PB and GB allow estimation of the mean electrostatic free energy but do not account for the (mostly) entropic effects arising from solute-imposed constraints on the organization of the water

or solvent molecules. This is known as the hydrophobic effect and is a major factor in the folding process of globular proteins with hydrophobic cores. Implicit solvation models may be augmented with a term that accounts for the hydrophobic effect. The most popular way to do this is by taking the solvent accessible surface area (SASA) as a proxy of the extent of the hydrophobic effect. Most authors place the extent of this effect between 5 and 45 cal/(^{2} mol). [82] Note that this surface area pertains to the solute, while the hydrophobic effect is mostly entropic in nature at physiological temperatures and occurs on the side of the solvent.

2.8.4.2 Viscosity

Implicit solvent models such as PB, GB, and SASA lack the viscosity that water molecules impart by randomly colliding and impeding the motion of solutes through their van der Waals repulsion. In many cases, this is desirable because it makes sampling of configurations and phase space much faster. This acceleration means that more configurations are visited per simulated nanosecond, on top of whatever CPU acceleration is achieved in comparison to explicit solvent. But it can lead to misleading results when kinetics are of interest. Viscosity may be added back by using Langevin dynamics instead of Hamiltonian dynamics and choosing an appropriate damping constant for the particular solvent [83].

2.8.4.3 Hydrogen-bonds with water

The average energetic contribution of protein-water hydrogen bonds may be reproduced with an implicit solvent. However, the directionality of these hydrogen bonds will be missing.

Chapter 3

Metal-Binding Membrane Proteins

In this chapter, I will present ionic channels and transporters introducing a modelling study we have recently published [84]. Then, I will give an overview of the current knowledge on metal transporting ATPases, subject of the major part of this thesis.

3.1 Ionic Channels and Transporters

3.1.1 Overview

Membrane transport proteins are crucial for life. They regulate the fluxes of ions, nutrients and other molecules across the membranes of all cells, and their activities underlie physiological processes as diverse as brain electrical activity, muscle contraction, water and solute transport in the kidney, hormone secretion and the immune response. Mutations in membrane transport proteins, or defects in their regulation, are responsible for many human diseases. Consequently, these proteins are targets for widely used therapeutic drugs.

Traditionally, membrane transport proteins have been divided into two groups: channels and transporters. Channels are membrane-spanning water-filled pores through which substrates passively diffuse down their electrochemical gradients whenever the regulatory gate is open. Transporters undergo a cycle of conformational changes linked to substrate binding and dissociation on opposite sides of the membrane. This conformational cycle can be coupled to energy sources like pre-existing ion gradients or ATP hydrolysis, thus allowing substrates to be moved "uphill" against their concentration gradients, as in nutrient and ion accumulation into the cell or export from the cell of ions, drugs or xenobiotics.

All transporters must effectively have two "gates" that control access from either side of the membrane to the substrate-binding sites as well as a conformational cycle that prevents both these gates from being open at the same time. It is obvious that if both gates were open simultaneously, the protein would then operate as a channel. And, owing to the orders-of-magnitude higher flow rates through channels than through transporters, even a fleeting moment of channel-like behaviour would render a transporter useless. To obviate any such occurrence, the conformational cycles of many transporters incorporate occluded states in which both gates are shut, enclosing the bound substrate, before one of the gates opens to release it.

3.1.2 Special ATP-binding cassette transporters

ATP-binding cassette (ABC) transporters are ubiquitous membrane proteins that couple the energy of ATP hydrolysis to translocation of diverse substrates across cell membranes. It has long been recognized that the sulphonylurea receptor SUR and the cystic fibrosis transmembrane conductance regulator CFTR are exceptional among ABC proteins in that they do not serve as pumps. Instead, they have hijacked the ATP-binding and hydrolytic activity of the nucleotide-binding domains (NBDs) to gate an intrinsic chloride channel (CFTR) or to regulate the gating of a separate inward-rectifier K_{ATP} potassium channel (SUR). Recent crystal structures of bacterial ABC transporters have suggested a common molecular mechanism by which binding and hydrolysis of ATP are coupled to conformational changes in the membrane-spanning domains, as discussed by [85].

3.1.3 K_{ATP}-channels

ATP-sensitive potassium (K_{ATP}) channels are non-voltage-dependent, potassium-selective channels gated by the intracellular nucleotides ATP and ADP. Gating is complex and thought to reflect the static and dynamic nature of the cellular metabolic status. Thus K_{ATP} channels are postulated to act as sensors of intracellular metabolism, tuning the potassium permeability, and therefore the electrical activity, of a cell to its energetic balance. These channels are present in most excitable cells. In pancreatic β cells, they play a key role in coupling insulin secretion to plasma glucose. In muscle and neuronal cells, their function is not as firmly established but evidence is strengthening for an implication in the protective response to various metabolic insult.

The K_{ATP} channel is made up of two proteins: the $\sim$ 160 kDa sulfonylurea receptor SUR which is a member of the ATP binding cassette (ABC) transporter family, and a smaller $\sim$ 40 kDa protein Kir6.2 (Uniprot entry IRK11-HUMAN) or Kir6.1 which belongs to the inward rectifier K^+ channel family. Four Kir6.x subunits assemble to form a K^+-selective pore, which is constitutively associated to four SUR subunits, (Fig. 3.1).

The ultrastructure of the K_{ATP} channel is unique for an ion channel as it is the only known instance of such an intricate partnership between members of the ion channel and ABC protein gene families. Kir6.x channels are already endowed with an inhibitory site for ATP, which is modulated by phospholipids. SUR adds two nucleotide binding sites able to bind and hydrolyze ATP. The resulting channel complex is predicted to have 96 transmembrane helices and 12 nucleotide binding sites. Another peculiar trait of the K_{ATP} channel is its pharmacological repertoire, which is the richest of all known potassium channels and includes both blockers and openers.

3.1.4 Biomolecular sensors based on SUR

Protein-based bio molecule sensors engineered by covalent assembly of natural ion channels and receptors forming nanoscale electrical biosensors, are promising tools for diagnostics and high-throughput screening systems. The electrical signal (current through the channel) allows label-free assays with high signal/noise ratio and fast real-time measurements that are well adapted to microelectronic chips. The challenge in developing such biosensors lies in functionally connecting a molecule detector to an electrical switch. Advances in this field has relied on synthetic ion-conducting pores and modified

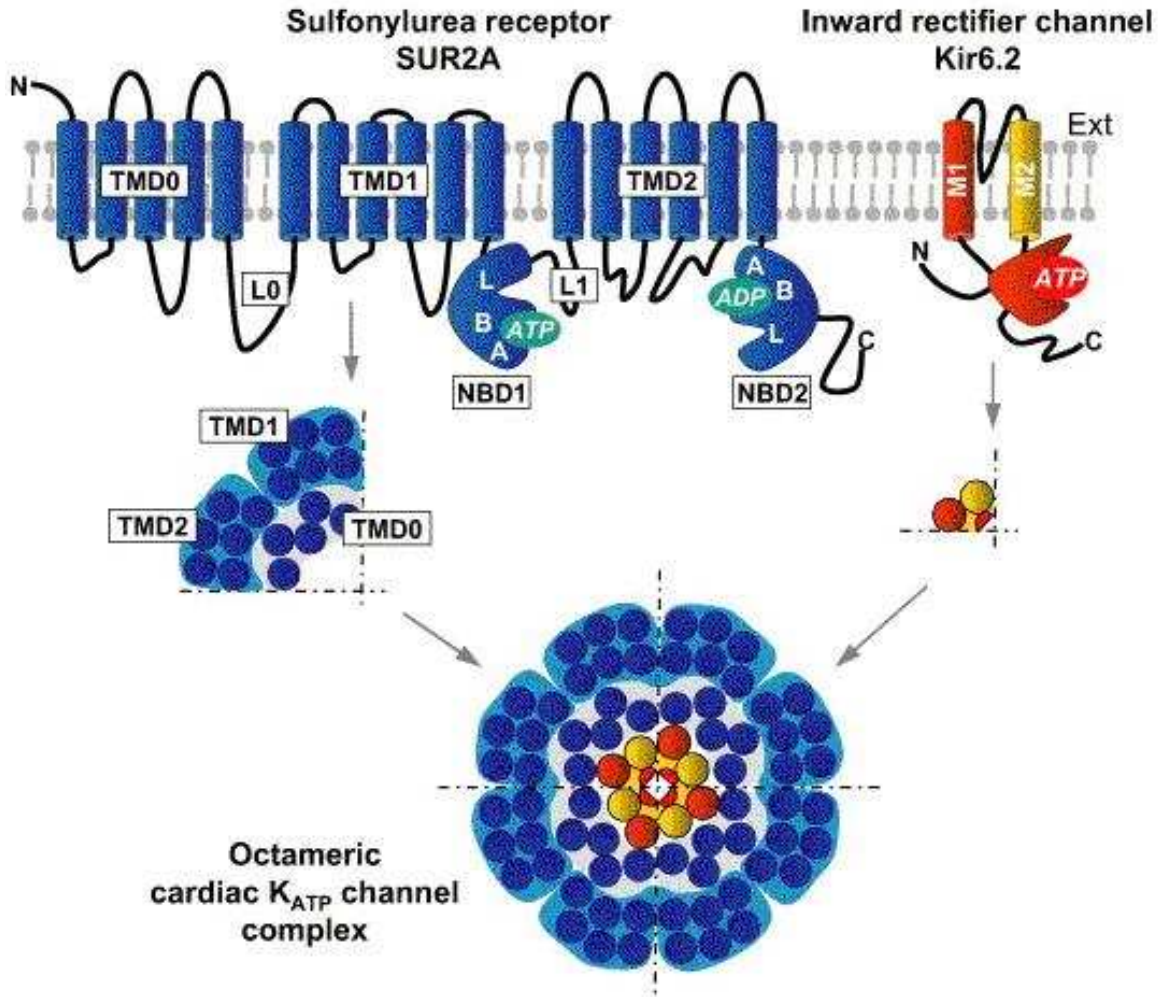

Figure 3.1: Hypothetical structural organization of the cardiac K_{ATP} channel. Kir6.2 (390 amino acids) and SUR2A (1549 amino acids) are the constitutive subunits of the cardiac K_{ATP} channel. Four subunits of the inwardly-rectifying K^+ channel, Kir6.1 or Kir6.2, associate with four ABC proteins, SUR, to form a functional K_{ATP} channel octamer. Kir6.x has two transmembrane helices M1 and M2, and a large cytoplasmic domain harboring an inhibitory binding site for ATP. SUR possesses three transmembrane domains (TMD0, 1 and 2), and two cytoplasmic nucleotide binding domains (NBD1 and 2) Bottom representation is a bird's eye view of the assembled complex.

ion channels that are not yet suitable for biomolecule screening. Based on the example of the channel coupled to SUR, researchers in M. Vivaudou's team (Institut de Biologie Structurale, Grenoble) have designed and characterized novel bioelectric sensing elements engineered by coupling an ion channel to G-Protein-Coupled Receptors (GPCRs). These Ion Channel Coupled Receptors (ICCRs) may potentially detect a wide range of ligands recognized by natural or altered GPCRs. GPCRs are major pharmaceutical targets, and ICCRs could form the basis of a unique platform for label-free GPCR-drug screening.

As discussed earlier, the ICCR model, the inward-rectifier K^+ channel Kir6.2 and the ATP- binding-cassette (ABC) protein SUR are structurally unrelated membrane proteins that associate naturally to form the ATP-sensitive K^+ channel (K_{ATP}), a sensor of cellular metabolism. In the K_{ATP} channel, Kir6.2 and SUR are functionally coupled so that binding of metabolites or pharmacological agents to SUR modulates the opening and closing of Kir6.2. Inspired by this rather unique design, the idea was that if SUR could be replaced by another receptor such as a GPCR, an ion channel sensitive to GPCR ligands could be created and utilized to convert chemical information into an electrical signal (Fig. 3.2)

3.1.5 Model of an ICCR

The GPCR which was chosen for this study is HM2 (Uniprot entry ACM2-HUMAN), the human muscarinic receptor which counts 466 amino acids, while Kir6.2 was chosen as the K^+ channel. Experimentally, the binding of acetylcholine to HM2 effectively induces a current flowing through Kir6.2.

To shed some light on the mechanism of signal transduction between HM2 and Kir6, models of HM2, Kir6 monomer, Kir6 tetramer and HM2 monomer linked to Kir6 tetramer were built.

Initial models of HM2 monomer and Kir6 (mono- and tetra-mer) were first built by homology modeling with the program Modeller (see section 2.2.3). HM2 was modeled from the structure of the adrenergic receptor (PDB code 2RH1) (See Fig. 3.3) and Kir6 was modeled from a 3D structural alignment with 3 known structures: chimera between kirBac and Kir3.1 (PDB code 1N9P), cytoplasmic parts of Kir2.1 (PDB code 2GIX) and Kir3.1 (PDB code IN9P) (See Fig. 3.4).

10 models of each protein were built and the model with lowest objective function retained. Initial structures of HM2 and Kir6 were then refined with program CHARMM with several stages of energy minimization with decreasing restraints on the atomic position of backbone atoms. In particular, two constraints were added in the Kir6 model to maintain the two β strands between residues V36 and S37 for first strand and between N41 and C42 for the second.

Then the two proteins were linked, C-terminal residue of HM2 bonded to the N+25 terminal residue of Kir6. The choice of deleting the 25 NTER residues of Kir corresponds to a successfully coupled experimental construct. The two proteins were oriented with their transmembrane helical region perpendicular to the XY plane (membrane region). HM2 and Kir6 were rotated around the Z axis so that their sliding helices (residues 447 to 456 for HM2 and 57 to 66 for Kir6) lay In the same X axis orientation. HM2 was translated away from Kir6 so that the linker region (residues 457 to 522 of the complete construct) was totally extended. Indeed, in this simple mechanism, we assumed that the "signal" is transmitted from HM2 to Kir6 by mechanical work on the completely extended and almost stiff linker. After model building sketched in Fig. 3.5, we used the program AMD (See section 2.4) to

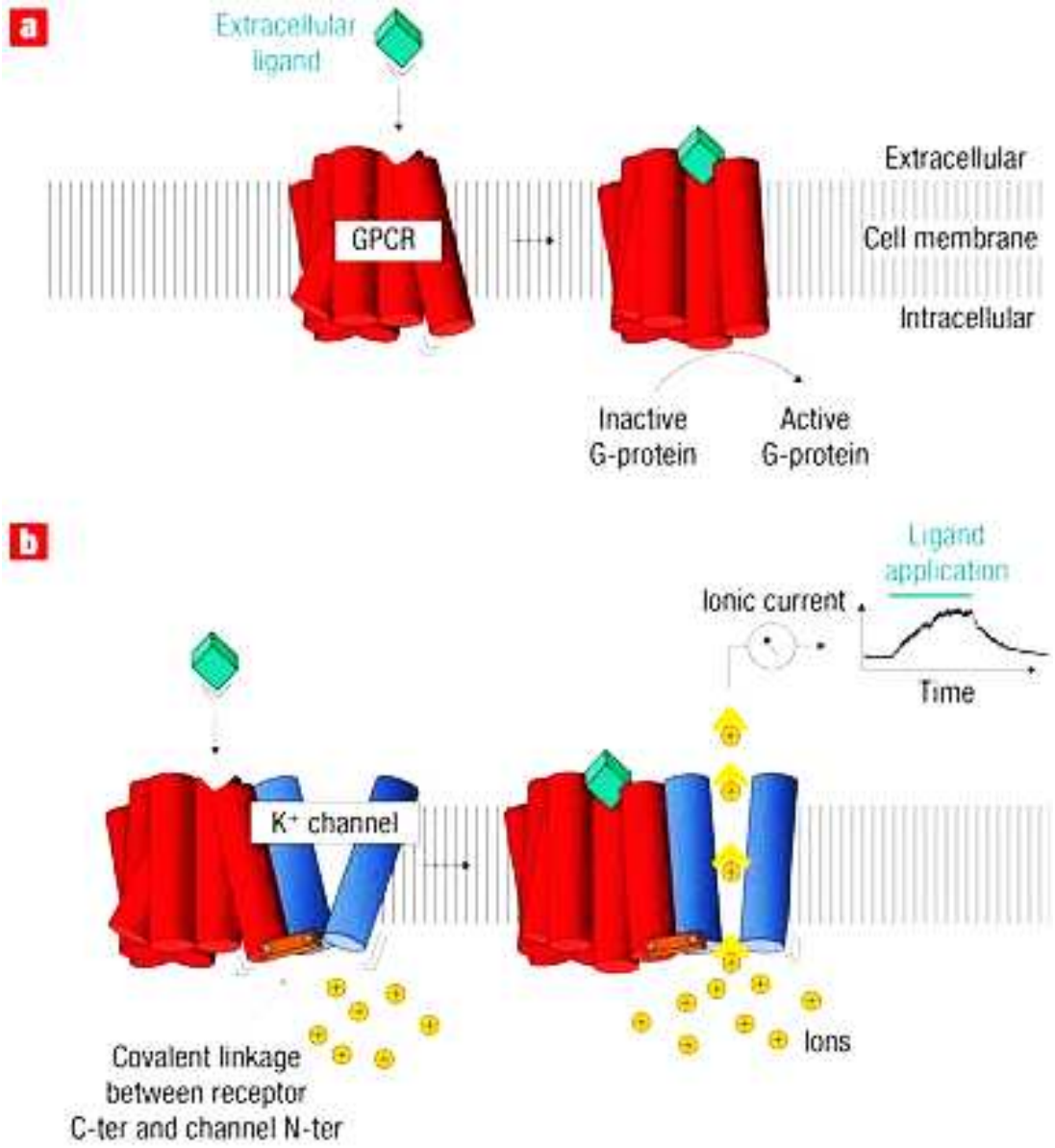

Figure 3.2: a, Upon binding of its ligand at an extracellular site, a transmembrane G-protein-coupled receptor (GPCR) adopts a new conformation that triggers activation of intracellular G-proteins. b, In an ICCR, the GPCR is attached to an ion channel in such a way that both proteins are mechanically coupled. When the GPCR binds a ligand and changes conformation, this change is directly transmitted to the channel and results in a change in gating and in the ionic current through the channel

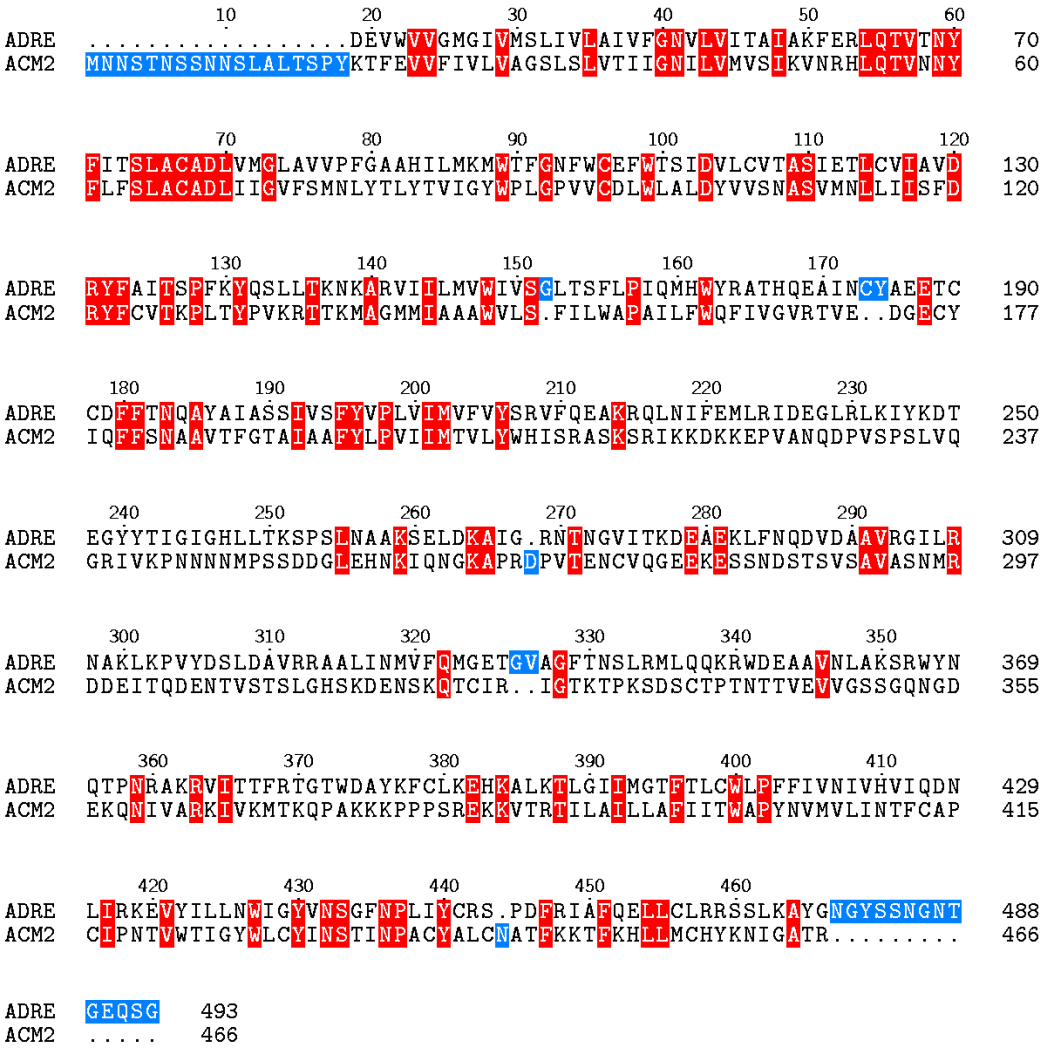

Figure 3.3: Alignment between HM2 (ACM2) and the human adrenergic receptor (ADRE) done with the program CLUSTALW [86].

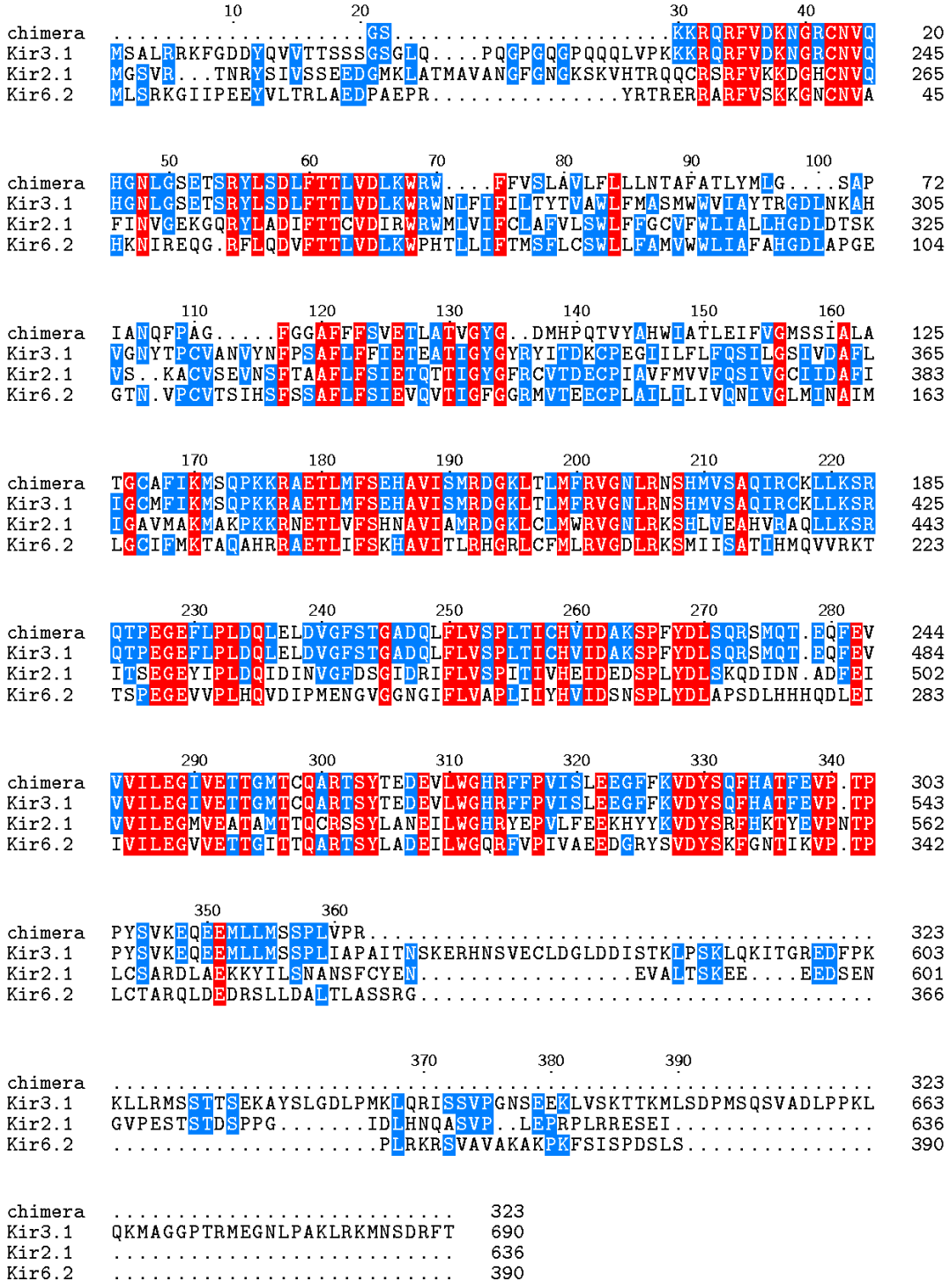

Figure 3.4: Alignment between Kir6.2 a chimera between kirBac and Kir3.1, Kir2.1 and Kir3.1 done with the program CLUSTALW [86].

interactively apply forces on HM2 transmembrane helix VII (residues 424-443) and on the linker.

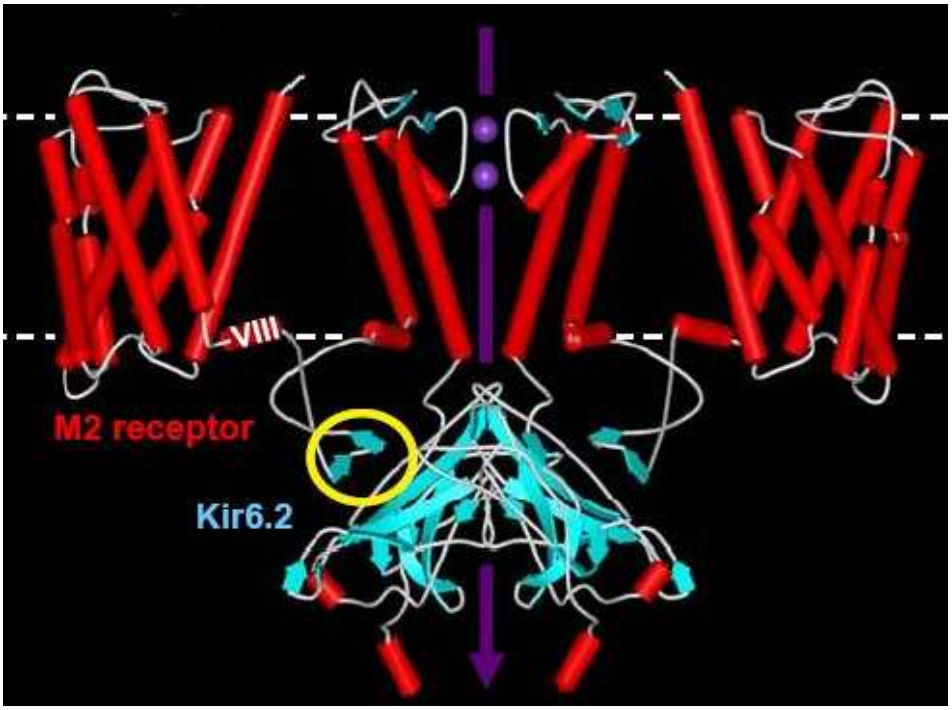

Figure 3.5: Hypothetical mechanism of intermolecular gating control. Transmembrane view of a molecular model of the HM2-Kir6.2 channel complex where the front and back subunits of the tetramer have been removed. GPCR conformational changes upon ligand binding could trigger a movement of helix VII that would pull on the Kir6.2 N-ter. The exerted force could be transmitted to the Kir6.2 C-ter, via the region of interaction between Kir6.2 C-ter and N-ter (circled in yellow), and to the attached transmembrane inner helix that controls channel gating.

A clear effect on the gating of Kir6 could not be observed for several possible reasons:

- The model of HM2 or Kir6 may not be optimal; especially the interface region
- The structure of the linker region, crudely modeled as extended, may be wrong
- Transmission of the signal may be more complicated than just mechanical work on the linker
- Forces applied were probably too strong. Smaller forces should be added continuously and for larger times (not interactively).

This project will be pursued in the coming years with M. Vivaudou's team. The program AMD, under development in S. Redon's team (INRIA, Grenoble), will be more performing with a lot of new functionalities, like accounting for symmetries in the systems and improved user interface.

Let's now come to other ATP-dependant ion transporters: P-Type ATPases

3.2 Generalities about P-Type ATPases

3.2.1 Introduction

Homeostasis and extrusion of solutes were processes developed early by the cells in order to survive [87]. The plasma membrane is the essential permeability barrier that separates the inside of the cell from the outside. However, it should also allow the transport of desired solutes. Transmembrane transport proteins are ubiquitous proteins that regulate intracellular concentrations of solutes by either extrusion or accumulation in sub-cellular organelles. P-Type ATPases are a ubiquitous family of transmembrane proteins involved in the cation transport against electrochemical gradients . The main structural characteristics of these enzymes are six to ten transmembrane α-helices (H1-H10) and one ATP binding domain. A highly conserved sequence (DKTGT) is present in the large cytoplasmic loop, and phosphorylation of the aspartic acid residue drives the key conformational changes in the protein [88].

The P-type ATPase is an extensive and expanding family of proteins that is divided into five groups (and some sub-groups) based on sequence alignments and putative ion specificity [89]. In this way, the group P1 is divided into: A: K^+ and B: heavy metals (Cu^+, Ag^+, Cu^{2+}, Cd^{2+}, Zn^{2+}, Pb^{2+}, Co^{2+}) transporters. Group P2: A: Ca^{2+}, Mn^{2+} including SERCA pumps; B: Ca^{2+}; including PMCA pumps; C: Na^+/K^+; H^+/K^+ and D: possible Na^+; Ca^{2+} pumps. Group P3: A: H^+ and B: Mg^{2+} transporters. P4 and P5 are groups with unknown selectivity but it was proposed that P4 could be related to the transport of lipids.

3.2.2 Catalytic Mechanism of P-type ATPases

Active transport of the metal by P-type ATPases follows the E1-E2 Albers-Post model by alternating the affinities of intracellular metal binding sites from high (E1) to low (E2) (Fig. 3.6) [90].

In the E1 state the ATPase has a high affinity for the metal and the Trans-Membrane Binding Domains (TMBDs) are accessible only from the cytoplasmic side. In contrast, an enzyme in the E2 state has low affinity for the metal and in this conformation the metal binding site faces the opposite side of the membrane. According to this model, the enzyme in E1 state is phosphorylated by Mg-ATP with metal ion binding to the TMBD from the cytoplasmic side (E1.ATP.nM^{+n}). The phosphorylation occurs with the transfer of the terminal phosphate of ATP to a conserved Asp residue located in the P-domain followed by the subsequent release of ADP (E1.P.nM^{+n}). This phosphorylation causes occlusion of the bound metal ion at the TMBD. The enzyme is unstable in the E1.P state and converts rapidly to the E2.P state. This transition leads to the release of the metal ions into the extracellular/luminal compartment. Finally, dephosphorylation takes place and the enzyme returns to the unphosphorylated and metal free form (E2). The enzyme then returns to the E1 conformation upon ATP (mM) binding to E2. Biochemical studies with eukaryote, prokaryote and archeal P-type ATPases have provided evidence for individual steps of the catalytic mechanism. ATPase activity, phosphorylation, dephosphorylation and metal transport studies have been carried out with isolated or membrane preparations of ATPases [91] [92].

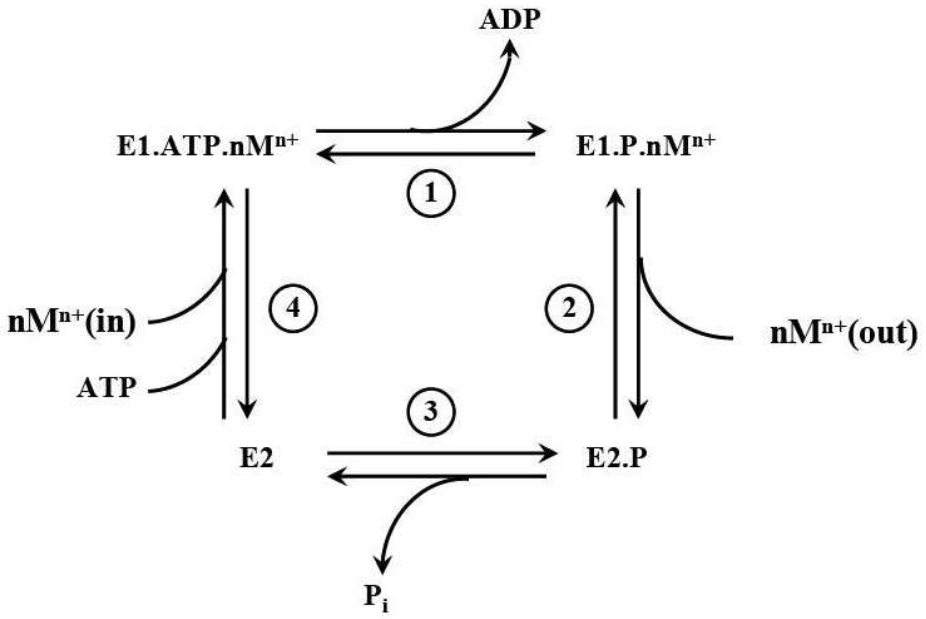

Figure 3.6: Catalytic Mechanism of P-type ATPases. E1 and E2 represent the different conformations of the enzyme. M^{+n} represent the different metals that are transported by P-type ATPases. n indicates the uncertainty on the stoichiometry of the metal transport. M^{+n} (in) represents the cytoplasmic and M^{+n} (out) represents the extracellular or luminal localization of the transported metal.

3.3 SERCA : A P-type ATPase of known tri-dimensional structure

3.3.1 Structure of SERCA

The Ca^{2+} ATPase from rabbit skeletal muscle sarcoplasmic reticulum (SERCA1a) is an integral membrane protein of 110 kda and structurally and functionally the best characterized member of the P-type ion translocating ATPases superfamily (more specifically P2A). The sarcoplasmic reticulum Ca^{2+} ATPase transports Ca^{2+} from the cytoplasm of muscle cells into the Sarcoplasmic Reticulum (SR) lumen. Transport of two Ca^{2+} is coupled to the hydrolysis of one molecule of ATP. In exchange for Ca^{2+}, two to three protons are counter-transported. Crystal structure of most major transport intermediates or analogs thereof have been obtained [93]. The structures corresponding to the four states in the catalytic cycle are shown on Fig. 3.7. The ATPase consists of a transmembrane domain (M) with 10 transmembrane helices and three cytoplasmic domains: phosphorylation (P), nucleotide binding (N) and actuator (A) domain. The P domain contains the phosphorylation site ASP-351 and the N domain interacts with ATP. In the Ca^{2+}-ATPase SERCA, the Ca^{2+} binding sites in the M domain are formed by residues in transmembrane helices TMH4, TMH5, TMH6 and TMH8 which requires that TMH4 and TMH6 are unwound near the Ca^{2+} binding sites. Helices TMH7 to TMH10 keep their position in the different transport intermediates and seem to anchor the protein in the membrane. In contrast, TMH1 to TMH6 move considerably upon Ca^{2+} binding and dissociation and upon nucleotide binding (See Fig. 3.7).

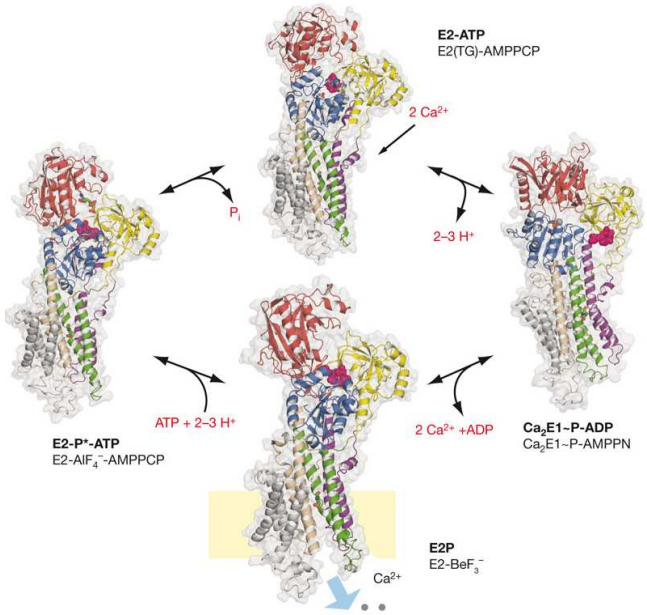

Figure 3.7: Structures of Ca^{2+} ATPase in the 4 states of its catalytic cycle. From work by Nissen *et al.* [94].

3.4 P1B-type ATPases

3.4.1 Structural Features of P1B-type ATPases

P1B-type ATPases, a subfamily of P-type ATPases, transport a variety of monovalent and divalent heavy metals across membranes using the energy of hydrolysis of the terminal phosphate bond of ATP [89] [95]. They are thought to appear in early evolution and are key proteins in the maintenance of metal homeostasis in a number of organisms including archea, bacteria, fungi, and eukarya [89].

Analysis of P1B-type ATPases sequences suggests that most have 8 transmembrane helices (TM) (Fig. 3.8) [96]. However, a small subgroup of P1B-type ATPases appears to have 6 TMs [96]. The presence of 8 TMs has been experimentally confirmed for two bacterial enzymes, *Helicobacter pylori* CadA and *Staphylococcus aureus* CadA [97] [98]. The conserved residues in TMs H6, H7 and H8 form the transmembrane metal binding domain (TMBD) and provide signature sequences that predict the metal selectivity of P1B-type ATPases [96] [99]. A large cytoplasmic loop responsible for ATP binding and hydrolysis is located between TMs H6 and H7. This loop, referred to as ATP binding domain (ATP-BD), encompasses the nucleotide binding (N) and phosphorylation (P) domains (Fig. 3.8) [100].

These two domains are separated by a "hinge" region. The smaller cytoplasmic loop between TM H4 and H5 forms the actuator (A) domain [100]. In the Ca^{2+}-ATPase, SERCA1, this loop has been shown to interact with the P-domain during the catalytic cycle [104] [93]. In addition to the TMBD, most P1B-type ATPases have regulatory cytoplasmic metal binding domains located in the N-terminus (N-MBDs), C-terminus (C-MBDs), or both. Most N-MBDs are characterized by one to six copies of a highly conserved domain containing the CXXC sequence [96]. These Cys residues can bind both monovalent and divalent cations (Cu^{+}, Cu^{2+}, Zn^{2+} and Cd^{2+}) [105] [106]. In some cases these sequences are replaced by His repeats or other diverse sequences [96] [107].

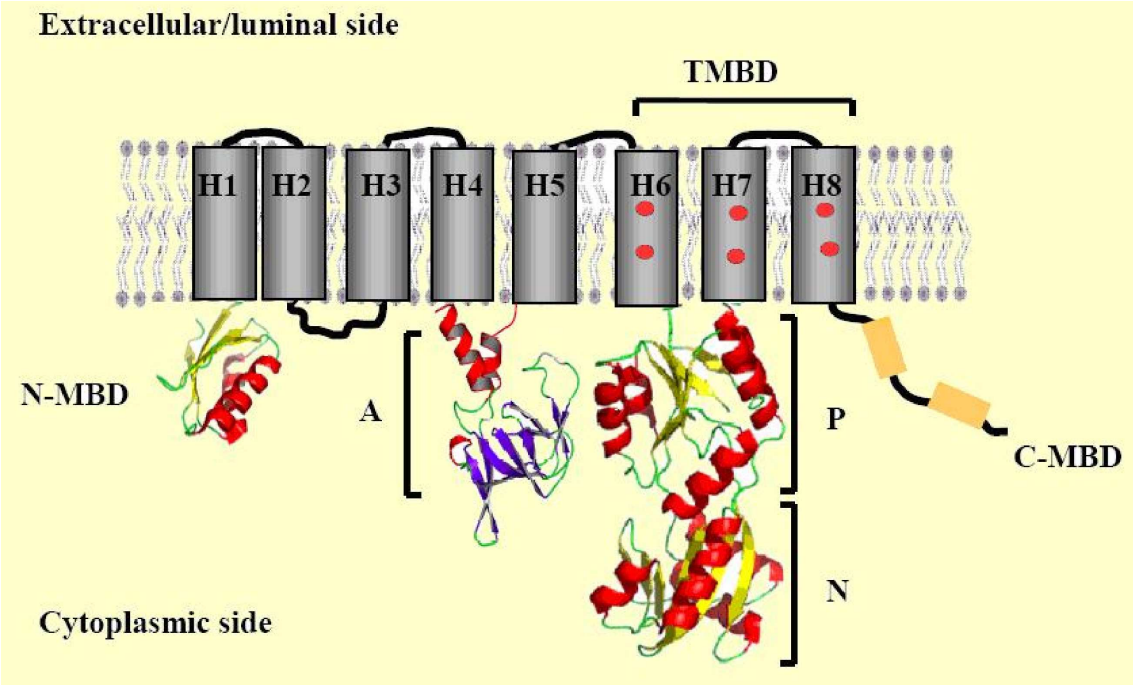

Figure 3.8: Schematic representation of the membrane topology of P1B-type ATPases. Transmembrane helices, H1-H8, are indicated. The relative locations and structure of *Archaeoglobus fulgidus* CopA actuator (A) domain and phosphorylation (P) and nucleotide (N) domains [101] [100] are shown. To represent one of the repeats present in the N-terminus the human Menkes disease protein (MNK) fifth N-terminal metal binding domain (N-MBD) [102] is depicted. The conserved amino acids in H6, H7 and H8 forming the transmembrane metal binding sites (TMBDs) are symbolized by red dots. The C-terminal metal binding domains (C-MBDs) with likely diverse structures are represented by yellow rectangles. (Adapted from a figure by Arguello *et al.* [103])

3.4.2 Physiological Roles of P1B-Type ATPases

P1B-type ATPases were first identified in bacteria like *Staphylococcus aureus* plasmid pI258 [108], *Rhizobium meliloti* [109], *Escherichia coli* [110], *Enterococcus hirae* [111] and *Synechococcus* PCC 6803 [112]. Gene knockout studies with most bacterial P1B-ATPases resulted in sensitivity of bacteria to high concentrations of metals indicating these ATPases have roles in maintaining metal quotas in the organism [111] [113]. These studies together with functional complementation assays enabled the initial determination of substrate specificity of P1B-type ATPases. Cu^{+}-ATPases, Zn^{2+}-ATPases and a Co^{2+}-ATPase have been identified in bacteria. Interestingly, functional and biochemical assays showed that these ATPases can also transport non-physiological substrates. For instance, Cu^{+}-ATPases also transport Ag^{+}. Similarly, Zn^{2+}- ATPases can transport Cd^{2+} and Pb^{2+} [110] [114]. *Archaeoglobus fulgidus* has two P1Btype ATPases, CopA and CopB, that transport Cu^{+} and Cu^{2+} respectively, suggesting the need to extrude alternative Cu forms depending on the organism's redox status [115]. In yeast the Cu^{+}-ATPase Ccc2p drives Cu^{+} export to a late- or post-Golgi compartment in the secretory pathway. Exported Cu^{+} is eventually incorporated into a multi-copper oxidase Fet3p, which translocates to the plasma membrane and works in conjunction with the iron permease to mediate high affinity Fe uptake [116] (See Fig. 1.3). In humans there are two genes (ATP7A and ATP7B) coding for Cu^{+}-ATPases: Menkes disease protein (MNKP) and Wilson disease protein (WNDP), that are associated with genetic Cu transport disorders. MNKP and WNDP mutant proteins manifest distinct phenotypes due to their differential expression patterns in human tissues. MNKP is expressed in almost all the cells except the hepatic cells. Mutations in MNKP lead to poor Cu uptake from the intestine resulting in severe neurological disorders and connective tissue abnormalities. WNDP is mainly expressed in hepatocytes and mutations in this ATPase result in high Cu levels in the liver, blood and brain causing consequent neurological disorders and cirrhosis. In the cell, both proteins are localized in a trans-Golgi compartment and undergo Cu-dependent trafficking [117]. Under conditions of high Cu, MNKP is located to the plasma membrane in various tissues [117] while WNDP is targeted to vesicles proximal to the plasma membrane of liver canicular cells where they function in Cu efflux [118].

3.4.3 Catalytic Mechanism of P1B-type ATPases

Transport experiments indicate that P1B-type ATPases drive the metal efflux from the cytoplasm [91] [92]. This is in agreement with a mechanism where the enzyme binds to ATP and the metal in the E1 state (TMBDs are open to the cytoplasmic site). Some earlier reports suggested that some Cu^{+}-ATPases might drive metal influx into the cytoplasm [111]. However, this would require an alternative mechanism where the binding of another substrate would be required in the E1 state to trigger ATP hydrolysis and enzyme phosphorylation followed by subsequent conformational changes to allow metal influx. The stoichiometry of transport has not been determined for any of the P1B-type ATPases. However, a study of *Escherichia coli* Zn^{2+}-ATPase ZntA has shown that Zn^{2+} binds to the TMBD with a stoichiometry of 1 metal bound per enzyme [119]. Although it can be argued whether all the TM metal binding sites were occupied, this is the first study towards determination of stoichiometry of metal transport by P1B-type ATPases.

3.4.4 Transmembrane Metal Binding Sites and Classification

Structural and functional characterization of Ca^{2+}- and Na^+/K^+-ATPases of P2-type ATPases indicate that conserved amino acids in their TMs H4, H5 and H6 are responsible for ion binding during transport [120] [121]. Sequence alignment and homology studies show that TMs H6, H7 and H8 of P1B-type ATPases are structurally similar to P2-type ATPases TMs H4, H5 and H6 [96]. Most P1B-type ATPases contain a CPX signature sequence in their TM H6 where the Pro is conserved in all P-type ATPases. This sequence has been proposed to participate in metal binding and transport [96] [89]. Mutations in the CPC sequence of some Cu^+-ATPases including *Caenorhabditis elegans* CUA-1, *E.coli* CopA, *Archaeoglobus fulgidus* CopA and *Saccharomyces cerevisiae* Ccc2p yielded proteins that were either unable to complement for the Cu^+-ATPase deficient yeast mutant Rccc2 or had no ATPase activity [92] [122]. For some of these proteins, it was shown that they still bind to ATP but are unable to hydrolyze it, suggesting that the turnover of the enzyme is prevented by the lack of metal binding to the TMBD [92, 123]. Alternative sequences (SPC, CPS, CPT, CPA, CPG, CPD and CPH) have also been observed in some P1B-type ATPases. P1B-type ATPases contain conserved amino acid residues in TMs H7 and H8 which were proposed to form transmembrane metal binding domains (TMBDs) together with CPX sequences in the TM H6 and determine the metal specificity of these enzymes [96]. The signature sequences in TMs H6, H7 and H8 allow the classification of P1B-type ATPases into 5 subgroups with distinct metal selectivity and functional characteristics.

Subgroup 1B-1 involves the Cu^+/Ag^+-ATPases. Some of the Cu^+ transporting members of this group have been well characterized. Proteins belonging to this subgroup are found in eukaryotes, prokaryotes and archea. These include Menkes and Wilson disease proteins which are associated with genetic Cu transport disorders in humans [124, 125], *Arabidopsis thaliana* RAN1 [126, 127], *E. coli* CopA [92, 110] and *A. fulgidus* CopA [99]. These Cu^+- ATPases have been shown to transport non physiological substrate Ag^+ and drive the efflux of the metal from the cytoplasm [92, 99, 110]. Proteins in subgroup 1B-1 have a conserved CPC sequence in TM H6. In addition, these proteins contain the conserved residues, Asn, Tyr in TM H7 and Met, Ser in TM H8. The participation of these residues in metal transport has been shown by site directed mutagenesis studies in AfCopA [99]. Mutant AfCopAs were not phosphorylated by ATP in the presence of Cu^+ and therefore were inactive. However, these were phosphorylated by inorganic phosphate (P_i) indicating that they retained the overall structure and could undergo major conformational transitions. These observations suggested the participation of conserved residues (two Cys of conserved CPC and Asn, Tyr, Met and Ser in H7 and H8) in Cu+ transport by P1B-type ATPases.

Subgroup 1B-2 involves the Zn^{2+}-ATPases. Interestingly, these ATPases have been found in archaea, prokaryotes and plants, but not in other eukaryotes [96]. Proteins in this subgroup also have the conserved CPC sequence in TM H6. However, conserved residues in TM H7 and H8 are clearly distinct from that of subgroup 1B-1 proteins. Subgroup 1B-2 ATPases contain a conserved Lys in TM H6 and Asp and Gly in TM H8. Mutations of Asp714 (Asp714His and Asp714Glu) of *E.coli* ZntA yielded inactive enzymes that were still able to bind Zn^{2+} and undergo phosphorylation in the presence of P_i [128]. However, mutant proteins could not be phosphorylated by ATP in the presence of Zn^{2+} supporting that conserved residues in TMs H6, H7 and H8 contribute to TMBD.

Members of Subgroup 1B-3 are Cu^{2+}-ATPases that are found in archaea and bacteria but not in

eukaryotes. These have a CPH sequence in TM H6 while amino acids in TM H7 and H8 are similar to those in subgroup 1B-2. These ATPases also transport Cu^{+} and Ag^{+} [115, 111]. However, studies with AfCopB indicate that Cu^{2+} produces five times more activation of this enzyme compared to that driven by Cu^{+} [115]. This is not surprising considering that imidazolium (in conserved CPH; a hard Lewis base) prefers to bind Cu^{2+} (an intermediate Lewis base) rather than Cu^{+} [96, 115]. A mutation in the CPH sequence (CPH U SPH) in Enterococcus hirae CopB resulted in lack of activity suggesting probable contribution of this sequence to TMBD [129].

Members of Subgroup 1B-4 have only 6 putative TMs [96]. In these, the large cytoplasmic loop is located between TM H4 and H5. TM H4 (corresponding to TM H6 of other P1B-type ATPases) contains a conserved SPC sequence. The substrate specificity of these enzymes has not been characterized. One member of this subgroup, Synechocystis PCC6803 CoaT, seems to be involved in Co^{2+} transport since disruption of coat gene reduced *Synechocystis* Co^{2+} tolerance and increased cytoplasmic Co^{2+} levels [130]. Another subgroup 1B-4 protein, *Arabidopsis thaliana* HMA1, is implicated in Cu^{+} transport into the chloroplast [131]. Further characterization studies of other members are necessary to delineate the metal specificity of subgroup 1B-4 proteins.

Finally, Subgroup 1B-5 includes only a few proteins that appear to be P1B-type ATPases based on the presence of typical structural characteristics including the cytoplasmic phosphorylation site (DKTGT), an APC, CPC, or CPS sequence and significant sequence similarity (30-40%) to other P1B-ATPases. Further studies are required to reveal the metal selectivity of these enzymes and the residues that participate in metal coordination.

3.4.5 Cadmium ATPases

The focus of one of my studies has been to understand the functional role and the structural functional relationships of another P1 type ATPase the Cadmium ATPase from *Listeria monocytogenes* (See chapter 4). The cadmium cation is toxic to most micro organisms, probably by binding to essential respiratory proteins and through oxidative damage by production of reactive oxygen species. Cadmium enters bacterial cells by the transport systems for essential divalent cations such as Mn^{2+} or Zn^{2+}. Microbial resistance to cadmium is usually based on energy- dependent efflux mechanisms. One of the best characterized bacterial cadmium resistance mechanisms is determined by the cadmium transporting ATPase found initially in gram positive bacteria. The cadmium transporting ATPase is a P1-type ATPase. It is widespread in *Staphylococcus* and *Listeria monocytogenes*. The ATPase is encoded by cadA, which is usually plasmid-borne and associated with transposons in *Listeria monocytogenes*. The tertiary structure of the Cadmium ATPase is not known and has been one of my research topics (See chapter 4). The experimental studies carried out so far on the transport site of Cadmium ATPase reveals the amino acids involved in the transport site of CadA. The two cysteines of the Cys-Pro-Cys motif (TM6) act at distinct steps of the transport process, Cys 354 being directly involved in Cd^{2+} binding, whereas Cys356 is required for Cd^{2+} occlusion; Asp 692 in TM8 would be directly involved in Cd^{2+} binding; Glu 164 in TM4 would be required for Cd^{2+} release. In addition it has been proposed that two Cd^{2+} ions are involved in the reaction cycle of CadA.

3.4.6 Cytoplasmic Metal Binding Domains

3.4.6.1 Structure of cytoplasmic MBDs

In addition to TMBD most P1B-ATPases have 1-6 cytoplasmic metal binding domains (MBD) located either in the N-terminus (N-MBD) (Fig. 3.9) or C-terminus or both. Most typical ones are the N-MBDs observed in Cu^{+}-ATPases and bacterial Zn^{2+}- ATPases of subgroups IB-1 and IB-2. These are usually 60-70 amino acid domains characterized by a highly conserved CxxC sequence [96]. Both conserved Cys have been shown to bind both monovalent and divalent cations including Cu^{+}, Cu^{2+}, Zn^{2+} and Cd^{2+}. The high resolution structures of several of the Cu^{+}-ATPases N-MBDs show a $\beta\alpha\beta\beta\alpha\beta$ fold that is similar to the well-described Cu^{+}-chaperones like human Atox1, yeast Atx1 and prokaryote CopZ [105]. N-MBDs have been shown to receive the metal from these chaperones. So far no Zn^{2+}-chaperone has been identified. Some Zn^{2+}-ATPases of subgroup IB-2 have His rich MBDs [(Hx)n (n=2-3)] alone or together with the typical N-MBDs (Fig. 3.9). Similar sequences have been observed in ZIP and Cation Diffusion Facilitator (CDF) families located in loops joining TMs. In bacterial Zn^{2+}-ATPase ZntA a unique CCCDGAC motif in the N-terminus has been shown to coordinate Pb^{2+} indicating that different metals might occupy different coordination environments in the same protein. Eukaryotic (plant) Zn^{2+}-ATPases contain unique sequences in both N and C-termini. All plant Zn^{2+}- ATPases lack the typical N-MBDs. In these the conserved CxxC sequences are replaced with CCxSE (x = S,T,P) sequences (Fig.3.9). In addition, Zn^{2+}-ATPases have unusually long C-termini that are either rich in His or CysCys repeats or both. Cu^{2+}-ATPases of subgroup IB-3 and only a few members of subgroup IB-4 have distinct His rich MBDs instead of the typical N-MBDs [96]. These domains contain His stretches instead of HX repeats.

3.4.6.2 Regulatory Roles of cytoplasmic MBDs

The absence of cytoplasmic MBDs in some P1B-ATPases suggests that these are likely regulatory domains [96]. Removal of N-MBD by truncation or inhibition of metal binding capability by mutation results in reduced enzyme activity with small or no changes in metal affinity [115]. It has been shown that N-MBDs of *Archaeglobulus fulgidus* Cu^{+}-ATPase CopA and Cu^{2+}-ATPase CopB control the enzyme turnover rate through the rate limiting conformational change associated with metal release/dephosphorylation. In yeast, there is no doubt, on the other hand, on the involvment of at least one of the MBDs in the transfer of the metal to another binding site in the protein *in vivo*. Morin *et al.* showed recently that the N-terminus of the ATPase plays a dual role of receiving copper from the metallochaperone Atx1 and conveying it to another domain of Ccc2 [132]. Other studies have shown the Cu^{+} dependent interaction of Wilson's disease protein N-MBDs with the large ATP binding cytoplasmic loop [133]. Studies of the human Cu^{+}-ATPases, Menkes and Wilson Disease proteins showed that at least one intact N-MBD is required for targeting of these ATPases to the plasma membrane and a vesicular compartment, respectively. Similar to Cu^{+}-ATPases, truncation of Zn^{2+}-ATPase ZntA N-MBD results in a decrease in overall rate of the enzyme without altering metal binding affinity. Morin2009

Type	Group	Length	Sequence	Protein
CXXC consensus	1B-1 1B-2	60-80	MVKDTYISSASKTPPMERTVRVTGMT **CAMC**VKSIETAVGSLEGVEEVRVNL ATETAFIRFDEKRIDFETIKRVIEDLGY GV	*A. fulgidus* CopA
CCXSE consensus	1B-2	90-100	MASKKMTKSYFDVLGI**CCTSE**VPLIE NILNSMDGVKEFSVIVPSRTVIVVHDT LILSQFQIVKALNQAQLEANVRVTGE TNFK	*A. thaliana* HMA2
$(HX)_n$ (n = 2-6)	1B-2	100-150	MNQPVS**HEHKHPHDHAH**GDDD**HGH** **A**A**HGHS**CCGAKAAPPLVQLSETASA QAQLSRFRIEAMDCPTEQTLIQDKLSK LAGIEQLEFNLINRVLGVRHTLDGTA DIERAIDSLGMKAEPIAAQDDGSASVP QPAKA	*P. putida* CadA-2
His rich	1B-3 1B-4	30-100	MNNGIDPENETNKKGAIGKNPEEKIT VEQTNTKNNLQE**H**GKMENMDQ**HH**T **H**G**H**MER**H**QQMD**H**G**H**MSGMD**H**S**H**M D**H**EDMSGMN**H**S**H**MG**H**ENMSGMD**H** SM**H**MGNFKQK	*E. hirae* CopB

Figure 3.9: Cytoplasmic N-terminus Metal Binding Domains of P1B-ATPases. Subgroups 1B-1, 1B-2, 1B-3 and 1B-4 refer to higher affinity for Cu^{+}, Zn^{2+}, Cu^{2+} and Co^{2+}, respectively. Menkes (Q04656) and Wilson (P35670) disease proteins belong to subgroup 1B-1 (from [103]).

3.4.7 The ATP Binding (ATP-BD) and Actuator (A) Domains

The crystal structures of the isolated N-domain of human Cu^+-ATPase Wilson's disease protein (WNDP) and A-domain and ATP-BD of *Archaeoglobus fulgidus* Cu^+- ATPase CopA (AfCopA) have been recently solved. AfCopA A-domain shows a 10 β-strand core with 2 α-helices connecting the TMs and its folding shows significant similarity to that of the A-domain of SERCA1, the P2-type Ca^{2+}-ATPase in spite of their little sequence homology [93].

In both, the conserved (S/T)GE(P/S) appears to be located at the tip of a solvent accessible loop on the side of the A-domain. In SERCA1, it has been shown that this loop interacts with the P-domain during phosphorylation/dephosphorylation driving the rotation of the A-domain with subsequent rearrangement of TMs [93]. This arrangement results in metal release. Structural similarity of the P1B-type ATPases A-domain with that of SERCA1 might point to a similar mechanism for metal release. The ATP-BD domain structure shows that the P- and N-domains are joined by two short loops called the hinged region [100]. The P-domain consists of a 6 stranded parallel β-sheet sandwiched between 3 short α-helices. This domain contains the conserved DKTGT sequence and shows similar folding to the P-domain of SERCA1 [121]. The N-domains of both AfCopA and WNDP consists of 6 antiparallel β-sheets flanked by 4 α-helices. Although the N-domains of both proteins show a basic similar folding to that of SERCA1, the sequence analysis reveals that the ATP binding site of P1B-type ATPases is distinct from that of P2-type ATPases. Structural analysis of the N-domains of WNDP and KdpB, a P1A-type ATPase, in the presence of nucleotides shows that these have unique homologous ATP binding sites [134, 100]. The residues that participate in nucleotide binding have been identified in WNDP (His1069, Gly1099, Gly1101, Gly1149, and Asn1150) and the involvement of some of these residues in ATP binding is supported by mutagenesis studies in WNDP, *Enterococcus hirae* Cu^+- ATPase CopB [129] and *Escherichia coli* Zn^{2+}-ATPase ZntA.

3.4.8 Human Copper ATPases

3.4.8.1 Overview

The human Cu-ATPases ATP7A and ATP7B are essential for intracellular copper homeostasis. The Cu-ATPases use the energy of ATP hydrolysis to transport copper from the cytosol into the secretory pathway and thus supply the metal for subsequent biosynthetic incorporation into various copper-dependent enzymes. ATP7A is required for formation of functional tyrosinase [135], peptidyl-α-monooxygenase [136], lysyl oxidase [137], and possibly some other enzymes [138], while ATP7B is essential for the biosynthesis of holo-ceruloplasmin, a copper- dependent ferroxidase [139]. In addition to their biosynthetic role, human Cu-ATPases participate in the export of excess copper from the cells. Overexpression of ATP7A in transgenic animals is associated with the decrease of copper content in tissues, which is particularly apparent in the heart and the brain [140]. The essential role of ATP7A in copper export from intestinal epithelium is best illustrated by the phenotype of Menkes disease. In this lethal human disorder, the functional ATP7A is lost due to various mutations in the corresponding gene, resulting in greatly impaired export of copper from the enterocytes [141].

In hepatocytes, a copper exporting role belongs to ATP7B [142]. Liver is the major organ of copper homeostasis in the body and is involved in removal of excess copper [143]. Copper is exported from

the liver into the bile and then to the feces in a process that requires the activity of ATP7B. Genetic inactivation of ATP7B results in accumulation of copper in the liver and a severe human disorder, Wilson disease. The disease is characterized by a spectrum of liver pathologies ranging from hepatitis and cirrosis to liver failure [143]. In both Menkes disease and Wilson disease, the lack of functional Cu-ATPase is also associated with the disrupted delivery of copper to the secretory pathway. The lack of copper incorporation into ceruloplasmin in Wilson disease is utilized as a biochemical marker for diagnosing the disease. In Menkes disease, the deficiency of active copper-dependent enzymes, for example lysyl oxidase, greatly contributes to the severity of the disease phenotype [144].

Two functions of human copper-transporting ATPases can be described as biosynthetic (the delivery of copper to the secretory pathway for metallation of cuproenzymes) and homeostatic (the export of excess copper from the cell). These two functions are associated with the distinct intracellular targeting of the transporters (Fig. 3.10). The localization in the trans-Golgi network (TGN), which is observed for both ATP7A and ATP7B under low copper conditions, reflects their role in the delivery of copper to copper-dependent enzymes. Such enzymes as tyrosinase, peptidyl-α-monooxygenase, and ceruloplasmin have been shown to co-localize with Cu-ATPases in the TGN and require the ATPase-mediated copper transport for formation of holo-enzyme.

It is not known whether the metallation of cuproenzymes occurs only in the TGN or if small quantities of Cu-ATPases are also present along the secretory pathway for re-metallation of secreted enzymes, if the latter looses copper. Such a scenario is possible in the case of ATP7A, since this Cu-ATPase (unlike ATP7B) can migrate towards the basolateral membrane in the same direction as secreted proteins. In fact, in the rat parotid acinar cell ATP7A is found not only in the TGN (predominant localization), but also in immature and mature secretory granules [146], where it may participate in copper delivery to peptidyl-α-monooxygenase and/or other copper-binding proteins.

The second function of Cu-ATPases - the export of copper from the cell for further utilization in the blood, milk, or for removal into the bile - requires trafficking of Cu-ATPases from the TGN to vesicles (Fig. 3.10). This relocalization occurs in response to copper elevation, hormone release, or other signaling and developmental events [117, 147]. It is thought that in response to these signals the Cu-ATPases sequester copper into the vesicles. The vesicles then fuse with the membrane releasing copper into the extracellular milieu [148]. Therefore, the regulation of intracellular localization of Cu-ATPases represents the key mechanism that determines whether the Cu-ATPases perform their homeostatic or biosynthetic function at a given moment.

Another level of regulation of copper transport must exist in cells where both Cu-ATPases are simultaneously co-expressed. While certain cells have only one Cu-ATPase (for example, ATP7B in hepatocytes), a number of cells and tissues (such as brain, mammary gland, and placenta) express both ATP7A and ATP7B. In this latter case, it is not known whether a preference exists in the distribution of copper between ATP7A and ATP7B and whether or not the same mechanisms regulate the Cu-ATPases function. Recent data from several laboratories suggest that two human Cu-ATPases differ in their enzymatic characteristics, trafficking properties, interacting partners, and regulation. It is also clear that unique sequence elements are present in the structure of two human copper pumps that may contribute to their distinct properties. Below, I summarize what is currently known about structure, function, and regulation of ATP7A and ATP7B, and speculate about possible contribution of unique sequence elements in the Cu-ATPase to their activity and regulation.

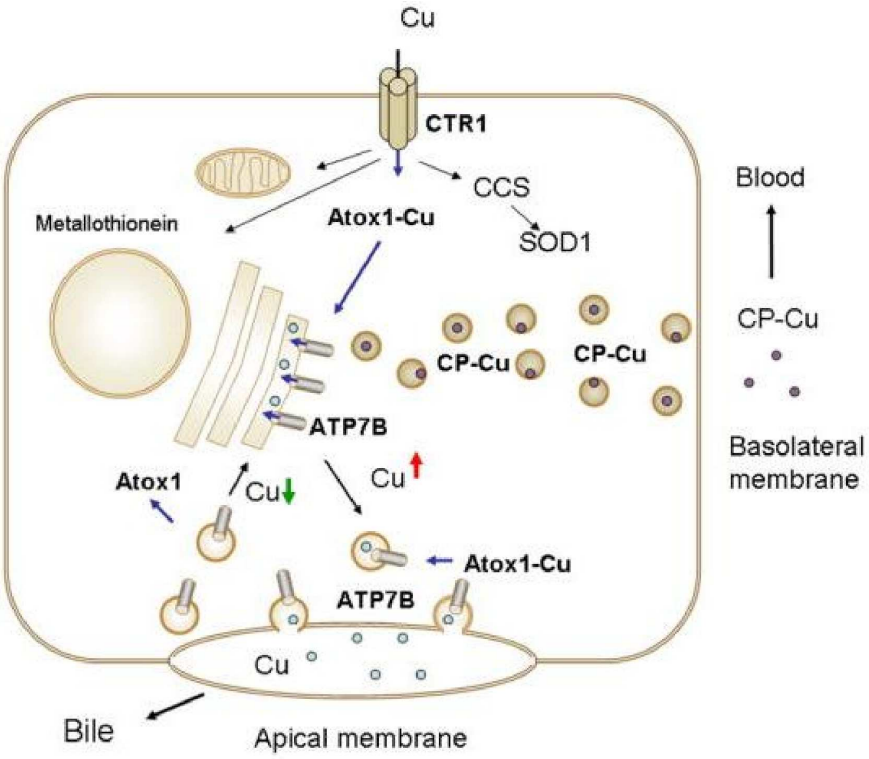

Figure 3.10: The dual role of copper-transporting ATPase ATP7B in hepatocyte. From work by Lutsenko *et al.* [145]. Copper enters the cell from the basolateral membrane via high-affinity copper transporter Ctr1 and is delivered to various cell targets with the help of copper chaperones. Atox1 (Hah1) transfers copper to ATP7B located in the trans-Golgi network (TGN). ATP7B transports copper into the lumen of TGN, where copper is incorporated into ceruloplasmin (CP), which is subsequently excreted into the blood. When copper is elevated (arrow), ATP7B traffics to vesicles. Vesicles filled with copper fuse with the apical (canalicular) membrane, copper is exported, and ATP7B is rapidly endocytosed. When copper is decreased ATP7B returns back to the TGN. It is possible that Atox1 regulates both copper delivery to ATP7B when copper is high and copper removal from ATP7B when copper is low.

3.4.8.2 ATP7A and ATP7B are representatives of the P1B-family of ion-transporting ATPases

At the biochemical level, the function of Cu-ATPases is to translocate copper across the membrane from the cytosol into the lumen of appropriate intracellular compartment (either TGN or vesicles). The vectorial copper translocation across the membranes is driven by the hydrolysis of ATP; the number of copper ions transported per one hydrolyzed ATP is currently unknown. Both ATP7A and ATP7B belong to the P1B-subfamily of the P-type ATPases (See section 3.4). Similarly to all members of the P-type ATPase family, the human Cu-ATPases hydrolyze ATP with the formation of a transient acyl-phosphate intermediate (Fig. 3.11). Phosphorylation takes place at the invariant Asp residue in the signature motif DKTG (Fig. 3.11A). The reaction requires the transfer of copper from the cytosol to the intra-membrane portion of the transporter; while the release of copper to the opposite side of the membrane is accompanied by dephosphorylation [149]. Cu-ATPases can also be phosphorylated by inorganic phosphate at the same aspartate residue within the DKTG motif. This reaction is reverse to the dephosphorylation step and is inhibited by copper binding to the intra-membrane site(s) from the luminal milieu [150].

During the catalytic cycle, the Cu-ATPases are likely to undergo significant conformational changes [151]. By analogy with other P-type ATPases, the binding of copper from the cytosolic side is thought to take place when the protein is present in the so-called E1-state, which is characterized by high affinity of the intramembrane sites for the transported ion. The measurements of copper-dependence of catalytic phosphorylation yielded apparent affinity of these sites for copper in the range of 0.7-2.5 μM [152]. Copper binding to the intra-membrane sites is associated with the transfer of γ-phosphate of ATP to the Cu-ATPase and transient stabilization of the phosphorylated state of the protein E1P (Fig. 3.11B). In this state, access to intra-membrane sites from the cytosol is blocked and copper is sequestered in the "occluded" form. Subsequently, the Cu-ATPase undergoes conformational change to the E2P state, and the affinity for copper is decreased. It is thought that in this state copper is released from the transporter and is taken up by an acceptor protein. Intermediate proteins are not required for copper transfer from the transporter to acceptors [153], however interaction between ATP7A and SOD3, a putative target protein of ATP7A activity, has been reported. This latter observation suggests that the donor-acceptor interaction, although not obligatory, may facilitate metal transfer to the copper-requiring enzymes in the secretory pathway. After copper is released, the Cu-ATPase dephosphorylates (E2 state) and then undergoes conformational transition into a high-affinity state, E1, for initiation of the next transport cycle.

3.4.8.3 Functional activity of human Cu-ATPases is coupled to their ability to traffic

The conformational changes, that take place when copper is bound to and released from the transport site(s), appear to be intimately linked to the ability of the Cu-ATPase to traffic between the intracellular compartments. Mutation of the invariant sequence motif TGE>AAA stabilizes the protein in the E2P-like state and also triggers the redistribution of Cu-ATPases from the TGN to the vesicles mimicking the response of the transporter to elevated copper [154]. Additional mutation of catalytic Asp to Glu in the background of the TGE>AAA mutant of ATP7A abolishes this effect and disrupts the trafficking from TGN under either low or high copper conditions. Similar inhibition of trafficking

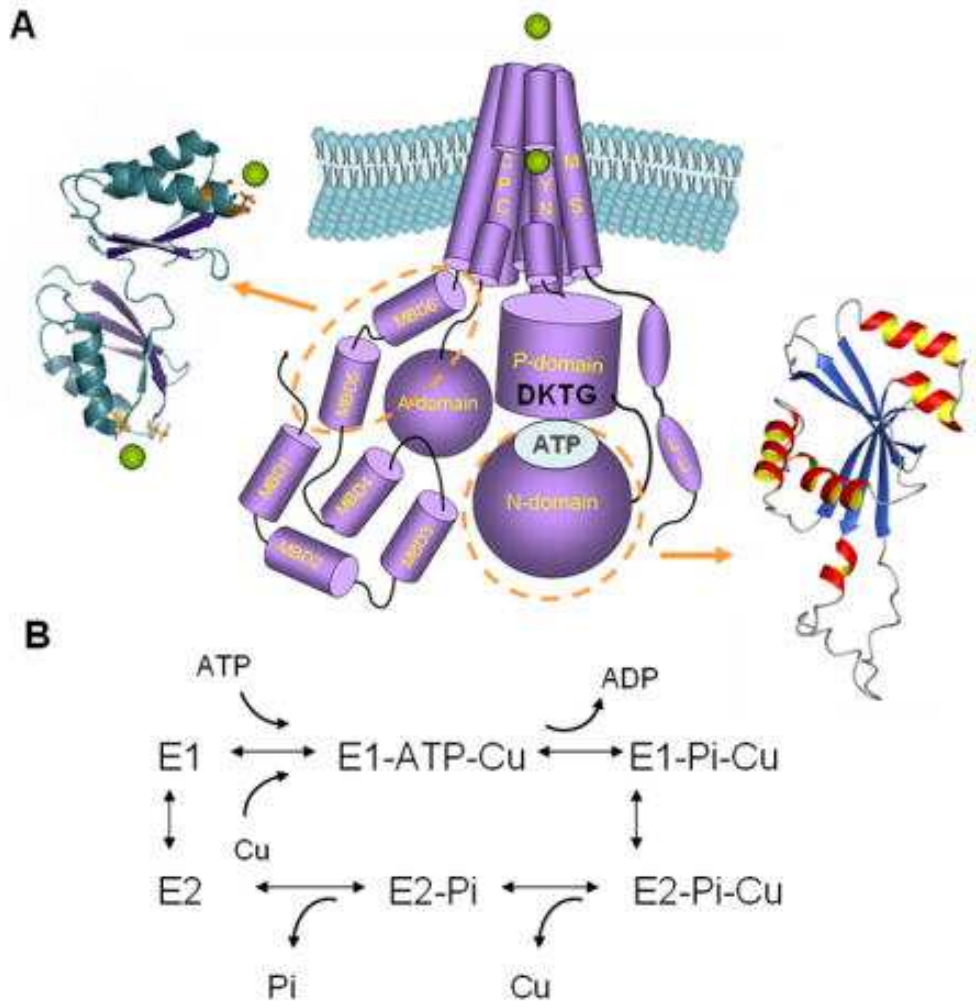

Figure 3.11: Transmembrane organization and catalytic cycle of human Cu-ATPases: From work by Lutsenko *et al.* [145]. (A) Cartoon illustrating the major functional domains of Cu-ATPases. The N-terminal domain contains six copper-binding MBDs (MBD1-6). The transmembrane portion has eight TransMembrane Segments (TMS); the position of residues predicted to be involved in copper coordination (CPC, YN, and MxxS) is indicated. The A-domain may link changes in the N-terminal domain with those in the ATP-binding domain and in the transmembrane portion. The ATP-binding domain consists of the P-domain and the N-domain. The domains of ATP7B for which structure has been experimentally determined are indicated by dashed circles and corresponding structures are shown. Two Leu residues in the C-terminal tails required for endocytosis and/or return to TGN are indicated by "LL". (B) The simplified catalytic cycle of human Cu-ATPases. Two major conformational states associated with high affinity for ATP and Cu (E1) and lower affinity for these ligands (E2) as well as phosphorylated intermediates (E1-Pi-Cu and E2-Pi-Cu) are shown

is observed when catalytic Asp is mutated to Glu in the wild-type background. Consequently, it has been proposed that the copper-dependent exit of Cu-ATPases from the TGN "requires a catalytically active enzyme and is associated with formation of the phosphorylated catalytic intermediate" [154].

Subsequent experiments utilizing the TGE>AAA mutant demonstrated that this mutant is hypersensitive to copper and traffics even when all known copper-binding sites necessary for catalytic activity and phosphorylation are inactivated [155]. Altogether, these studies pointed to the presence of additional copper binding sites in Cu-ATPases that may act as copper sensors and be involved in the initiation of trafficking. Such sites could be present in the luminal loops of the transporter and stabilization of a copper-bound conformation of Cu-ATPase (that may resemble the E2P state) could be critical for the ability of Cu-ATPases to relocate from the TGN to vesicles.

3.4.8.4 Domain organization of human Cu-ATPases

The molecular architecture of human Cu-ATPase is similar to that shown in Fig. 3.8. Both ATP7A and ATP7B have 8 predicted transmembrane segments (TMS or indefferently TMH for TransMembrane Helix) that form a copper translocation pathway. The highly conserved residues CPC in TMS6 contribute to the intramembrane copper-binding site(s). The experimental data obtained for prokaryotic copper-transporting ATPases, the disease-causing mutations, and the modeling of TMS1-6 using homology with Ca^{2+}-ATPase suggest that other possible candidates for copper coordination are residues YN in TMS7 and MxxxS in TMS8 [99]. Mutation of the corresponding residues in Cu-ATPase CopA from *A. fulgidus* disrupts copper-dependent phosphorylation of CopA in the presence of ATP, while phosphorylation by inorganic phosphate remains intact in all mutants except Y682S [99]. These observations strongly support the role of corresponding residues in copper coordination within the membrane (although in the absence of high-resolution structure indirect conformational effects of mutations cannot be fully excluded).

The transmembrane segments of Cu-ATPases are connected by loops of different length, which are fairly short at the luminal side of the transporters. The major bulk of Cu-ATPases and their key functional domains (the N-terminal domain, the ATP-binding domain, the A-domain, and the C-terminal tail) are all cytosolic (Fig. 3.11A). The N-terminal domain contains six copper-binding sites and serves as a regulatory center of human Cu-ATPases. The ATP-binding domain and the A-domain are essential for enzymatic activity of ATP7A and ATP7B and contain sequence motifs common to all P-type ATPases. Beyond these common sequence motifs, the Cu-ATPases show little primary sequence similarity to other well-known ion-pumps, such as Ca^{2+}-ATPase or Na^+, K^+-ATPase. Nevertheless, the three-dimensional fold of their ATP-binding domain and the A-domain is very similar to the corresponding domains of other P-type ATPases. This has been experimentally demonstrated by the recently determined solution structure for the nucleotide-binding domain of ATP7B [156] and by crystal structures of the ATP-binding domain and A-domain for bacterial Cu-ATPase CopA [101].

ATP7A and ATP7B share significant structural homology with each other, particularly in the core portion of their molecules, which is common to all P-type ATPases. At the same time, there are several regions in which ATP7A and ATP7B are quite different, while their orthologues display high sequence conservation in the same regions. It is tempting to speculate that the structural variability between ATP7A and ATP7B may contribute to their differences in enzymatic activity, regulation,

and/or trafficking to different membranes in polarized cells. In addition to amino-acid variations (for example, in the C-termini of ATP7A and ATP7B), the structural differences between human Cu-ATPases are manifested by the presence of well-defined inserts containing sequences unique for each ATPase. Such inserts are observed in the N-terminal domain, in the first luminal loop, and within the nucleotide-binding domain of ATP7A and ATP7B.

3.4.8.5 ATP7A and ATP7B have distinct functional properties

When compared under identical conditions in vitro, the membrane-bound ATP7A shows faster phosphorylation from ATP and it also dephosphorylates more rapidly. This observation suggests that the ATP7A turnover and copper transport rates are probably higher. Although transport rates have not been compared directly for ATP7A and ATP7B, it appears that ATP7B is indeed a slower transporter [157].

Which structural features are responsible for the distinct rates of phosphorylation and dephosphorylation of ATP7A and ATP7B? The role of the N-terminal domain in regulating enzyme turnover has been suggested for bacterial Cu-ATPases [158], however, the step of the cycle which is affected by the N-terminus remains in dispute. For human ATP7A, it was shown that simultaneous mutation of all metal binding sites within the N-terminal domain slows the rate of ATP7A dephosphorylation. This effect could be due to disregulation of conformational transitions and/or inefficient delivery of copper to the transport site(s) when cytosolic metal-binding sites are mutated. Consistent with the role of the N-terminal domain in conformational transitions, the CxxC>AxxA mutation within MBD6 of ATP7B results in the apparent shift of the E1-E2 equilibrium towards E1; the shift is evident from the increase in the apparent affinity of intra-membrane sites of ATP7B for copper [159]. These results support the notion that the MBDs that are closest to the transmembrane portion of Cu-ATPase may modulate the affinity of intra-membrane sites for copper by regulating conformational transitions of the enzyme. The sites unique for human Cu-ATPases (MBD1-4) are not involved in regulation of affinity for copper, but appear to play an auto-inhibitory role via inter-domain interactions [159].

3.4.9 Characteristic transmembrane hairpin TMS1,2 in P1B-ATPases

Six trans-membrane segments (TMS3-8) in the membrane portion of Cu-ATPases have equivalents in the structure of other P-type ATPases; while the first transmembrane hairpin (TMS1,2) is unique for the P1B-ATPases and is not found in other P-type pumps. TMS1,2 is directly linked to the large N-terminal copper-binding domain via TMS1. The sequence of the TMS1,2 hairpin is not conserved in various Cu-ATPases, therefore this hairpin is unlikely to play a direct role in copper coordination within the membrane during copper occlusion and phosphorylation steps. At the same time, the hairpin is important for the Cu-ATPase folding or function as evidenced by the "toxic milk" phenotype resulting from the Gly712Asp substitution in TMS2 [160].

3.4.10 The diverse roles of the metal-binding sites within the N-terminal domain of Cu-ATPases

The N-terminal domain of human copper transporting ATPases consists of 6 repetitive sequences (Fig. 3.11A) that are characterized by an invariant GMxCxxC motif. Each of these sequences fold into

individual domains and bind single copper ion in the reduced Cu(I) form via two cysteine residues. The role of individual metal-binding domains (MBDs) has been a subject of intense investigations [161] and a consensus has begun to emerge. It seems firmly established that the very N-terminal metal-binding sites, MBD1-4, are not required for transport function. In fact, the deletion of these MBDs in ATP7B does not alter the apparent affinity of the transporter for copper but stimulates copper-dependent catalytic phosphorylation. This result is consistent with the inhibitory role of this region in regulating the activity of Cu-ATPase. Such regulation is likely to be mediated through the domain-domain interactions within the transporter. The interactions between the N-terminal domain and the ATP-binding domain of ATP7B, which are weakened by copper binding to the former, were experimentally demonstrated [133]. Although MBDs1-4 are not essential for the transport activity of Cu-ATPases, their regulatory role is important and may contribute to a fine-tuned regulation of ATPase trafficking. The role of MBD5 and MBD6 seems to influence the affinity of the intra-membrane copper-binding site(s), most likely by shifting the E1-E2 equilibrium upon copper binding/dissociation. Copper binding to isolated MBD5,6 of ATP7B does not significantly alter their structure [162] and only small rearrangements are detected in proximity to the metal binding site. Nevertheless, mutations of metal-coordinating Cys to Ala in a single MBD6 alter the apparent affinity of intra-membrane site(s) for copper. These observations suggest that the metal-binding site of MBD6 is located in close proximity to (and perhaps interacts directly with) the other domains of ATP7B, thus influencing conformational changes of the entire protein and the affinity of intramembrane sites.

An important part of my work has been the study of the MBDs of the Menkes protein (ATP7A) (See chapter 5). I have chosen to study the relations between structure and function of these domains, for which NMR structures have been obtained, with a theoretical approach involving computer simulations.

Chapter 4

Structure and dynamics of The TransMembrane region of a Cd^{2+} ATPase

4.1 Introduction

The cadmium cation, Cd^{2+}, is toxic to most organisms and this toxicity is an interesting topic of reserach for several laboratories in my organization (CEA). In our laboratory, experimentalists have been studying the Cadmium ATPase from *Listeria monocytogenes*i, a P1-type ATPase encoded by the gene cadA (hence the name CadA I will use for the protein) and involved in Cd^{2+} transport (See section 3.4.5). Understanding the mechanism of ion translocation in CadA requires knowledge of the membrane topology of the pump. The present work aims at identifying the TransMembrane Helices (TMH) (or membrane topology) of CadA, finding the 3D structure of these helices and especially the spatial arrangement of amino acids involved in Cd^{2+} binding during the transport cycle. Experimental studies have shed some light on the structure of CadA especially on the residues involved in Cd^{2+} binding but the three-dimensionnal (3D) arrangement of the helices remains speculative.

4.2 Secondary structure prediction

Obtaining a correct sequence alignment is the cornerstone of success in all homology modelling procedures. Sequence alignment methods are ubiquitous tools for the prediction of structure and function; they are primarily used to identify related sequences via database searches and to detect template structures needed for the construction of homology models (10). Here we have done sequence alignment between Serca and CadA and the result has less than 25 % similarity. Though Serca and CadA are functionally homologous.

4.2.1 From WEB servers

We have used online transmembrane prediction servers, listed in Table 4.1 to predict the transmembrane helices of CadA. Most transmembrane prediction servers use the evolutionary information from multiple sequence alignments either directly or indirectly (11) and they often predict different sets of sequences as Transmembrane helices.

Server	Reference	URL
S1 DAS-TMfilter	[82]	http://mendel.imp.ac.at/sat/DAS/DAS.html
S2 HMMTOP 2.0	[86]	http://www.enzim.hu/hmmtop/
S3 MEMSAT3	[150]	http://bioinf.cs.ucl.ac.uk/psipred/
S4 MINNOU	[151]	http://minnou.cchmc.org/
S5 OrienTM	[152]	http://o2.biol.uoa.gr/orienTM/
S6 Phobius	[100]	http://phobius.cgb.ki.se/
S7 SOSUI	[83]	http://sosui.proteome.bio.tuat.ac.jp/sosuiframe0.html
S8 Split4	[88]	http://split.pmfst.hr/split/4/
S9 THUMBU		http://phyyz4.med.buffalo.edu/service.html
S10 TMAP	[90]	http://bioinfo.limbo.ifm.liu.se/tmap/index.html
S11 TMHMM	[84]	http://www.cbs.dtu.dk/services/TMHMM/
S12 TOP-PRED	[153]	http://bioweb.pasteur.fr/seqanal/interfaces/toppred.html

Table 4.1: Servers used for predictions

The Quality or accuracy of prediction of these servers in the particular case of P-ATPase, is calculated by comparing the position of TMH in the X-ray determined structure of Serca with the server prediction. PDB structure 1SU4 for Serca with two bound calcium ions has been used. To avoid introducing biases due to the fact that some servers were not specialized in TMH predictions, we deleted from the pdb all ATOM lines corresponding to residues which were not listed as belonging to a TMH (leaving 4 extra residues before and after each identified TMH).

The formula for calculating the quality of the prediction is :

$$Q = \frac{RTMPTM - (RTMNPTM + NRTMPTM)}{RTM}$$

where

[-] RTMPTM = Number of AA in Real TMH Predicted in TMH

RTMNPTM = Number of AA in Real TMH Not Predicted in TMH

- NRTMPTM = Number of AA in Not Real TMH Predicted in TMH

- RTM = Total Number of AA in Real TMH

The quality of prediction (Q) for Serca TMH is listed in Table 4.2. Q has a maximum value of 1 when all AA in TMH are predicted in TMH and a minimum value of -TOTR/RTM if TOTR is the total number of residues in the sequence. In the case of Serca, the total number of residues was 304 and RTM was 204, so that Q could theoretically vary between -1.39 and 1.

The same servers were used to predict the number of TMHs of CadA which we know by experiment should be 8. The 5 servers with prediction quality $Q > 0$ found 8 TMH for CadA.

The prediction results for the TMHs of CadA by the 3 servers with highest Q value (S4, S6 and S7) are given in Table 4.3. The results from the 3 servers agree for TM1, 2, 3 and 5 but problems are visible for correctly identifying the other 4 TMHs. However Server S6 seems to show a good consensus

Server	Predicted number of TMH	PTM	RTMPTM	RTMNPTM	NRTMPTM	Q
S1	8	115	91	59	24	0.04
S4	11	259	182	25	77	0.39
S6	8	181	132	20	49	0.31
S7	5	255	177	27	78	0.35
S12	10	210	132	52	78	0.009

Table 4.2: Server predictions for TMH in Serca. Only those servers are listed with Q value greater than zero. Total number of amino acid residues in real TM of serca (RTM) is 204. PTM = Total number of residues which were predicted as belonging to TMHs.

between all 3 servers: it agrees with S7 for TM4 and 6 and with S4 for TM7 and 8. Therefore, results from S6 only will be discussed in the following.

S4	S6	S7
W89 - Q105 (TM1)	W89 - M107(TM1)	R90 - Q105 (TM1)
D111 - L138 (TM2)	Y113 - F131(TM2)	Y113 - G134 (TM2)
E146 - S182 (TM3)	F143 - G160(TM3)	L139 - G160 (TM3)
T311 - L334 (TM5)	S166 - M183(TM4)	E164 - R180 (TM4)
E340 - P355 (TM6?)	Y315 - F335(TM5)	A313 - F335 (TM5)
A357 - K375 (TM6?)	L347 - G371(TM6)	G346 - N372 (TM6)
I660 - G681 (TM7)	I659 - I679(TM7)	P611 - A625 (TM7?)
L683 - R705 (TM8)	L685 - L704(TM8)	N663 - A688 (TM8?)

Table 4.3: Sequence ranges of the predicted TMHs of CadA by the three servers S4 (Minnou) S6 (Phobius) and S7 (Sosui). TMHs which could not be identified clearly due to absence of consensus are marked with ? signs.

4.2.2 From experiments

Plasmid Pl258 carries the cadA gene that confers resistance to cadmium, lead and zinc. The membrane topology of CadA was experimentally determined by constructing fusions with the reporter genes phoA or lacZ [98]. A series of C-terminal truncated CadAs were fused with one or the other reporter gene and the activity of each chimeric protein determined. Alkaline phosphatase and β-galactosidase activity assays allowed the authors to mark some specific residues as cytoplasmic or periplasmic and thus have a rough estimate of the TMH location as shown in Fig. 4.1. From these results, the authors identified TMHs as shown in Table 4.4 column 4.

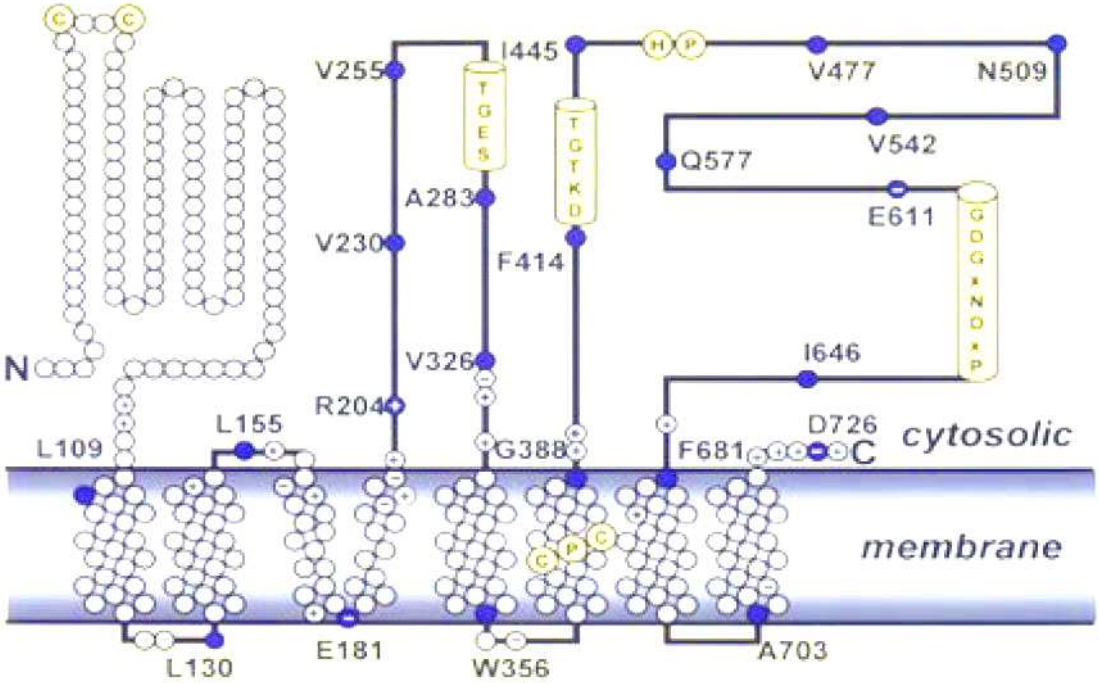

Figure 4.1: Membrane topology of CadA. The cylinders represent motifs conserved in all P-type ATPases. The filled circles indicate the residues identified as extra-membranous in the P1258 CadA-ATPase. (This is Fig.5 of Ref; [98])

4.2.3 By homology with CopA

CopA, a P1B-type ATPase plays a major role in the resistance of the cell to copper by effecting the export of the metal across the cytoplasmic membrane. CopA is predicted to have eight transmembrane segments and two N-terminal soluble domains.

The topology of the TMHs of CadA, "where they begin and end in sequence", has been derived from a homology model built from CopA and discussed in detail in section 4.3.2. The results are summarized in Table 4.4 column 5. Results between the Web server and the CopA model are comparable. TMH8 is predicted significantly longer in the CopA model than from the server (26 amino acids instead of 20). This is due to a large tilt of TMH8 in Serca used to build the CopA model. Definition of the TMHs of CadA from the Web server (column 3) have been used in the 3D models of CadA defined in the next sections.

4.3 3D structure prediction by homology with other ATPases

As already mentioned, the sequence identity between CadA and other proteins of known 3D structure is low and falls below the authorized region for homology modelling described in Fig. 2.3. If we restrict ourselves to the sole TM part, functional and structural homology can be found between CadA and Serca or a model of the copper ATPase CopA whose TMHs are themselves deduced from a structure of Serca.

TMH	Uniprot	Web server	P1258 CadA	CopA model
TM1	89 - 109	89 :107	105:123	87:111
TM2	111 - 131	113:131	131:151	113:133
TM3	151 - 171	143:160	164:180	141:160
TM4	-	162:178	182:192	162:182
TM5	317 - 337	315:335	332:356	315:335
TM6	347 - 367	347:371	363:391	346:371
TM7	669 - 689	659:679	677:697	661:683
TM8	-	685:704	699:719	684:709

Table 4.4: Predictions of CadA TransMembrane Helices (TMH): Column 2 : results from the Uniprot web server (http://www.uniprot.org) for accession number Q60048 (CadA2-LISMO); Column 3 : result from best TMH prediction server S6 (See Table 4.3); Column 4 : result from the experimental setup by Rosen *et al.* [98] and Column 5 : prediction from a homology model based on CopA (See Table 4.6 and section 4.3.2.

4.3.1 Homology with Serca

Sequence alignments between Serca, the Ca^{2+} ATPase and CadA are unaccurate because the homology is very low between the 2 proteins. A slightly larger homology has been found between the TMHs 1 to 6 in Serca and TMHs of CadA. In Serca, amino acids in helices 4, 5, 6 and 8 constitute the 2 binding sites for Ca^{2+}. These helices would then be functionaly homologous to TMHs 3, 6, 7 and 8 in CadA (See section 4.4.1). Model building of CadA directly from the structure of Serca was not attempted. We used instead results from other authors who built the structure of a copper ATPase: CopA whose TMH part was inferred from the coordinates of Serca TMHs. We wrote a CHARMM script *findwidth* as yet another method to find the limits of the TMHs of a protein knowing its 3D structure and the assumed width of the membrane. This script is explained in the legend of Table 4.5 where it is applied to the structure of Serca with two bound Ca^{2+} ions (PDB code 1SU4). Comparison between columns 2,3 and 4 show that the results of "findwidth" are valid for finding TM helix ranges in a TM protein. They are close to results from sequence analysis and the residues predicted in TMH are really part of a TMH as confirmed by the analysis of the PDB.

4.3.2 Homology with CopA

Alignments between all types of P1B ATPases in the region of TMHs 3, 6, 7 and 8 of CadA are shown in Fig. 4.2. Some homology is visible between CadA (line CadALm) and CopA (line CopAEh). The conserved CPC motif in helix 6 is clearly visible. As mentioned above, the structure of CopA from *Archaeoglobus Fulgidus* has been solved experimentally but for its extra-membrane part only [163]. The coordinates of the amino acids in the TM part have been adjusted from those of corresponding residues in Serca. TMH1 and TMH2 of CopA have no counterpart in Serca whereas TMH4 to TMH8 of CopA have been "fitted" to TMH2 to TMH6 of Serca in its E2 form (See Table 4.6). The assignment of TMH1 to TMH3 by Wu *et al.* [163] was then somewhat arbitrary; they were placed such as to bridge

TMH	Uniprot	PDB	"findwidth"	Cα_z
1	49 : 69	48:54 + 54:79	50:70	-14.3 to 13.5
2	90 : 110	85:118 + 119:123	90:109	-13.4 to 14.7
3	254 : 273	247:274 + 275:279	256:276	-15.7 to 14.7
4	296 : 313	289:307 + 310:329	293:311	-15.2 to 14.4
5	758 : 777	739:782	762:782	-14.0 to 14.2
6	788 : 808	788:800 + 801:808	784:805	-15.1 to 13.8
7	829 : 851	830:857	834:857	-14.1 to 13.7
8	898 : 917	887:892 + 893:916	895:914	-14.2 to 14.1
9	931 : 949	926:929 + 930:950 + 951:957	933:952	-14.0 to 14.0
10	965 : 985	963:975 + 975:992	960:982	-13.3 to 15.2

Table 4.5: Definition of Serca TMHs and the *findwidth* script: Col 2 Uniprot prediction for entry P04191 Serca1A; Col 3 From information contained in the PDB file: 1SU4; col 4 calculated by "findwidth" from 1SU4. The *findwidth* script works as follows: The coordinates of the protein are read from the PDB and, if necessary, a "reduced" model of the TMHs is built, where big extramembrane loops of the protein are deleted (See section 4.6.1.2 in the case of Serca). 1 (or a few) helix:TMHc, in the resulting TMH bundle, is then chosen as having its principal axis closest to the normal to the membrane (z). This can be done in an iterative way by calculating the helix tilt angle *a posteriori* and checking that the chosen helix was indeed a good "model" of the membrane normal. Then the bundle is oriented by aligning the principal axis of TMHc with z and placing the center of mass of the bundle in (0,0,0). Finally, and for a membrane width of w Å , the Cα atoms with Z-coordinate closest to $-w/2, w/2$ are selected as belonging to the residues defining the "ends" of the TMHs. These coordinates are shown in column 5 in the case of Serca and for a membrane thickness $w = 28$ Å .

the metal binding domain (MBD) reasonably to the rest of the molecule.

TMH	Serca	"findwidth" on Serca	CadA "equivalent"
1	Y36 - W77 (1)	V53 - W77	K87 - D111 (25)
2	W967 - A988 (10)	T967 - I987	Y113- E133 (21)
3	W832 - W854 (7)	L833 - W854	F141- G160 (20)
4	T86 - A115 (2)	T86 - V106	W162- S182 (21)
5	E243 - H278 (3)	K252 - W272	Y315- F335 (21)
6	I289 - L336 (4)	T295 - T317	G346- G371 (26)
7	E749 - A780 (5)	M757 - A779	K661- L683 (23)
8	P789 - F809 (6)	P784 - P809	T684- V709 (26)

Table 4.6: Prediction of the range of residues constituting CadA TMHs from a homology with CopA. The structure of the TMHs of CopA were themselves directly deduced from the structure of Serca (PDB code 2EAR). Col2: sequence of Serca used in the CopA model (TMH number in parentheses); Col.3: selected range of residues forming Serca TMHs found with the *findwidth* script for a membrane width of 28 Å ; Col.4: corresponding range of residues in CadA (number of residues in parentheses).

We built a model of CadA TMHs by homology with the TMHs used in the CopA model and thus indirectly from the structure of Serca in its E2 form with bound thapsigargin (PDB code 2EAR). The matching between amino acids in Serca and in CadA is done according to Table 4.6. We ran the "findwidth" script on the TMH bundle of the 8 helices of Serca (TMH1 to 7 + TMH10) (See Table 4.6), for a membrane width of 28 Å . Then we used the coordinates of the residues in the bundle as template to build the model of CadA. As many atom coordinates as possible were directly transfered from the Serca sequence to the CadA sequence (all backbone atomes + Cβ atoms + ...) In the case of TMH3 and TMH6 a direct matching was not satisfactory. TMH3 was a little bit shortened because acidic residues K140 and E161 (See Fig. 4.9) are likely to lie outside the membrane. TMH6, on the contrary was a little bit extended due to the big kink found in the Serca structure.

4.4 *Ab initio* Structure prediction

Ab initio 3D structure prediction of a protein is a tremendous task. The problem of predicting the TMHs of a TM protein, although much simpler, remains very difficult in practice. Considering that all TMHs are orthogonal to the membrane plane, the 3D problem can be turned into a 2D equivalent where the X-Y center of each TMH occupies a point on a 2D-grid as shown in Fig. 4.3. Our choice of a triangular grid comes from the inspection of the projection of the TMHs of Serca on the X-Y plane shown in Table 4.7(Left). The corresponding 2D-grid representation is shown in Table 4.7(Right).

Each helix seems to interact mainly with 2 to 5 or 6 neighbour helices which can be easily modelled on a triangular or hexagonal grid. This is confirmed by an analysis of the interaction energies between Serca helices calculated using the CHARMM force field in vacuum and shown in Table 4.8. TMH1 interacts mainly with TMHs 4 and 2. TMH7 interacts mainly with TMHs 5 and 8. TMH 8 interacts with TMHs 5,6,7,9 and 10...

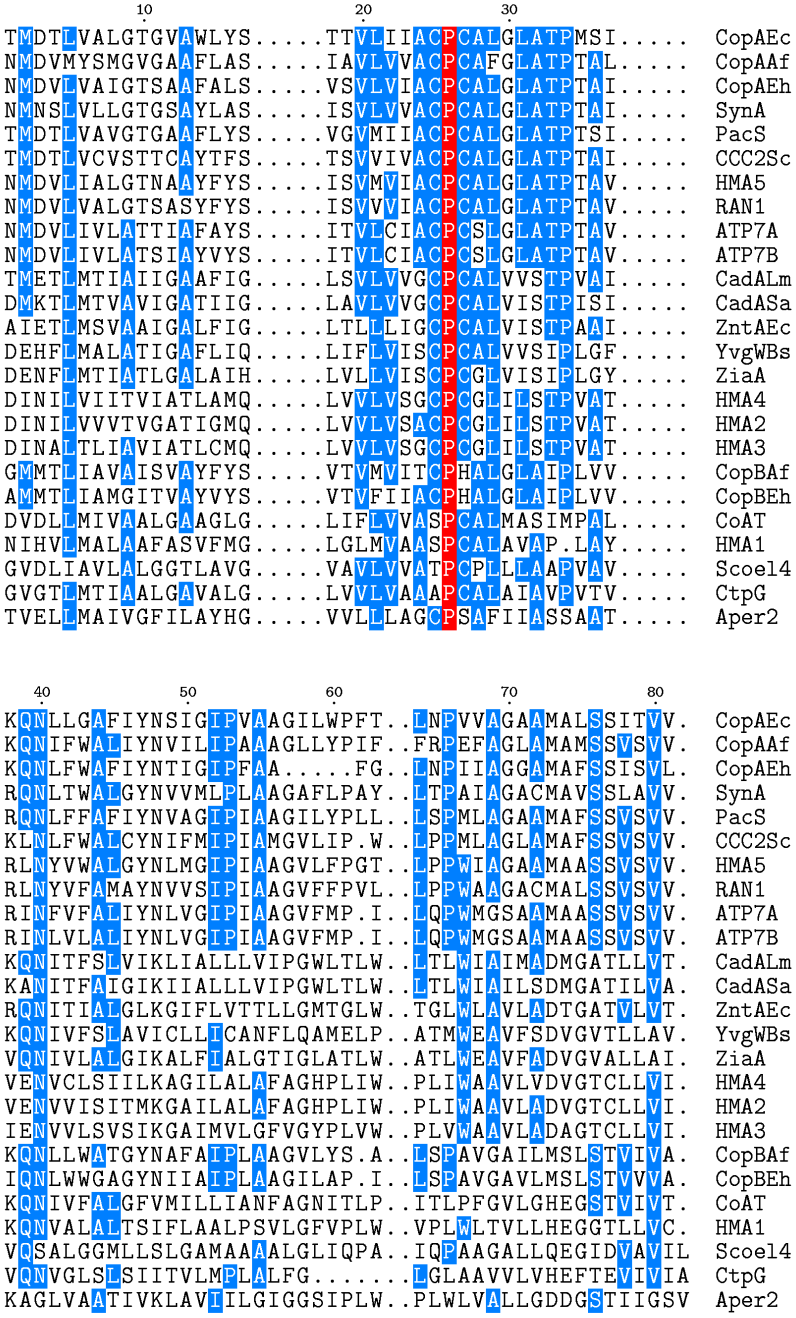

Figure 4.2: Alignment of TMHs 3, 6, 7 and 8 of several members of P1B-type ATPases of subgroup IB-1 (CopAEc to ATP7B), IB-2 (CadALm = CadA to HMA3), IB-3 (CopBAf to CopBEh), IB-4 (CoAT to HMA1) and IB-5,6 (Scoel4 to Aper2). (See classification in section 3.4.4). The proteins of main interest in this work are named: CadALm for CadA from *Listeria monocytogenes*, CopAAf for CopA from *Archaeoglobus Fulgidus* and ATP7A for the Menkes ATPase.

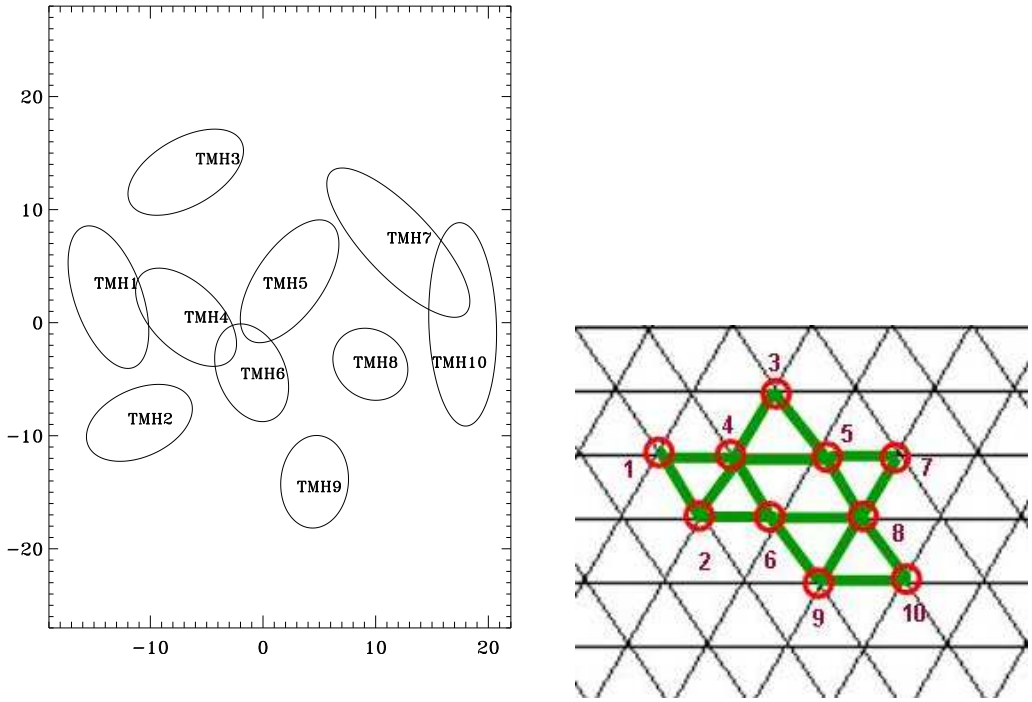

Table 4.7: Left: Schematic representation of the projection of Serca TMHs on the X-Y membrane plane. The center of each ellipse corresponds to the center of the helix principal axis calculated from the coordinates of C_α atoms. The short axis of each ellipse is constant for all TMHs. The long axis corresponds to the projection of the principal axis on the X-Y membrane plane. Right: Grid representation of Serca X-ray structure.

TMH	2	3	4	5	6	7	8	9	10
1	-13.90	-11.74	-44.07	-0.24	-0.51	0.00	0.00	0.00	0.00
2	-	0.00	-2.82	0.01	-30.16	0.00	0.00	-0.18	0.00
3	-	-	-18.55	-10.26	-1.11	-0.35	0.00	0.00	0.00
4	-	-	-	-24.52	-21.78	-0.00	0.25	0.00	0.00
5	-	-	-	-	-27.33	-32.13	-30.17	-0.23	-6.39
6	-	-	-	-	-	-3.30	-25.24	-27.15	-0.04
7	-	-	-	-	-	-	-28.97	0.00	-25.40
8	-	-	-	-	-	-	-	-16.04	-38.66
9	-	-	-	-	-	-	-	-	-20.49

Table 4.8: Interaction energies (kcal/mol) between TMHs of Serca calculated from the X-ray structure with two bound Ca^{2+} ions (PDB 1SU4).

4.4.1 Core TM bundle responsible for metal binding

From the example of Serca, we have learnt that each TMH interacts significantly with at least two other helices. Then, what are the possible arrangements of 8 TMHs on a 2D-grid under the constraint that each TMH should interact with at least two other TMHs? An upper bound to the number of these arrangements (topologies) can be easily estimated. Starting from the single arrangement of 3 TMHs forming a triangle; a fourth helix can be added at 3 different positions from the 3 vertices of the triangle; similarly a fifth helix can be added at 4 different positions... Finally the number of arrangements will be slightly less than (N_{TMH} - 1)!/2 = 2520 in the case of N_{TMH} = 8 TMHs. Building and testing all these models thoroughly would be difficult to achieve. Fortunately, in the case of P-type ATPases, experimental results allow us to further simplify the problem. Indeed, experiments show that 4 TMHs are always involved in metal binding. Moreover, these helices can be placed in the 2D-grid representation, on the summits of a diamond as shown in the case of Serca TMHs (3, 4, 6 and 8) in Table 4.7.

Similarly and based on experimental results by Catty *et al.* [164], CadA TM residues M149 in TMH3, C354, P355 and C356 in TMH6 and D692 in TMH8 play a significant role in cadmium binding. E164 in TMH4, on the other hand is probably important for cadmium transport. Very recently, Lubben *et al.* proposed a structure of CopA from *Enterococcus hirae* [165] by combining topological information obtained by intramolecular cross-linking with molecular modelling. These authors find important residues for Cu^{+} binding in TMH7 of CopA. Under the hypothesis of 2 cadmium binding sites in the TM part of CadA, C354, C356 and D692 could form site I while M149, C354 and D692 proposed to be at the interface with the extracellular medium would constitute site II. From these results, we constructed models of CadA around a basis constituted of TMHs 3,6,7 and 8. Starting from this basis of 4 TMHs which bind the metal, a homemade C^{++} program "buildtopo" was run to find all possible arrangements of the other 4 TMHs of CadA. For instance, as shown in Fig. 4.3, starting from TMH labelled A, part of the basis (A,B,C,D), 4 possible locations are accessible to a TMH helix defined as neighbour to A.

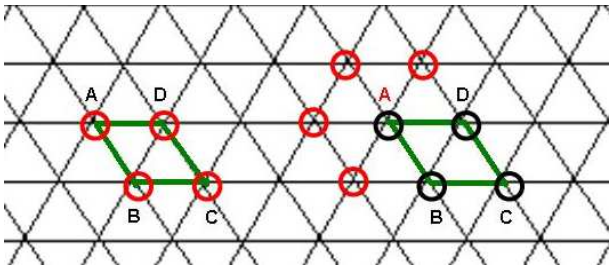

Figure 4.3: Selection of possible model topologies from a 4-TMH basis. In this grid representation, basis TMHs are located at the summits of a diamond. If A is one of these basis helices, 4 possible locations are accessible to a TMH helix defined as neighbour to A.

4.4.2 Topology of CadA TM bundle: program "buildtopo"

The flowchart of the program "buildtopo" is shown in Fig. 4.4.

The program uses two arrays: c_M[i][j] which contains the coordinates of helix i (integer coordinates

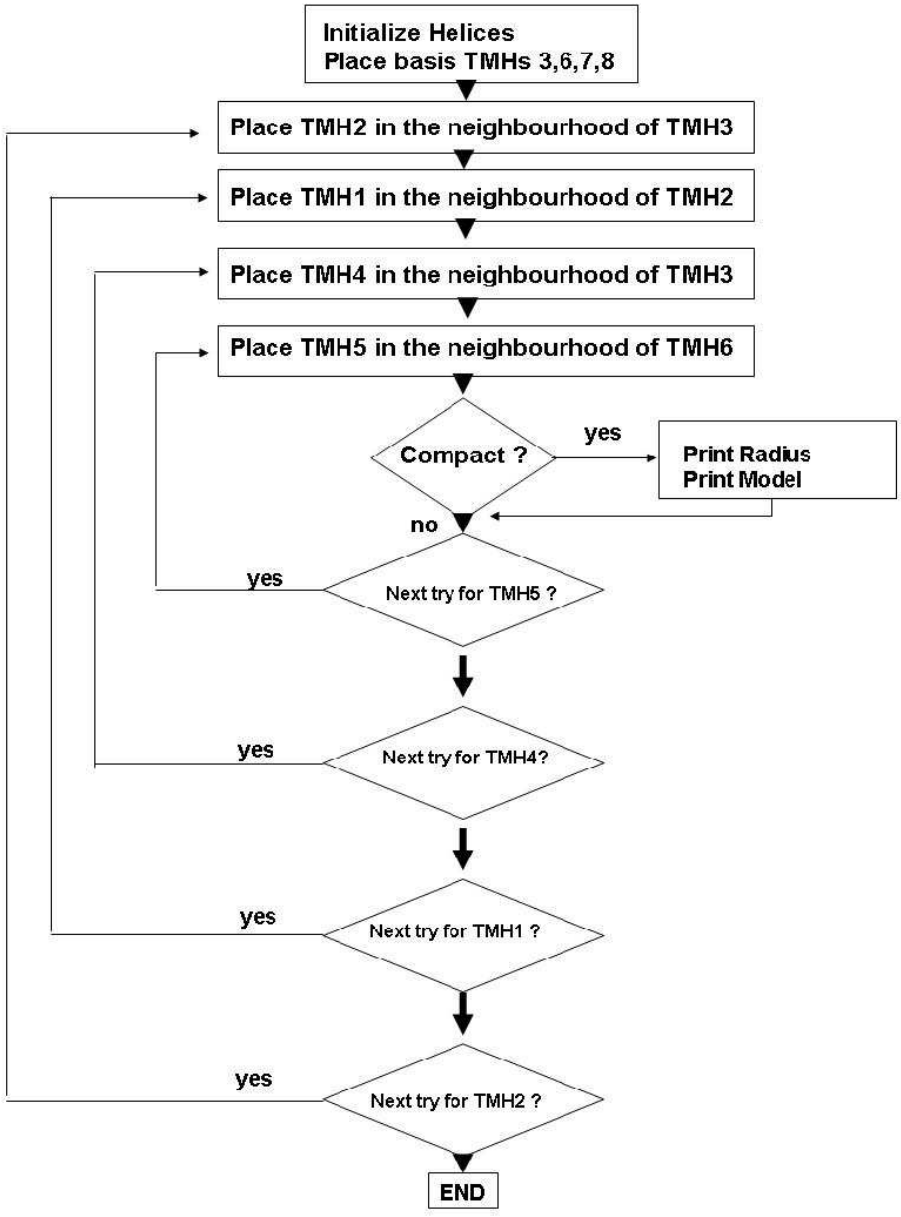

Figure 4.4: Flowchart of the program "buildtopo" applied to the structure of CadA with basis helices 3, 6, 7 and 8.

in grid space, line coordinate for $j = 0$, column coordinate for $j = 1$). and c$_G$[i][j] which contains the helix number found in line i and column j of the hexagonal grid. Each helix on the grid located at position (c$_M$[i][0],c$_M$[i][1]) may have a maximum of 6 neighbour helices (at the same arbitrary distance 1 of i) located at positions (c$_M$[i][0],c$_M$[i][1]+1), (c$_M$[i][0],c$_M$[i][1]-1), (c$_M$[i][0]+1,c$_M$[i][1]), (c$_M$[i][0]-1,c$_M$[i][1]), (c$_M$[i][0]+1,c$_M$[i][1]-1), (c$_M$[i][0]-1, c$_M$[i][1]+1). The program is initialized with the fixed coordinates of the basis (TMHs 8,3,6,7). Then the program attempts to place TMH2 neighbour to TMH3, TMH1 neighbour to TMH2, TMH5 neighbour to TMH6 and finally TMH4 neighbour to TMH3. Every time, the attempt is a success if the helix just placed has at least two neighbour helices. Finally, the radius of the attempted model (function averageRadius) is calculated as the average euclidian distance of all helices to the center of gravity of the bundle. Each side of the triangle measures 1 in arbitrary units. We first need to transform helix coordinates from grid space to real world. The first column on the grid scheme is shifted to the right by 1/2 every new line (See program output in chapter 7). The abscissa hx(i) of helix i in real world is then c$_M$[i][0]/2 + c$_M$[i][1] while its ordinate hy(i) is c$_M$[i][0].$\sqrt{3}$/2 ($\sqrt{3}$/2 being the height of a triangle). The coordinates of the center of gravity are then

$$x_G(\text{resp.}y_G) = 1/8 \sum_i hx(i)(\text{resp.}hy(i))$$

Then the radius R is given by

$$R = 1/8 \sum_i \sqrt{((hx(i) - x_G)^2 + (hy(i) - y_G)^2)}$$

A listing of the program is given in the supplementary material section together with the first page of output of the grid schemes. In the case of CadA, 27 models were found and are shown in the supplementary material section and 12 out of 27 can be considered as more "compact" with radius $R \leq 1$. I present in Fig. 4.5 a grid diagram of 4 selected models.

4.4.3 Derivation of model restraints for X-PLOR

As mentionned in a previous chapter (See section 2.1.1), NOE-type distance restraints can be derived from the topologies of the models.

Each TMH corresponds to a regular N_A amino acid long α-helix by definition. This allows the definition of $N_A - 4$ hydrogen bonds and thus distance restraints between O atom of residue i and H atom of residue $i + 4$. In the case of TMH1 of Serca, for instance, extending from residue W5 to C25 (See Fig. 4.7), we define distance restraints between W5-O and I9-H, E6-O and E10-H... until L21-O and C25-H. The form of the restraining potential is given in Eq. 2.1, with $S = 1, d = 2\text{Å}\,, d_{minus} = d_{plus} = 0.2\text{Å}\,$.

Similarly and equivalent to the NMR-J-coupling derived restraints, we can define for the same N_A amino acid long right-handed α-helix 2.($N_A - 1$) angle restraints corresponding to ϕ and ψ dihedral angles. In the example of Serca TMH1, we define restraints for dihedrals formed by atoms W5C-E6N-E6Cα-E6C (ϕ), E6N-E6Cα-E6C-L7N (ψ)... until A23C-A24N-A24Cα-A24C (), A24N-A24Cα-A24C-C25N. The form of the restraining potential is given in Eq. 2.2, with $\phi_o = -60$ and $\psi_o = -50$ degrees, $\Delta\phi = 5$ degrees.

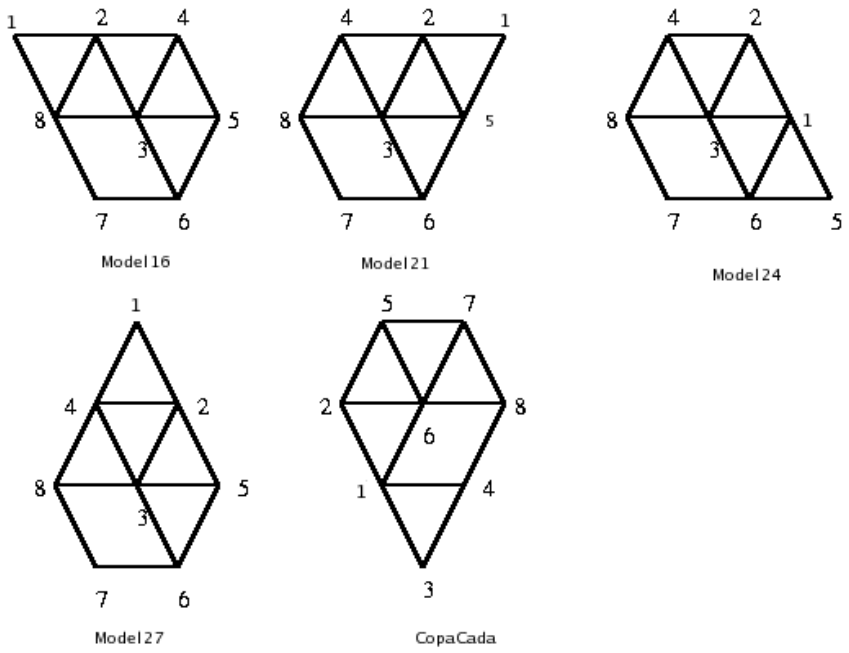

Figure 4.5: 2D-grid diagrams for CadA: we present 4 topologies extracted from the 27 models calculated with the program "buildtopo", namely models 16, 21, 24 and 27. All these models are based upon the work by Catty *et al.* [164] which leads to a basis for binding Cd^{2+} constituted of TMHs 3,6,7 and 8. The Last model (noted CopACadA) is based on the contrary on the work by Wu *et al.* [163] which appears to place TMHs 8,4,1 and 6 as basis for binding Cu^{+} (See Fig. 6 in their work).

NOE-like distance restraints were also introduced between atoms which are known to bind the metal. These restraints are listed in Table 4.9 in the case of models of Serca and CadA. The form of the restraining potential is given in Eq. 2.1, with $S = 1, d = 5\text{Å}\,, d_{minus} = d_{plus} = 1\text{Å}\,$.

Serca	E116 OE* - D168 OD*	E139 OE* - D168 OD*	E233 OE* - D168 OD*	E116 OE* - E139 OE*	E116 OE* - E233 OE*	E139 OE* - E233 OE*
CadA	K181 NZ - D202 OD*	M60 SD - C144 SG	M60 SD - C146 SG	C144 SG- D202 OD*	C146 SG - D202 OD*	M60 SD - D202 OD*

Table 4.9: X-PLOR restraints on residues involved in metal binding: the first restraint, for example, means that a distance restraint is applied between ϵ-Oxygens of Glu 116 (E116 OE*) and δ-Oxygens of Asp 168 (D168 OD*). (SD = δ-Sulfur, SG = γ-Sulfur)

Most importantly, the distance restraints which define and differentiate the models were also applied between Cα atoms of terminal residues of TMHs. There is one such restraint per vertex in the 2D-grid representation of each model. For instance in the case of the Serca X-ray grid model (See Table 4.7 Right), there are 16 such restraints. The form of the restraining potential is given in Eq. 2.1, with $S = 1, d = 10(\text{resp.}11)\text{Å}\,, d_{minus} = d_{plus} = 1(\text{resp.}2)\text{Å}$ for the CadA and Serca models, respectively.

Finally, planar restraints were added during the simulated annealing and refinement phases in X-PLOR to ensure that the helices are built orthogonal to the membrane plane. These restraints were applied on TMH terminal C_α atoms using a harmonic potential described in Eq: 2.3.

4.4.4 Model building

A flow chart of our model building strategy is shown in Fig. 4.6.

Distance and dihedral angle restraints being properly set, the standard X-PLOR scripts for building models from distance geometry (See section 2.1.4) were run. For each topology, we constructed 100 models and kept the model with lowest X-PLOR energy. Protein models built with X-PLOR or CHARMM correspond to a continuous sequence of amino acids linked together through the peptide bond. ATPases are proteins with large cytoplasmic loops, we did not model in our studies. Consequently, so-called "reduced" models of the protein were used where only the TMHs are kept preceded and followed by 4 amino acids from the real sequence. All interhelical loops including the so-called "small" and "big" loops of the proteins are replaced by 3 glycine residues. The definitions of the TMHs of "reduced" sequences of Serca and CadA which were used in X-PLOR models and MD simulations are shown in Table 4.10. These definitions are, of course, directly deduced from our previous results: Table 4.5 col.4 for Serca and Table 4.4 col.3 for CadA. The corresponding sequences are shown in Fig.4.7 and Fig. 4.8 for Serca and CadA, respectively. The sequence using real numbering for CadA is also shown in Fig. 4.9.

4.4.5 Model refinement with CHARMM

Due to the application of planar restraints, X-PLOR models deliver TMH bundles already oriented along the Z-axis orthogonal to the X-Y membrane plane. In the case of simulations from X-ray

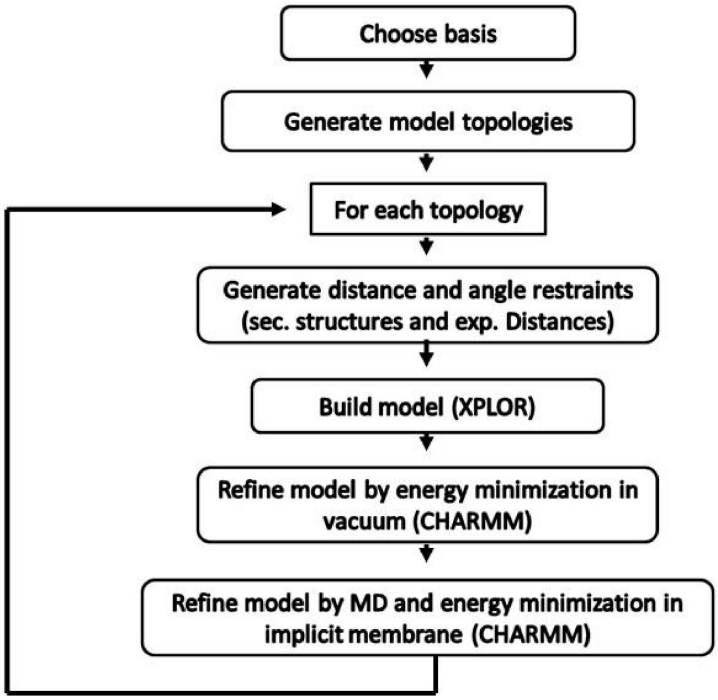

Figure 4.6: Flow chart showing the protocol of model building. A basis is first chosen that is a series of TMHs that will bind the metal and whose topology will be constant in all models. Then model topologies are generated using a homemade C^{++} program: "buildtopo". Then for each of these models, 3D coordinates are built using X-PLOR and refined using MD simulation with CHARMM inside an implicit membrane after the introduction of 2 Cd^{2+} ions.

TMH	Serca (211/304)	CadA (149/217)
1	5:25	4:22
2	37:56	28:46
3	68:88	58:75
4	100:118	77:93
5	130:150	105:125
6	152:173	137:161
7	185:208	173:192
8	220:239	195:214
9	251:270	-
10	278:300	-

Table 4.10: Ranges of residues defining the TMhs for Serca and CadA. These definitions of the TMHs have been used for defining restraints for X-PLOR modelling and refinement and CHARMM MD. 211 amino acids are in αhelix conformation out of 304 in Serca. 149 amino acids are in αhelix conformation out of 217 in CadA.

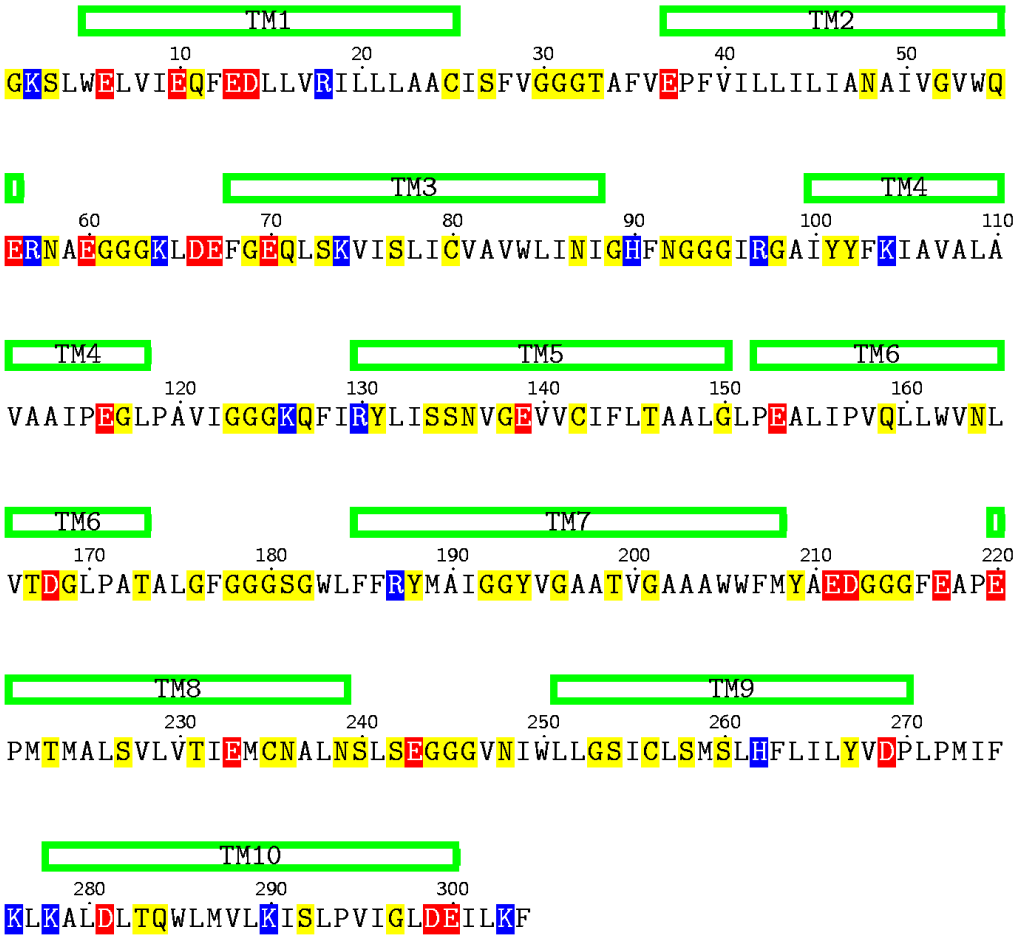

Figure 4.7: Reduced sequence of Serca used in MD simulations. TMHs are highlighted with colour boxes as well as charged or polar amino acids.

Figure 4.8: Reduced sequence of CadA used in MD simulations. TMHs are highlighted with colour boxes as well as charged or polar amino acids.

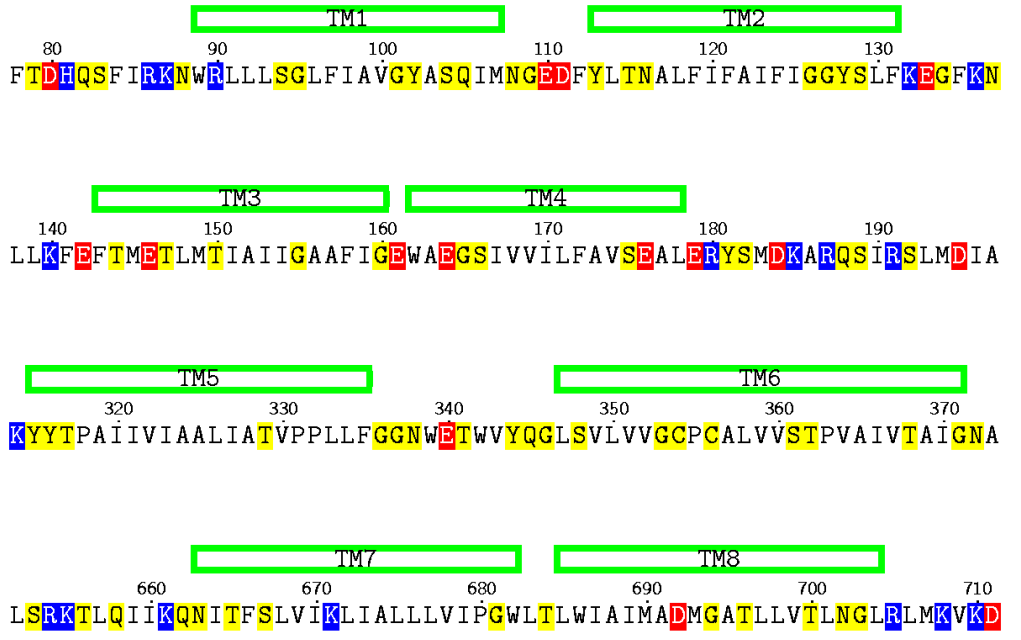

Figure 4.9: Extract of the sequence of CadA from *Listeria monocytogenes* with amino acid numbering of the full-length sequence. TMHs are highlighted with colour boxes as well as charged or polar amino acids.

structures the TMH bundle is oriented along the Z-axis with the CHARMM "coor orient" statement. This statement finds the principal axes of the protein and automatically orients the axis with largest eigenvalue towards Z. NOE-like and dihedral angle restraints are introduced in the force field. Two Cd^{2+} ions are introduced in the models 2 Å away from the 2 CYS sulfur residues of the CPC motif (C144 and C146). The same distance and dihedral angle restraints are applied as in the case of model building with X-PLOR. Then the structures are energy minimized in vacuum with CHARMM. In a second step, the implicit solvent method EEF1 (See section 2.8.2.5) is setup for a neutral (zwitterionic) membrane of 28 Å width. Drift of the protein in the X-Y plane is avoided with the introduction of a "MMFP" restraining potential (basically harmonic) applied on the X and Y components of the center of mass of the TMH bundle with force constant 100 kcal/mol/$Å^2$. Then 50 steps SD followed by ABNR minimizations (See section 2.7.3.2) down to an energy gradient of 0.1 kcal/mol/Å are run. Finally 5 ns MD simulations are run using Langevin dynamics (See section 2.3.2.3) with a time step of 2 fs. To check the influence of the restraints on the results 1 more ns MD simulation is run after removing all NOE-like and dihedral angle restraints.

4.5 Model checking using standard methods

4.5.1 Using Procheck

Quality of the models of CadA was checked using PROCHECK-NMR [166] The results are shown in Table 4.11. Models 11, 16, 21 and 24 seem particularly interesting in terms of good positioning of the residues in the Ramachandran plot. The disallowed residues are listed in Table 4.12. Y105 at the beginning of TMH5 seems problematic to most models. All disallowed residues belong to loop regions in the sequence (See Fig. 4.8).

4.5.2 Using stride

Program stride [167] was used to check the integrity of transmembrane helices after model building with XPLOR and refinement with CHARMM. First, the output of stride in the case of Serca models is summarized in Table 4.13. We can see a good conservation of the secondary structures of the TMHs even though a loss of α-helix conformation is sometimes observed: In the case of the X-ray model and TMH4, this is easily explained by the presence of a proline residue at position 115.

We used the same analysis in the case of CadA. Results of these analyses after CHARMM MD and minimization in the presence of restraints and after removing the restraints are shown in Tables 4.14 and 4.15, respectively. Clearly and not surprisingly, in the presence of restraints, all TMHs are preserved. TMH5 in model 17 is even 4 residue longer than imposed by the restraints. When the restraints are removed, many helices lose their α-helical conformation at the predicted ends of these TMHs, especially in the case of models 13, 16 and 17. Loss of α-helices is most of the time compensated by the appearance of turns, bridges and π-helices. Models 21 and 27 seem exceptionnally stable regarding the maintaining of the α-helical conformation of their TMHs.

Model	After X-PLOR		After CHARMM	
	Core+allowed (%)	disall.	Core+allowed (%)	disall.
7	91.2	11	95.6	8
8	91.6	9	95.6	5
11	95.2	1	98.4	4
12	94.5	5	94.6	12
13	97.3	3	95.7	5
14	94.6	2	95.2	8
15	90.7	9	95.6	8
16	96.7	3	97.9	4
17	98.9	1	97.3	6
19	97.3	2	96.8	6
21	96.7	2	97.3	5
22	95.7	6	97.3	6
23	96.2	5	97.8	4
24	95.7	0	97.3	4
25	94.9	5	96.7	6
27	91.6	10	94	8

Table 4.11: Procheck results for the 16 models of CadA: Percentage of core and allowed residues in the Ramachandran plot are shown in col.2 and 4 after X-PLOR refinement and CHARMM MD and minimization respectively. Corresponding number of disallowed residues are listed in col. 3 and 5.

Model	Residue name
Cada m16	R215, Y105, L137, L193
Cada m21	A164, K47, Y105, N52, W77
Cada m24	K47, L93, W129
Cada m27	E48, E57, E94, D26, Y105, L137, L93, W192
Serca m1	M273, L277
Serca m2	S3, K64
Serca m3	F68, F128
Serca m4	V21, I6, I88, I122, Q127, L241, L251, C25
Serca m5	L302, A174, K276, E153
Serca X-Ray	I301, E67, L184, W183
BR m1	Y131, Y133, D104, S193, A228
BR m2	D102, T170, T67, E161, I229
BR m3	R227, M163, W12, E166, Q75, V167, D102, I4, T128
BR m4	D102, T128, S162, L66, Y79
BR X-ray	-

Table 4.12: List of disallowed residues found by Procheck analysis for selected models of CadA, Serca and bR. Coordinates after CHARMM MD and minimization have been used. In the initial X-ray structure of Serca, we found 8 disallowed residues namely E67, R97, F128, W183, L184, F216, K276 and I301.

Model	TMH1 (5-25)	TMH2 (37-56)	TMH3 (68-88)	TMH4 (100-118)	TMH5 (130-150)
1	6-21	37-56	68-76 + 79-90	101-114	131-149
2	6-21	40-54	69-73 + 79-92	99-114	128-149
3	5-24	38-60	70-88	101-113	129-149
4	4-22	39-54	64-76 + 79-88	100-117	132-148
5	6-24	35-55	67-87	99-117	127-137 + 140-149
X-ray	3-22	35-58	68-90	99-114 + 119-122	127-149
Model	TMH6 (152-173)	TMH7 (185-208)	TMH8 (220-239)	TMH9 (251-270)	TMH10 (278-300)
1	151-172	185-212	221-232	251-268	277-303
2	156-166	186-207	221-238	255-269	278-300
3	156-175	184-208	222-234	251-268	277-299
4	154-165 + 168-172	185-207	221-238	255-268	279-301
5	154-170	185-209	221-240	255-268	279-299
X-ray	155-175	185-208	219-231	249-269	279-299

Table 4.13: Ranges of residues found in α-helix by the program stride for 5 models of Serca + the X-ray structure. Coordinates are taken from the output of CHARMM after MD and minimization after removing the NOE-like restraints. (Residue ranges in parentheses represent the definitions of the TMHs used in the simulations for setting up the restraints. (See Table 4.10)

Model	TMH1 (4:22)	TMH2 (28:46)	TMH3 (58:75)	TMH4 (77:93)	TMH5 (105:125)	TMH6 (137:161)	TMH7 (173:192)	TMH8 (195:214)
11	3-22	27-48	58-74	77-92	105-125	137-162	173-192	196-214
13	4-22	29-46	58-74	78-95	105-124	136-160	173-191	196-213
16	4-23	29-47	58-74	78-93	106-124	138-160	174-192	195-214
17	4-22	27-48	59-74	77-93	106-129	137-160	174-191	196-213
21	5-21	29-46	58-75	78-94	106-125	136-161	173-192	195-214
24	4-23	29-46	59-74	78-92	105-127	136-160	173-192	196-214
27	4-21	29-45	58-75	78-92	106-125	138-163	173-191	196-213

Table 4.14: Ranges of residues found in α-helix by the program stride [167] for 7 selected models of CadA. Coordinates are taken from the output of CHARMM after MD and minimization in the presence of NOE-like restraints.

Model	TMH1 (4:22)	TMH2 (28:46)	TMH3 (58:75)	TMH4 (77:93)	TMH5 (105:125)	TMH6 (137:161)	TMH7 (173:192)	TMH8 (195:214)
11	3-22	27-48	58-74	77-92	105-125	137-162	173-186	196-214
13	4-21	29-45	67-72	78-94	105-124	147-160	173-194	196-213
16	4-23	29-46	58-74	78-92	106-125	137-156	174-182	195-214
17	5-22	27-47	59-73	77-93	106-127	137-160	174-189	203-213
21	5-21	29-46	59-70	78-94	106-125	136-161	174-192	195-214
24	4-23	32-46	59-74	78-91	106-124	136-160	173-189	196-214
27	4-21	29-45	56-75	78-93	106-125	138-162	173-189	199-211

Table 4.15: Ranges of residues found in α-helix by the program stride for 7 selected models of CadA. Coordinates are taken from the output of CHARMM after MD and minimization after removing the NOE-like restraints.

4.6 Model checking using energy minimization and MD simulations

4.6.1 Validation using bR and Serca

In order to check the reliability of our protocol described in Fig. 4.6, to build TMHs of TM proteins of unknown structure, we have applied this protocol to the prototypical membrane protein: bacteriorhodopsin (bR) and the Ca^{2+} ATPase Serca.

4.6.1.1 Bacteriorhodopsin (BR)

BR is a transmembrane protein found in the cellular membrane of *Halobacterium salinarium*, which functions as a light-driven proton pump. The bR molecule contains seven helices that surround a channel through which ions can move. BR is a 247 amino acid protein of which 2 thirds constitute 7 TMHs as shown in Table 4.16. In the absence of large cytoplasmic loops, the total sequence of bR was used in the models and simulations.

TMH	PDB	"findwidth"
1	8-28	12-28
2	37-61	40-59
3	80-101	80-99
4	105-126	106-125
5	133-162	135-156
6	165-191	171-190
7	201-225	205-225

Table 4.16: Definition of TMHs in bR: col.2: from information in the PDB structure; col.3: from the "findwidth" script, and used for defining restraints in the models.

The refined topology of the TMHs in bR, "where do they begin and end" in sequence has been predicted using the "findwidth" CHARMM script described previously. We then followed the protocol for model generation and refinement shown in Fig. 4.6. Initial grid models for bR are shown in Fig. 4.10. A pseudo basis for bR is constituted of TMHs 3,4,5 and 6 (the 4 most buried helices). The models were built and refined using X-PLOR as explained in the case of CadA. Then MD simulations of bR inside an implicit membrane were run using CHARMM starting from the 4 models and also from the X-ray structure. A graph of the CHARMM total potential energy vs. time during the 5 ns MD simulation is shown in Fig. 4.11. Clearly, the energy of the bR X-ray structure is the lowest which validates our approach based on the idea that a low energy model should have a structure close to the native one and vice versa. Model 1, built with X-PLOR with a topology closest to the X-ray structure of bR equilibrates rapidly to a low energy value.

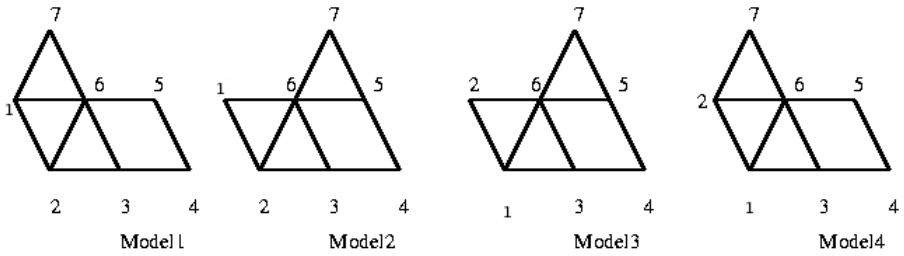

Figure 4.10: Grid models for bacteriorhodopsin. Model corresponding to X-ray structure is model 1

Procheck analysis for bR in Table 4.17 shows a perfect positioning of the residues of the X-ray structure in the Ramachandran plot which is conserved after MD simulation. Model 1 which corresponds to X-ray also performs relatively well according to Procheck.

	After X-PLOR		After CHARMM	
Model	Core+allowed (%)	disall.	Core+allowed (%)	disall.
1	97.4	2	97	5
2	96.5	1	97.5	5
3	96.9	3	95.4	9
4	99.0	0	96.9	5
X-ray	100.0	0	100.0	0

Table 4.17: Procheck results for bR models: Percentage of core and allowed residues in the Ramachandran plot are shown in col. 2 and 4 after X-PLOR refinement and CHARMM MD and minimization respectively. Corresponding number of disallowed residues are listed in col. 3 and 5. The measures on the X-ray structure are reported in the "X-ray" line and "After X-PLOR" columns.

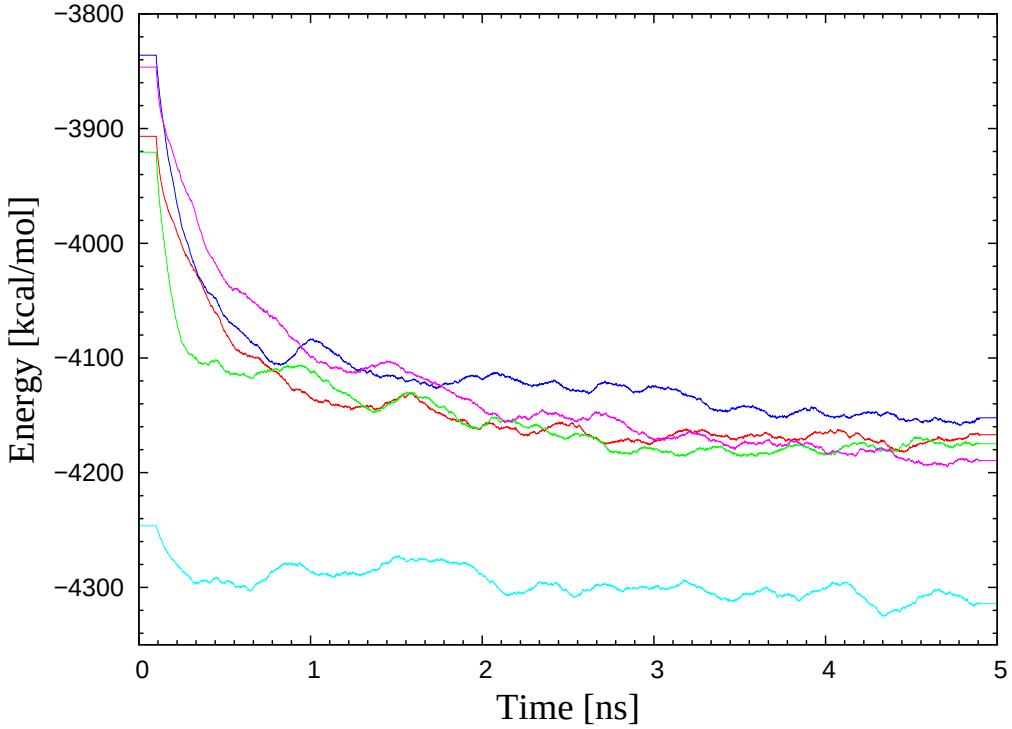

Figure 4.11: Total CHARMM Energy vs Time for Bacteriorhodopsin simulations: Model 1 : Red; Model 2 : Green; Model 3: Blue; Model 4: Pink while the simulation from X-ray structure is shown in cyan

4.6.1.2 Serca

Model building and MD simulations were run on a "reduced" model of the protein. The reduced sequence is shown in Fig. 4.7. Ranges of residues defining the 10 TMHs have been given already in Table 4.5 col.4 and Table 4.10 col.2 for the real and reduced sequences, respectively. The basis for Serca is constituted of helices 4,5,6 and 8 as shown in Fig. 4.12; the model corresponding to the X-ray structure and 4 variations have been tested. The same protocol as for bR was used for model building and refinement and the same analyses of the results were done.

It is surprising to see in Table 4.18 that, in the case of Serca, Procheck analysis reveals 9 aminoacids in disallowed regions of the Ramachandran plot for the X-ray structure. Model 1 with no disallowed residue and more than 99 % allowed looks almost perfect in terms of positioning of the residues in the plot. Model 3, the closest to X-ray, also looks good.

The atom coordinate Root Mean Square Deviation (RMSD) from the initial coordinates has been calculated after MD simulations with CHARMM for the models described in Fig. 4.12. Results are shown in Fig. 4.13. The large RMSDs denote big rearrangements occuring in the structures until a plateau and thus equilibration is reached. No significant increase in RMSD occurs after removing the restraints (last ns).

A graph of CHARMM energy vs. time for the 5 models of Serca + The X-ray structure is shown in Fig. 4.14. The results are summarized in Table 4.19. A big energy drop occurs when the restraints are removed and the systems hardly reach a new equilibrium: the X-ray structure does not give the lowest energy structure after 6 ns. On the other hand, when three models are equilibrated for a longer

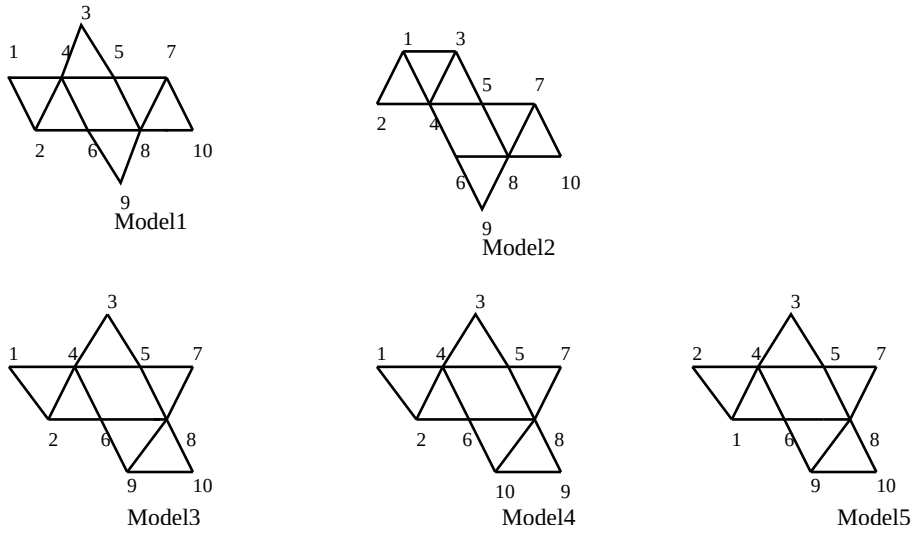

Figure 4.12: 2-D grid models of Serca. Model close to X-ray structure is model 3

	After X-PLOR		After CHARMM	
Model	Core+allowed (%)	disall.	Core+allowed (%)	disall.
1	99.3	0	98.8	2
2	98	2	98.8	2
3	98.8	2	98.4	2
4	98.4	2	96.5	8
5	98.4	1	97.6	4
X-ray	95.2	9	98.4	4

Table 4.18: Procheck results for Serca models: Percentage of core and allowed residues in the Ramachandran plot are shown in col. 2 and 4 after X-PLOR refinement and CHARMM MD and minimization respectively. Corresponding number of disallowed residues are listed in col. 3 and 5. The measures on the X-ray structure are reported in the "X-ray" line and "After X-PLOR" columns.

time of 11 ns, they reach an equilibrtium and the X-ray model yields almost the same energy as the X-PLOR built model which most resembles the X-ray structure: (average energy between 10 and 11 ns: -6039 kcal/mol for model 3 and -6032 kcal/mol for X-ray model). Contrary to the case of bR where we used the full structure, we used, in the case of Serca a "reduced" system where big loops have been deleted. This certainly has a consequence on the energy of the system: reintroducing the loops in the model might very well increase their energy higher than that of the X-ray model.

	With restraints		Without restraints	
Model	After mini.	< Last ns >	After mini.	< Last 200 ps >
1	-8153	-5812	-8340	-5926
2	-8139	-5798	-8258	-5903
3	-8189	-5812	-8388	-6019
4	-8185	-5840	-8309	-5930
5	-8179	-5815	-8307	-5942
X-ray	-8169	-5805	-8310	-5930

Table 4.19: CHARMM energies (kcal/mol) for 5 models of Serca + the X-ray structure: In the presence of distance and dihedral restraints: col.2 and 3 after 5 ns MD and energy minimization and averaged over the last ns, respectively. After removing the restraints: col. 4 and 5 after 1 ns MD, energy minimization and averaged over the last 200 ps, respectively.

4.6.2 Results on CadA

4.6.2.1 Energy

Average CHARMM potential energy over the last ns of 5 ns MD runs and minimum energy after minimization of the 16 models of CadA with radius lower than 1 are shown in Table 4.21. Based upon energy, models 16, 21, 24 and 27 have been chosen for further analysis. Note that models 16, 21 and 24 had already been selected as good models under the Procheck criterions. Models 11, 13 and 17 also display low energy and low number of disallowed residues and could be analyzed further. Model 27 is interesting because it gives the worst results according to Procheck but the lowest energies. Energies of selected models are reported in Table 4.22 together with average potential energy over 200 ps and minimum energy after energy minimization calculated from an 1 ns MD run after removing all restraints.

A graph of total CHARMM potential energy vs. time for the 4 "best energy" models of CadA plus 1 model coming from the topology of CopA is shown in Fig. 4.15.

4.6.2.2 RMSD

The atom coordinate RMSD from the initial coordinates has been calculated during the dynamics for the selected models. Results for backbone atoms are shown in Fig. 4.16. The same remarks as for Serca apply: RMSDs reach a plateau after around 1 ns MD and no big change is observed after

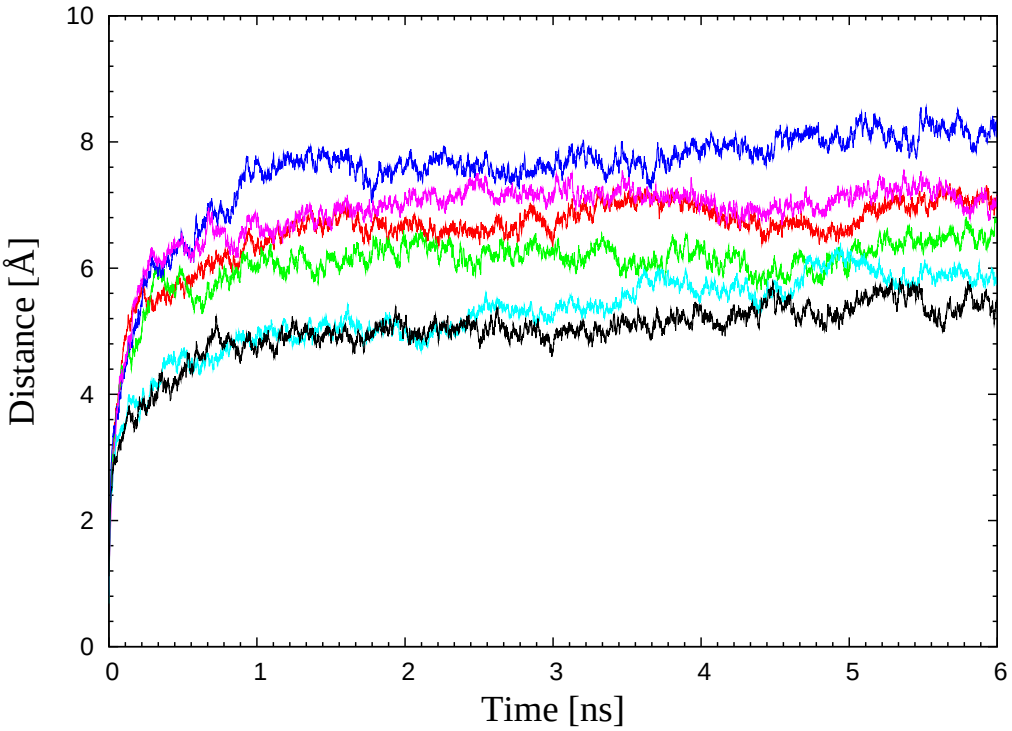

Figure 4.13: Backbone atom RMSD with respect to initial MD coordinates vs Time for Serca. Red : Model1; Green : Model2; Blue : Model3; Pink : Model4; Cyan : Model5; black : X-ray structure. Restraints have been removed after 5 ns MD.

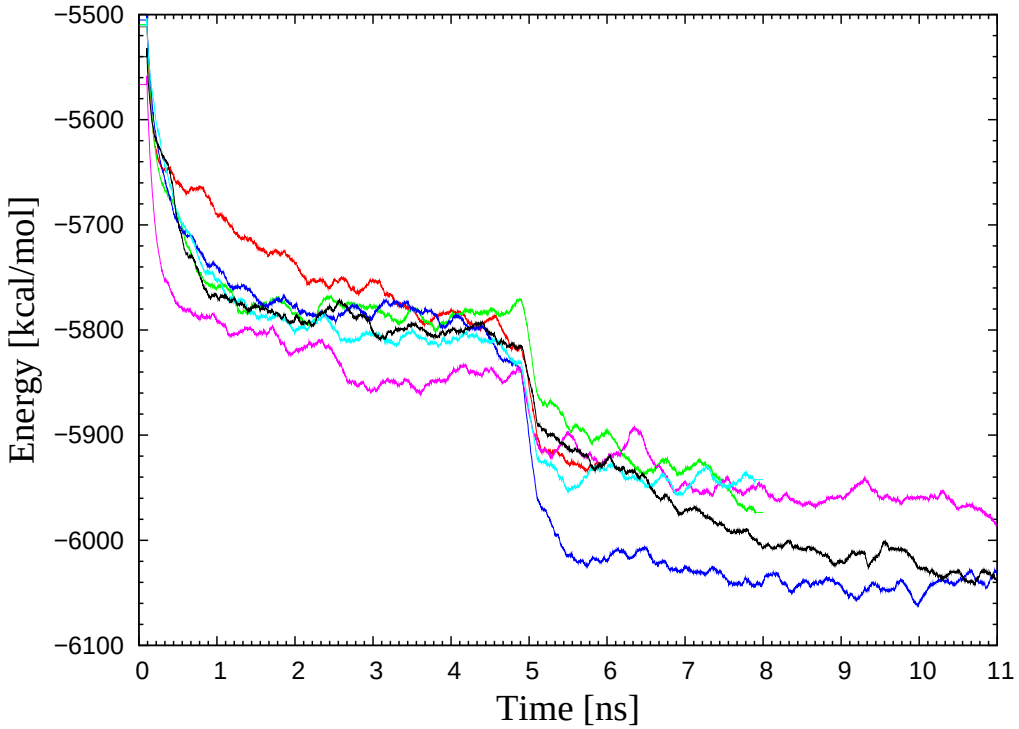

Figure 4.14: Total potential energy vs. time for Serca. Red : Model1; Green : Model2; Blue : Model3; Pink : Model4; Cyan : Model5; black : X-ray model. Model 3, 4 and X-ray have been simulated for 11 ns instead of 6 ns for the others.

TMH	Range	Extension (Å)
1	89:107 (4-22)	-12.6 to 10.1
2	113:131 (28-46)	-14.9 to 11.6
3	143:160 (58-75)	-11.8 to 11.8
4	162:178 (77-93)	-12.5 to 11.7
5	315:335 (105-125)	-17.8 to 14.4
6	345:371 (137-161)	-17.1 to 14.7
7	663:682 (173-192)	-12.7 to 12.2
8	685:704 (195-214)	-10.7 to 13.0

Table 4.20: Extension of CopA-CadA TMHs (script *findwidth*). The sequence range of each helix is given in col.2 following real sequence numbering and, in parentheses, numbering relative to the reduced sequence used for MD simulations and protocol checking. After orientation of the 10-helix bundle normal to the X-Y membrane plane, the Z-coordinates of Cα atoms of the two end-residues of each TMH are listed in col. 3. Taking the averages of these coordinates, we find that the width of the transmembrane part of CopA-CadA defined by its Cα atoms is $12.4 - (-13.7) = 26.1$ Å .

Model	Minimized	<Last ns MD>
7	-5939	-4278
8	-5947	-4260
11	-6009	-4300
12	-5960	-4263
13	-6003	-4294
14	-5983	-4285
15	-5917	-4235
16	-6011	-4323
17	-6012	-4303
19	-5965	-4281
21	-6041	-4343
22	-5960	-4281
23	-5928	-4266
24	-6046	-4350
25	-5982	-4290
27	-6097	-4405
CopA	-5911	-4218

Table 4.21: Total CHARMM energy (kcal/mol) of 16 CadA models + the model from CopA. Col 2. Energy minimized model after MD. Col. 3 average over the last 1 ns of MD.

Model	With restraints		Without restraints	
	After mini.	< Last ns >	After mini.	< Last 200 ps >
16	-6011	-4323	-6150	-4432
21	-6041	-4343	-6195	-4475
24	-6046	-4350	-6230	-4508
27	-6097	-4405	-6241	-4523
CopA	-5911	-4218	-6142	-4444

Table 4.22: CHARMM energy (kcal/mol) for the 4 "best" energy models of CadA + the model built from CopA: In the presence of distance and dihedral restraints: col.2 and 3 after energy minimization and averaged over the last ns, respectively. After removing the restraints: col. 4 and 5 after energy minimization and averaged over the last 200 ps, respectively.

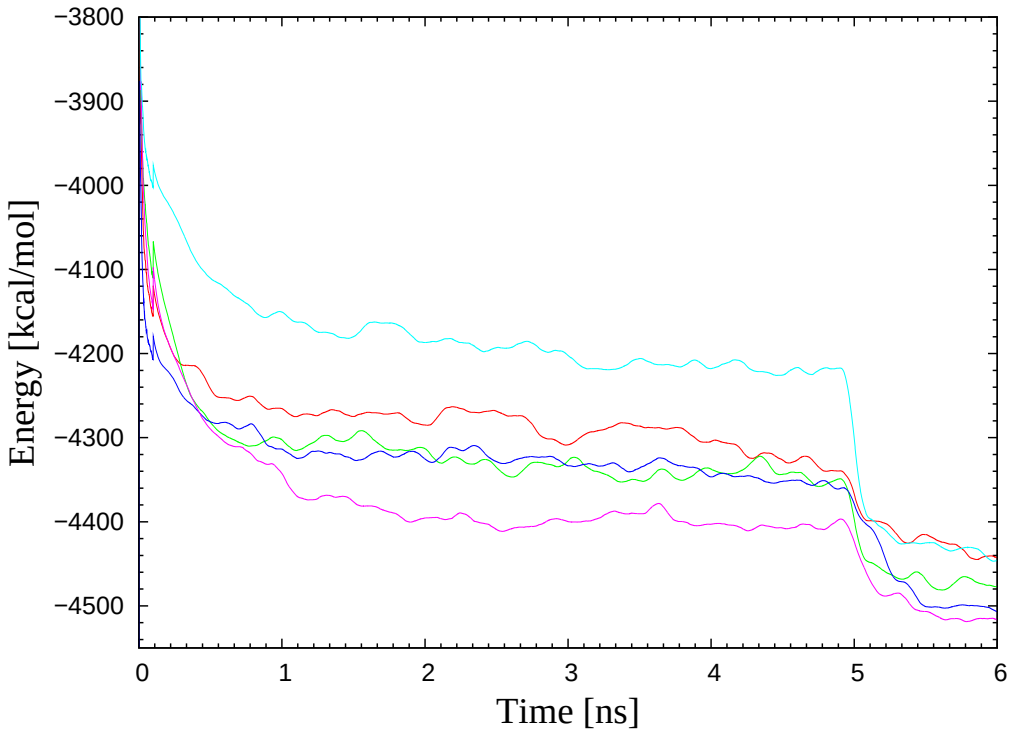

Figure 4.15: Total potential energy vs Time for CadA models. Model 16: Red, Model 21: Green, Model 24: Blue, Model 27: Pink, CopACadA: Cyan. Restraints have been removed after 5 ns MD.

removing the restraints (last ns). Note, however, that the RMSDs are, on average, higher than in the case of Serca. Model 27 with lowest energy undergoes the largest rearrangements of its structure together with the model from CopA with high energy which could need further refinement.

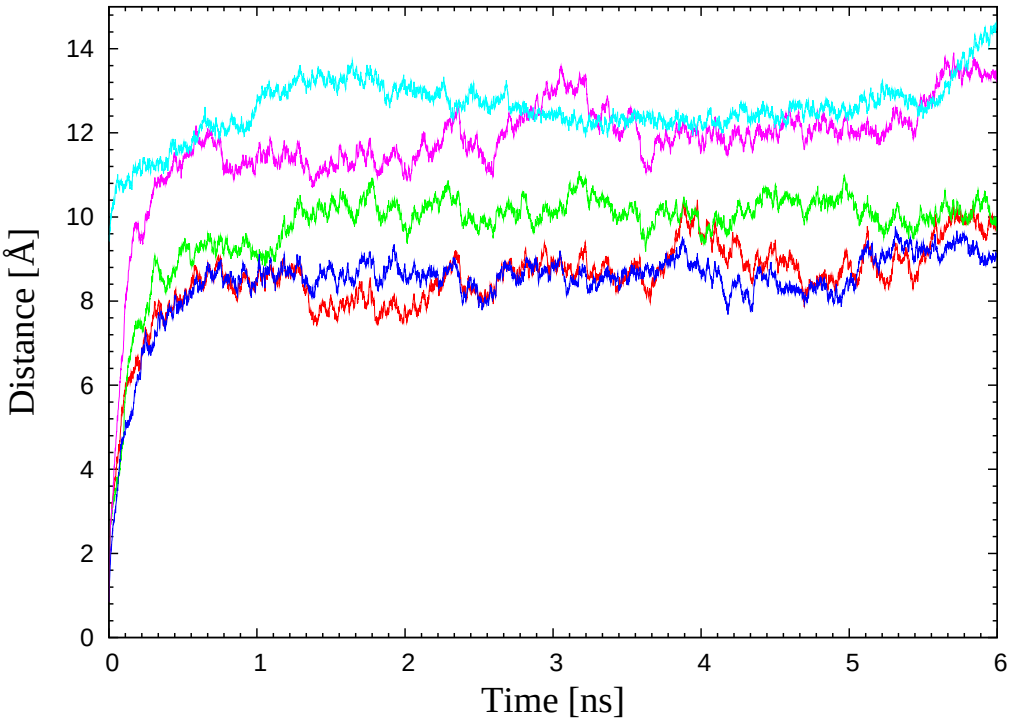

Figure 4.16: Backbone RMSD with respect to initial MD coordinates vs time for CadA models. Model 16: Red, Model 21: Green, Model 24: Blue, Model 27: Pink, CopACadA: Cyan. Restraints have been removed after 5 ns MD.

4.6.2.3 TMH tilt angles

Tilt angles of the TMHs of CadA with respect to the normal vector to the membrane have been calculated from minimized energy structures after 5 ns MD in the presence of restraints. Tilt angles vary from 8 to 30 degrees depending on the TMH. No clear difference between the models appears

4.6.2.4 Cadmium sites

To get an idea of the residues most probably involved in the binding of Cd^{2+}, the CHARMM interaction energies between Cadmium ions and their surrounding residues have been calculated. Results are summarised in Table 4.23. Removing the restraints dramatically lowers the interactions. This is due to electrostatics terms which, in the case of a charged ion, will vary sharply with distance: in the implicit membrane simulation approach with CHARMM an r-dependant dielectric constant is used ($\epsilon = r$) which makes the electrostatic energy vary like $1/r^2$. If we take the example of the Cd^{2+} ion interacting with a CYS Sulfur atom (charge -1 in the *toph19-eef1* CHARMM topology file), the electrostatic energy of the Cd-S pair will vary from -150 to -74 kcal/mol when the distance varies from 2.1 to 3 Å.

	Cd1		Cd2	
Model	NOE	No NOE	NOE	No NOE
16	-258	-419	-360	-411
21	-221	-377	-254	-346
24	-365	-494	-438	-523
27	-434	-532	-411	-505
CopA	-381	-393	-199	-279

Table 4.23: Total CHARMM interaction energies between cadmium ions (noted Cd1 and Cd2) and all other residues around them from the minimized energy structures after MD runs. Both the structures after 5 ns in the presence of restraints and after 6 ns, after releasing the restraints for 1 ns, have been used.

In models 24 and 27, Cd^{2+} ions seem to be more stably inserted in their sites. Individual interactions with the neighbouring residues are listed in Tables 4.24 and 4.25, for Cd1 and Cd2, respectively. Cd1 shows high interactions with C144 in all models and with C146 in models 24, 27 and CopA, only. Cd2 shows high interactions with C144 in models 16, 24 and 27 and with C146 in all models except CopA. A small interaction energy between M64 in TMH3 (eqv. M149 in the real sequence) and Cd1 in model 24 or Cd2 in model 16 is visible. Medium size interaction energies between both cadmium ions and D202 (eqv. D692) are detected. Cd1 in model 27 and Cd2 in model 24 interact significantly with carbonyl oxygens of L140, V141 and V142. Interactions between Cd2 and G143 are also clear in models 16 and 21.

Model	Residue (Interacting Energy)
16	C144(-157), C146(-23), L178(-4), K181(-45), D202(-30)
21	C144(-158),L195(-5),A198(-37),I199(-7),D202(-12)
24	M60(-4),M64(-5),L140(-7),V141(-7),V142(-7),C144(-151),C146(-157),D202(-27)
27	V139(-5),L140(-45),V141(-41),V142(-43),C144(-150),C146(-144)
CopA	S9(-6),F12(-38),13I(-6),G143(-36),C144(-162),C146(-151)

Table 4.24: CHARMM Interaction energies (kcal/mol) between Cd1 and other residues from the minimized energy structures of CadA models after MD in the presence of restraints.

4.6.2.5 Role of K181 (eqv. K671)

Lysine 671 (K181) in TMH7 is a conserved residue in Zn^{2+} and Cd^{2+} ATPases whereas in Cu^{+} ATPases, it is replaced by an asparagine (See position 48 in the sequence shown in Fig. 4.2). Largest interaction energies ($>$ 4 kcal/mol) between the important residue K671 in TMH7 and other residues of the different CadA models are listed in Table 4.26. K181 shows sizeable interactions with the carbonyl oxygens of its neighbours I180 and L182. A small interaction with D202 appears in model 16.

Model	Residue (Interacting Energy)
16	M60(-5), M64(-6), G143(-39), C144(-157), C146(-151), D202(-5)
21	V142(-42), G143(-43), C144(-21), C146(-146)
24	V139(-6), L140(-46), V141(-42), V142(-44), C144(-147), C146(-144)
27	L140(-7),V141(-7),V142(-7),C144(-149),C146(-156),D202(-53),T206(-28)
CopA	E91(-11),C144(-10),C146(-13),A147(-42),L148(-40),S151(-20),T152(-32),D202(-28)

Table 4.25: CHARMM Interaction energies (kcal/mol) between Cd2 and other residues from the minimized energy structures of CadA models after MD in the presence of restraints.

Model	Residue (Interacting Energy)
16	S177(-5), L178(-5), I180(-16), L182(-15), D202(-6)
21	T152(-4), S177(-5), L178(-5), I180(-16), L182(-14), L185(-6)
24	S177(-5), L178(-5), I180(-14), L182(-14), A184(-4), L185(-7)
27	S177(-5), L178(-5), I180(-14), L182(-14), A184(-4), L185(-6)
CopA	S177(-8), L178(-5), I180(-16), L182(-14), L185(-6)

Table 4.26: CHARMM Interaction energies (kcal/mol) between K671 in TMH7 (K181) and other residues from the minimized energy structures of CadA models after MD in the presence of restraints.

4.6.2.6 Role of D202 (eqv. D692)

In the same manner, Aspartate 692 (D202) in TMH8 is a well conserved residue in Zn^{2+} and Cd^{2+} ATPases whereas in Cu^{+} ATPases, it is replaced by a Methionine (See position 73 in the sequence shown in Table 4.2). In their work on D714 in ZntA (eqv. of D692 in CadA), Dutta *et al.* showed that four different substitutions at D714 resulted in complete loss of *in vivo* resistance activity and complete or large reductions in ATPase activity [128]. Largest interaction energies ($>$ 4 kcal/mol) between this important residue D692 and other residues of the different CadA models are listed in Table 4.27. In the same manner as K181, D202 shows moderate interactions with its neighbouring residues A201 and M203 and, of course, with K181, in model 16.

Model	Residue (Interacting Energy)
16	K181(-6), A198 (-6), I199(-5), A201(-13), D202(-8), M203(-14)
21	C144(-5), A198 (-4), I199(-8), A201(-13), M203(-16), T206(-8)
24	A198 (-7), I199(-6), A201(-13), D202(-6), M203(-15), T206(-8)
27	A198 (-6), I199(-8), A201(-14), M203(-16)
CopA	E91(-7), A198 (-8), I199(-6), A201(-13), M203(-14)

Table 4.27: CHARMM Interaction energies (kcal/mol) between D692 in TMH8 (D202) and other residues from the minimized energy structures of CadA models after MD in the presence of restraints.

4.7 3D-models of CadA

Close views of the Cadmium sites in models 16, 21, 24, 27 and CopA described and studied above are shown in Figs. 4.17 to 4.21. Distance between the 2 Cd^{2+} ions is 4.1, 8.9, 3.3, 3.3 and 8.7 Å in models 16, 21, 24, 27 and CopA, respectively. These distances are either short or very long compared to the distance between the 2 Ca^{2+} ions in the calcium ATPase Serca. In models 16, 24 and 27 the 2 cysteines of the CPC motif (C144 and C146) are in close proximity with their accompanying Cd^{2+} ion nearby. In model 21, the orientations of the 2 cysteine sidechains are different and thus the two ions are further apart. In the CopA model, Cd2 does not interact directly with any of the 2 cysteines which makes it unlikely to represent a valid model. The interaction between D202 and Cd2 is clear in model 27. In model 16, the side chain of K181 is pointing towards the interior of the TMH bundle and interacts with D202 for a good charge screening.

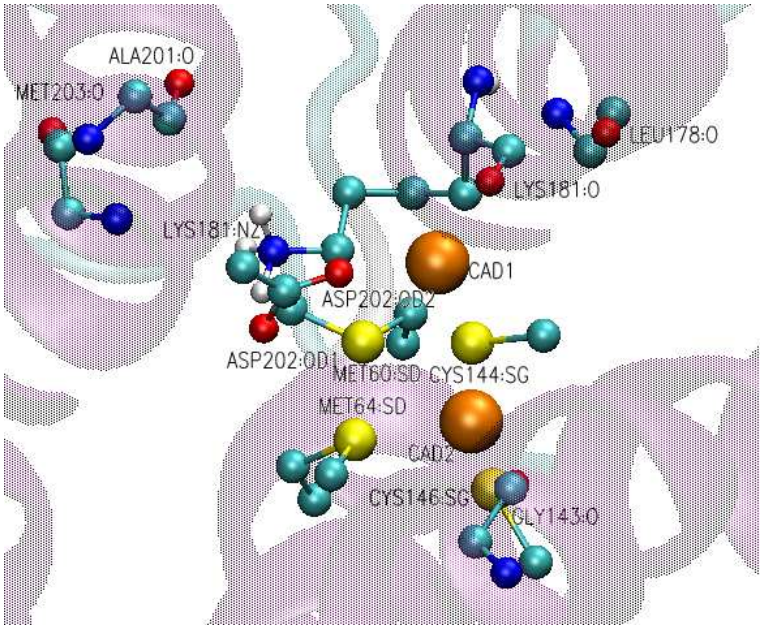

Figure 4.17: Close view of the cadmium sites from model 16. Structures, built and refined with CHARMM inside an implicit membrane, are shown after 5 ns MD in the presence of distance restraints and minimization. Cadmium ions are shown as orange spheres and important residues around the ions are represented in CPK models and labelled. A "glass" view of the TMHs to which they belong is also visible.

4.8 Discussion and perspectives

I have presented a novel method to predict the topology and structure of the transmembrane part of a transmembrane protein supposed to be completely α-helical. To cope with the difficulty of studying all possible combinations of the TMHs, I have assumed that 4 helices (the basis) were known from experimental data to be in close proximity in the center of the bundle. A direct application to the case of ATPases corresponds to the four helices suspected to be involved in metal binding and transport. The choice of these basis helices is still controversial. In their study of CopA from

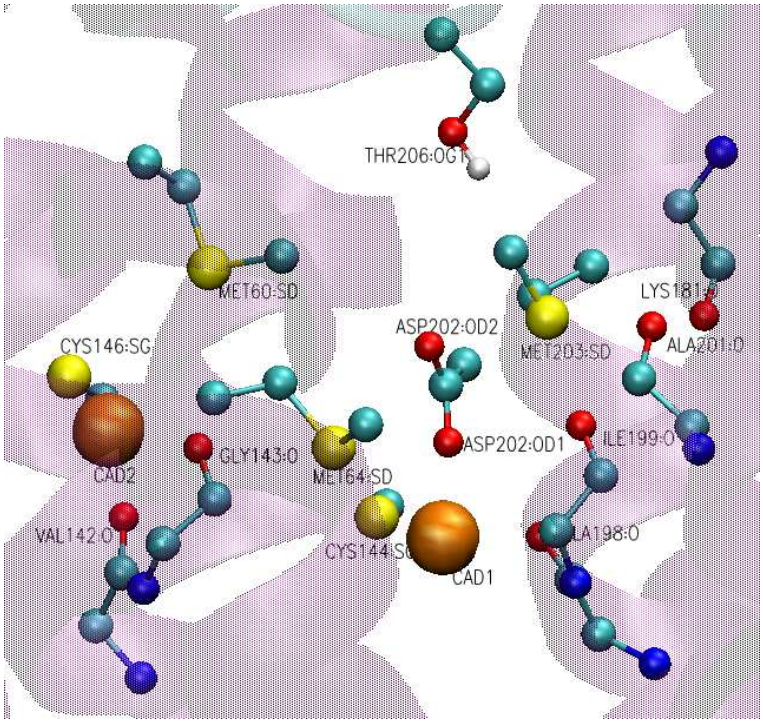

Figure 4.18: Close view of the cadmium sites from model 21. Same legend as in Fig. 4.17.

Enterococcus hirae [165], Lubben *et al.* used homology modeling based on the assumed structural similarity of CopA to the Ca-ATPase, Serca in its E2 form in the absence of calcium (PDB entry 1IWO). To this end, a binary comparison of both sequences was performed with CLUSTALW [86]. N-terminal transmembrane helices TMH1 and TMH2 of CopA are absent in Serca, thus TMH3-TMH8 of CopA were considered homologous to TMH1-TMH6 of Serca. This choice is not evident from the sequence alignments of the proteins but we used their basis helices in most of our models of the cadmium ATPase: CadA. We also built a model of CadA based on the structure of CopA solved (for its extramembrane part only) by another group (Wu *et al.*). These authors published their structure of CopA from *Archaeoglobus Fulgidus* under accession code 2VOY [163] and took the coordinates of the TM part of their protein (chains B,C,D,E,G,H,K,L) directly from the X-ray structure of Serca in its E2 form with bound thapsigargin (PDB code 2EAR). Their model leads to a different basis for metal binding but it is based on E2(TG)P21 crystals of Serca used to build structure 2EAR which diffracted to only 3.3 Å resolution [168]. The two models of bases satisfy a functional homology between Serca and P1B-type ATPases, which is the presence of so-called small and big loops in the proteins corresponding to the A, N and P domains. Small and big loops of Serca are located between TMH 2 and 3 and 4 and 5, respectively. In CadA, these loops are located between TMH 4 and 5 and 6 and 7, respectively, leading to a direct homology of TMHs 2 to 5 of Serca with TMHs 4 to 7 of CadA.

Starting from these bases, we produced and analyzed different models of CadA. We used several criteria to discriminate between these models: lowest energy, low RMSds or maximum number of allowed residues in the Ramachandran plot. These criteria seemed relatively efficient to discriminate between models when applied to bacteriorhodopsin or Serca, although, in the latter case, . removing

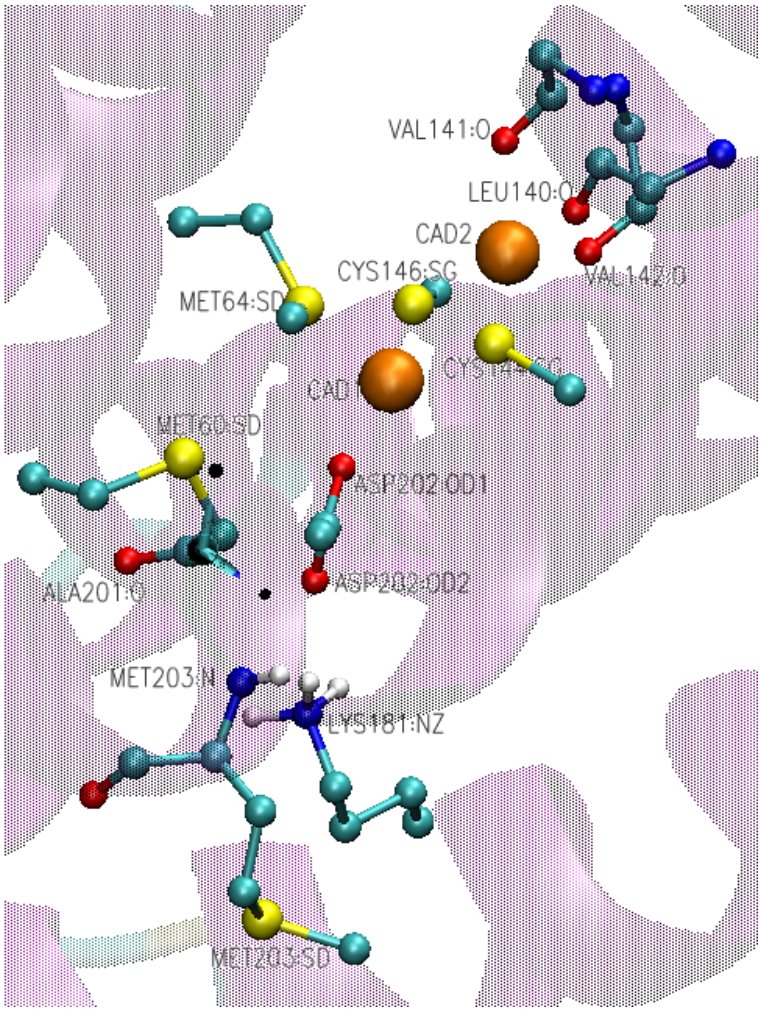

Figure 4.19: Close view of the cadmium sites from model 24. Same legend as in Fig. 4.17.

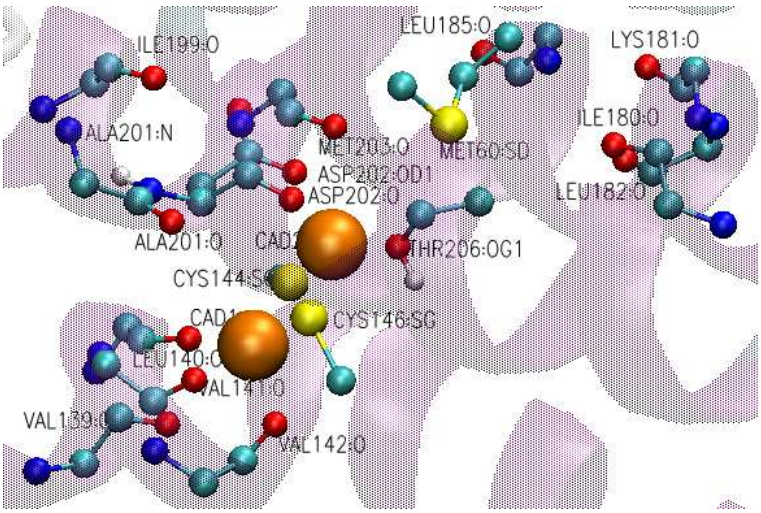

Figure 4.20: Close view of the cadmium sites from model 27. Same legend as in Fig. 4.17.

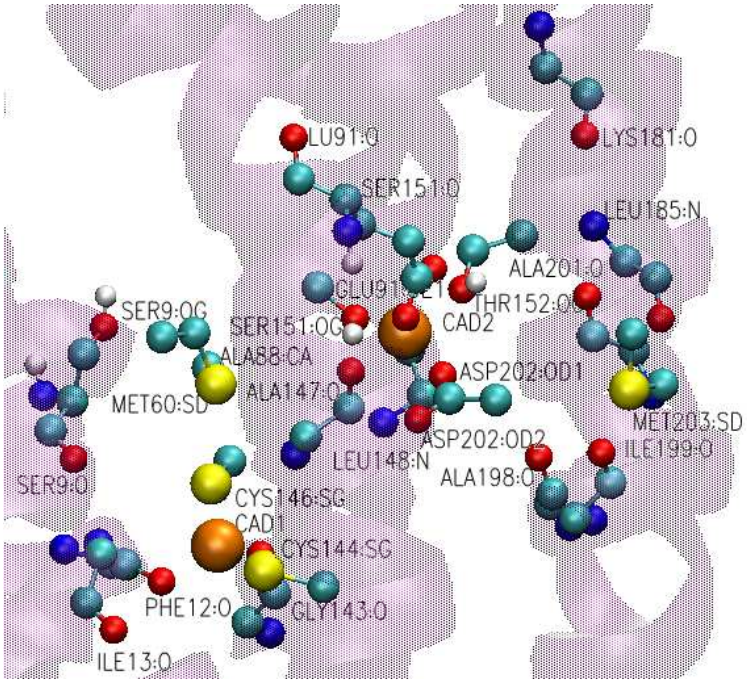

Figure 4.21: Close view of the cadmium sites from model CopA. Same legend as in Fig. 4.17.

large parts of the protein (small and big loop) may have an influence on results.

I think, however that the protocol described for model building and the produced models of CadA, themselves, will be of interest to the experimentalists for getting a clearer view of new possible mutants that could affect metal binding and transport.

The experimental studies carried out so far on the transport site of Cadmium ATPase reveal that two Cd^{2+} ions (hence our choice of two Cd^{2+} ions in our models) are involved in the reaction cycle of CadA. The two cysteines of the Cys-Pro-Cys motif (TMH6) act at distinct steps of the transport process, C354 (C144) being directly involved in Cd^{2+} binding, whereas C356 (C146) would be required for Cd^{2+} occlusion; D692 (D202) in TMH8 would be directly involved in Cd^{2+} binding and its interaction with cadmium is visible in several of our models. E164 in TMH4 would be required for Cd^{2+} release and is far from the cadmium sites in our models. Moreover, in their work on the ATPase ZntA which transports Zn^{2+}, Pb^{2+} and Cd^{2+}, Okkeri and Haltia [169] show experimentally that K693 and D714 (eqv. to K671 and D692 in CadA) are functionally essential and appear to contribute to the metal specificity of the protein, most probably by being parts of the metal binding site made up by the CPC motif. Our model 16 is compatible with this idea.

Our models and simulations indicate, among other results, that mutating K671 into Alanine in CadA seems an interesting idea to see the influence of this residue on Cd^{2+} binding and transport.

More generally, this project will go on by regular exchanges with experimentalists who will put up bonds for realistic models.

Chapter 5

Dynamics and stability of the Metal Binding Domains of the Menkes ATPase

5.1 Introduction

As discussed in a previous chapter, there are, in humans, two genes (ATP7A and ATP7B) coding for Cu^{+}-ATPases: Menkes disease protein (MNKP) and Wilson disease protein (WNDP), that are associated with genetic copper transport disorders. Each of these P1B ATPases (subgroup 1B-1), of unknown three-dimensional structure, possesses 8 TMHs, (like the Cadmium ATPase CadA) and 6 cytoplasmic Metal Binding Domains (MBD) (See section 3.4.6). The absence of these MBDs in some P1B-ATPases suggests that these are likely regulatory domains. Studies have shown that Menkes and Wilson Disease proteins require at least one intact MBD for targeting to the plasma membrane and a vesicular compartment, respectively. Each of the MBDs binds a single copper ion in the reduced Cu(I) form via two cysteine residues. The role of each MBD has been a subject of intense investigations but remains a point of great interest with no clear answer. I have chosen to study the relations between structure and function of the Menkes ATPase (ATP7A) MBDs, for which NMR structures have been obtained. I will present the results I have obtained from molecular dynamics (MD) simulations of the 6 MBDs of the Menkes protein, individually, in their Apo (no copper) and Holo (with copper) forms.

5.2 Models and simulation details

5.2.1 Simulation parameters

All MD simulations were performed with the CHARMM [48] program using the CHARMM27 all-atom force field [170]. To cope with the lack of parameters for metals in the CHARMM force field, researchers in our team have developed, in a previous work [59], special parameters for $Cu^{(I)}$ bound to cysteine sulfur atoms in proteins, in a fixed linear geometry (See Table 2.1). In our model, charge transfer between thiolate and metal is approximated by a covalent bond between metal and sulfur. The electrostatic field around the metal accurately reproduces *ab initio* calculations by a careful adjustment of partial atomic charges of thiolates and metals (See Table 2.2).

5.2.2 Model structures

Tridimensionnal atom coordinates of the Menkes Metal Binding Domains (MBD) were obtained from the NMR structures from the Protein Data Bank with PDB entries given in Table 5.1. Since the

	Apo			Holo	
Mnk	PDB ID	RMSD Å	Charge	PDB ID	RMSD Å
1	1KVI [171]	1.8	-5	1KVJ	2.0
2	1S6O [172]	1.3	+2	1S6U	1.2
3	2G9O [173]	1.03	-1	2GA7	1.2
4	1AW0 [174]	1.00	-5	2AW0	1.0
5	1Y3K [175]	1.7	0	1Y3J	2.4
6	1YJU [176]	1.6	0	1YJV	1.6

Table 5.1: List of Menkes Metal Binding Domains used in the simulations. PDB codes of the NMR structures are given in col.2 and 5 for the Apo and Holo forms, respectively. RMSDs of the Cα atom coordinates of the 10 models found in each PDB entry are given in col.3 and 5. Total charge of each MBD in its Apo form is given in col.4.

MBDs of the Menkes ATPase are part of a longer protein, N- and C-terminal residues have been capped with neutral acetylated N-terminus and amidated C-terminus (ACE and CT2 as defined in the CHARMM topology file). A view of the sixth MBD (Mnk6) in the presence of copper is shown in Fig. 5.1. The elements of secondary structure, the ferredoxin fold: $\beta\alpha\beta\beta\alpha\beta$ are highlighted.

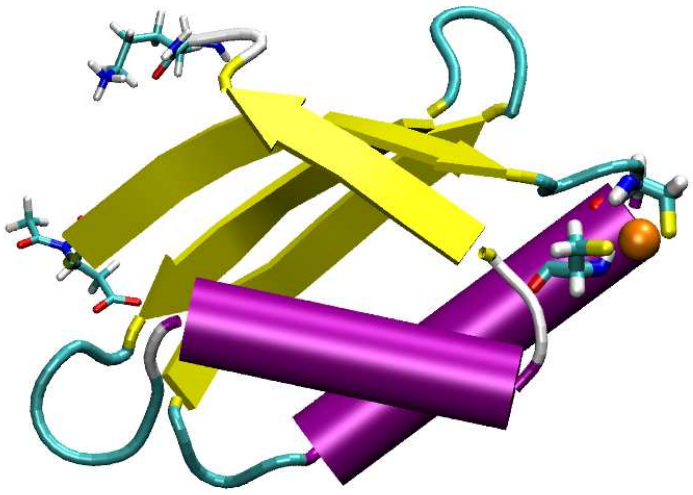

Figure 5.1: Structure of Mnk6-Holoi with Cu^{+} ion represented as an orange sphere. Nter and Cter capping residues, the $\beta\alpha\beta\beta\alpha\beta$ secondary structure motif and the two CYS residues binding the metal in a linear geometry are highlighted.

The first model of each of the solution structure PDB file of the proteins was used as a starting conformation for all MD simulations. (Models are structurally very close and this choice is likely

to have no incidence on the results reported here). Counterions (chloride or sodium) were added to the systems when necessary to keep them electrically neutral. (They were initially placed so as to minimize their electrostatic interaction with the protein). A view of Mnk6 in the presence of copper and with 1 sodium counterion is shown in Fig. 5.2. Charged amino acids are highlighted.

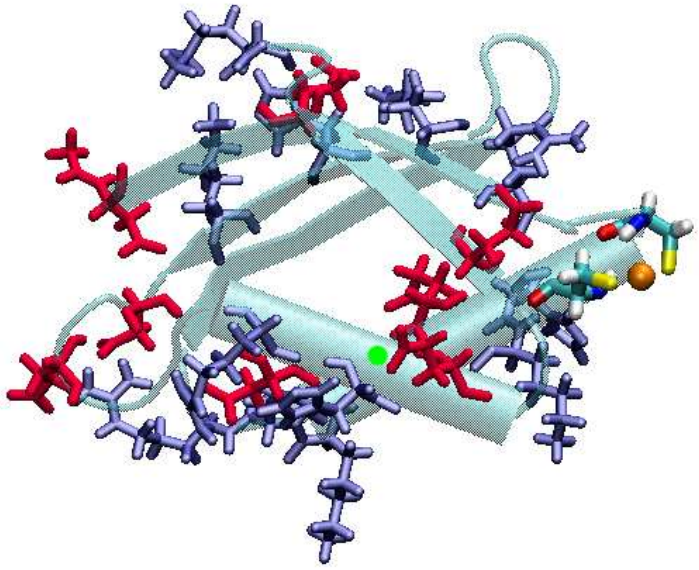

Figure 5.2: Structure of Mnk6-Holo with 1 sodium counterion (green sphere), The Cu^{+} ion is represented as an orange sphere and charged amino acid residues are highlighted in blue (positive LYS and ARG) and red (negative GLU and ASP).

Proteins were solvated with TIP3P water molecules [177] in an orthorhombic box. Water molecules with oxygen atom closer than 2.4 Å to any protein heavy atom (*i.e.* non-hydrogen atom) were deleted. The final simulated system including Cu^{+}, sodium counterion and a water box is shown in Fig. 5.3, still in the case of Mnk6.

5.2.3 Simulation details

The systems were energy-minimized and equilibrated for 30 ps at constant temperature (300 K) and pressure (1 atm) with periodic boundary conditions, and using the particle mesh Ewald method [178] for calculation of long-range electrostatic interactions. Langevin dynamics (See section 2.3.2.3) and Langevin piston were used to keep temperature and pressure constant, with a friction coefficient of $3\,ps^{-1}$ for all protein non-hydrogen atoms. The protein heavy atoms were first constrained with a harmonic force constant of $1\,kcal.mol^{-1}\,Å^{-2}$ for 5 ps. Then, it was decreased to zero in $0.2\,kcal.mol^{-1}\,Å^{-2}$ step every 5 ps. After the 30 ps equilibration, the temperature of the system was kept at 300 K with a Nosé-Hoover thermostat and the pressure was maintained around 1 atm with a Langevin piston. SHAKE [179] was used to constrain all covalent bonds involving hydrogen atoms at their equilibrium length. All the simulations employed the leapfrog Verlet algorithm (See section 2.3.2.2) with a 2-fs integration timestep. The largest box dimension (along the z axis) of each system was allowed to vary with time ($NPAT$ ensemble). All simulations consisted of 30 ps. equilibration followed by 15 ns.

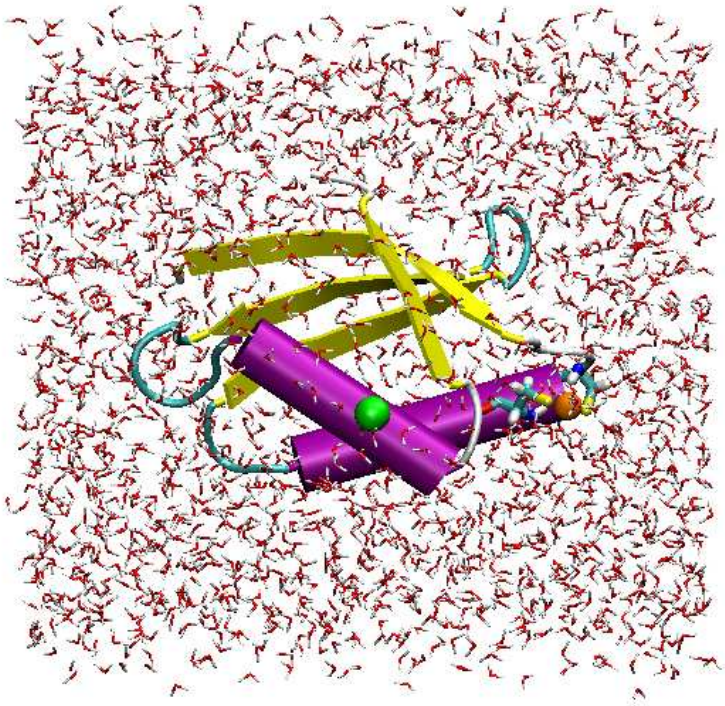

Figure 5.3: Image of the solvated system: Mnk6-Holo with sodium counterion inside a TIP3P water box.

production dynamics. The simulations were analyzed over the last 5 ns (unless otherwise specified). The criteria used to determine the convergence of the simulations were the total and potential energy terms, the crystal dimensions and the time-evolution of the Root Mean Square Deviation (RMSD) of the backbone atom position with respect to the corresponding experimentally resolved starting conformation. Simulations of MBDi of the Menkes ATPase will be noted Mnk$i$ in the following.

5.2.4 Restraints

Simulations were done in the presence of 4 mild distance restraints aimed at preventing the "opening" (fraying) of the structures between β strands. These restraints of the type "biharmonic with well" (See section 2.1.3) are treated as "NOE-like" restraints in CHARMM and have the following form:

$$E_{NOE} = \sum_{restraints,r} k_r \Delta^2$$

with

$$\Delta = \begin{cases} R - (d + d_{plus}) & \text{if } d + d_{plus} < R \\ 0 & \text{if } d - d_{minus} < R < d + d_{plus} \\ d - d_{minus} - R & \text{if } R < d - d_{minus} \end{cases}$$

where the distance between the two atoms is R, the target distance is d and the restraining potential is null for $d-d_{minus} < R < d+d_{plus}$ Values of $k_r = 20$ kcal/mol/Å ^{2}, $d = 2.0$ Å and $d_{minus} = d_{plus} = 0.2$ Å were used. These restraints correspond to H-bonds between atom O of 1 residue in β-1 and H of 1 residue in β-4 and between atom O of 1 residue in β-2 and H of 1 residue in β-3.

5.2.5 Additional sessions and restraints

As already explained, all Mnk MBD systems were first simulated during 15 ns with 4 mild NOE-type restraints aimed at preventing fraying of the β-strands: (let's call these runs session 1). After initial analysis of these results, we noticed large coordinate RMSD with respect to initial NMR structures for several systems. We then decided to run additional MD sessions where we would progressively add distance restraints until we have maintained all H-bonds found in the NMR PDB structure. These new sessions are explained in Table 5.2 for Apo and Table 5.3 for Holo Mnk simulations. In the case of Mnk1, session 2, for instance, 2 restraints were added between C17-O and I21-H within α-1 and between H33-O in β-2 and I46-H in β-3.

Session	Mnk1	Mnk2	Mnk3	Mnk4	Mnk5	Mnk6
1	4 (15)	4 (15)	4 (15)	4 (15)	4 (15)	4 (15)
2	6 (15)	-	-	7 (5)	-	-
3, 4	44 (5)	42 (5)	38 (5)	36 (5)	33 (5)	35 (5)

Table 5.2: Number of NOEs applied in each session and for all Apo Mnk systems. 4 dynamics sessions were run to correct for large RMSD found in some simulations. Session 1 corresponds to the "normal" 15 ns run with 4 NOE-like distance restraints as discussed previously. In Session 3 and 4, all H-bonds detected in the initial NMR structures were translated into distance restraints. In session 3, NOEs were kept for the total 5 ns runs while in session 4, they were removed during the last 2 ns. (Simulation time in ns is given in parentheses.)

Session	Mnk3	Mnk5
2	9 (15)	8 (5)
3	11 (6)	9 (5)
4	15 (5)	17 ($>$ 2.5)
5	17 ($>$ 2.5)	22 (2)
6	49 (9)	44 (5)

Table 5.3: Number of NOEs applied in each session and for Holo Mnk 3 and 5. Six MD sessions were run to correct for large RMSD found in some simulations. Session 1 corresponds to the "normal" 15 ns run with 4 NOE-like distance restraints. Simulation time in ns is given in parentheses. "$>$ 2.5' means that we stopped the simulation after 2.5 ns because a large RMSD between final and initial coordinates was reached. The only difference between sessions is the number of applied restraints. The starting structure remains the original NMR structure after 30 ps. equilibration MD. (Simulation time in ns is given in parentheses.)

5.3 Results

5.3.1 Sequence alignment

A sequence alignment of Menkes Metal Binding Domains Mnk1 to Mnk6 used in this work is shown in Fig. 5.4. The alignment was done with the program CLUSTALW [86]. Conserved sequences include the GMxCxSC (where x represents any residue) metal binding motif in loop L1, I21 and E22 in helix α1, V36 and L38 in β2, Y47 in β3, and I61 in α2. Clearly Mnk3 presents the sequence which shows largest variability compared to others. Mnk3 lacks T13, a positively charge amino acid in 27, G30, F49, G65, F66 and A68.

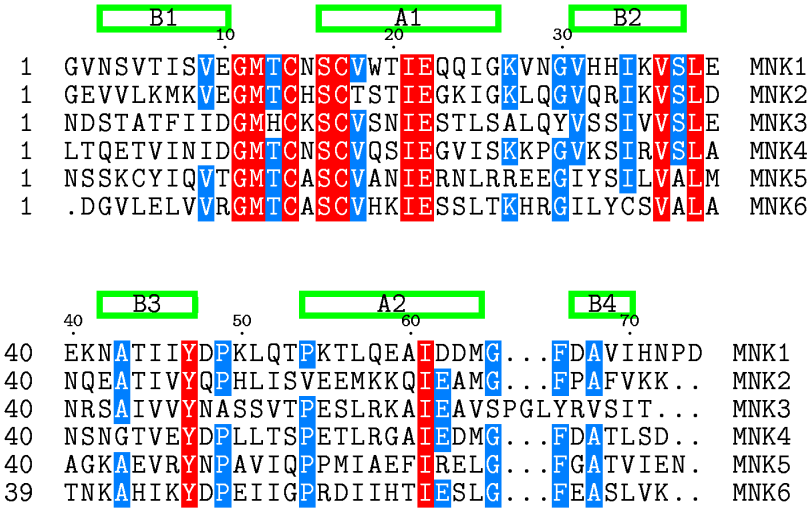

Figure 5.4: Sequence alignment of Menkes Metal Binding Domains Mnk1 to Mnk6 used in the MD simulations. Secondary structure labelling (A for α, B for β) corresponds to the sequence of Mnk4.

5.3.2 Time evolution of energy and box dimensions

5.3.2.1 Energy

Energy was correctly equilibrated for all systems after 4 ns. The standard deviations of the total CHARMM potential energy of the systems measured from the last 5 ns of 15 ns MD simulations are shown in Table 5.4. Standard deviations are similar for all systems.

Mnk1		Mnk2		Mnk3		Mnk4		Mnk5		Mnk6	
Apo	Holo	Apo	Holo	Apo	Holo	Apo	Holo	Apo	Holo	Apo	Holo
76.7	77.2	79.1	80.5	72.9	73.8	76.2	80.8	76.1	79.7	74.0	76.2

Table 5.4: Standard deviations (σ in kcal/mol) of the total CHARMM potential energy of the systems measured from the last 5 ns of 15 ns MD simulations.

5.3.2.2 Box dimensions

Two periodic box dimensions (b and c) were fixed during the simulations (NPAT ensemble). The variable box dimension (a) was rapidly equilibrated as shown in the example case of Mnk5 Holo in Fig. 5.5. Average dimension is 48.5 Å with a very low standard deviation of 0.05 Å .

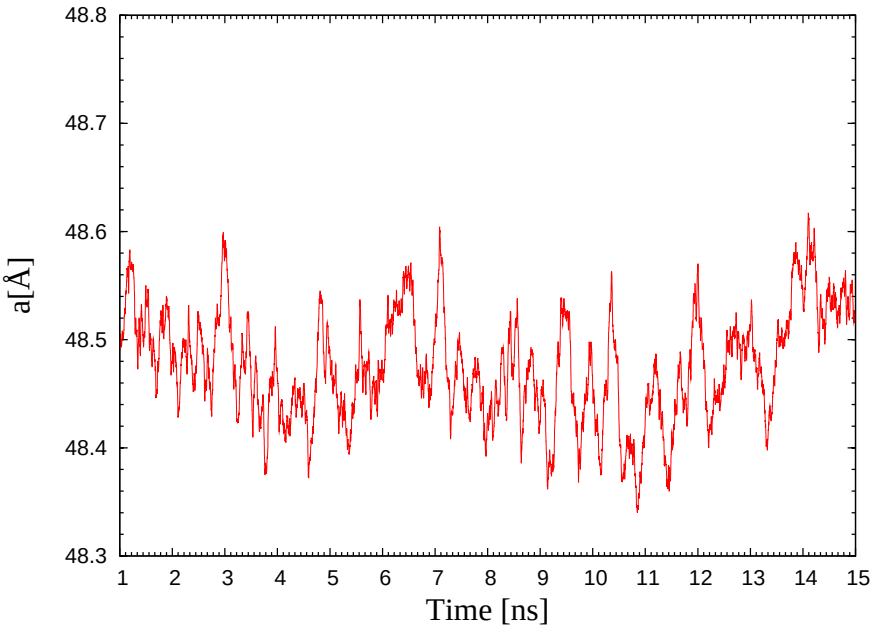

Figure 5.5: Variable dimension of the simulation box vs. time for the simulation of Mnk5 with copper.

5.3.3 Coordinate RMSD from initial structure

5.3.3.1 Session 1

Backbone atom coordinate RMSD from initial NMR structures during the 15 ns MD runs (Session 1) for Apo and Holo Mnk systems, are shown in Fig. 5.6 and Fig. 5.7, respectively. Quantitative analyses of the curves are given in Tables 5.5 and 5.6 for Apo and Holo Mnk systems, respectively. In the Apo form, only Mnk6 shows nicely low RMSDs around 1.5 Å from its NMR structure, while the 5 others depart significantly from the NMR structures with a maximum average deviation of 2.9 Å for Mnk1. This high value for Mnk1 agrees with the large experimental RMSD between models observed in structure 1KVI (See Table 5.1). In their Holo (Cu^+) forms, Mnk1, Mnk2, Mnk4 and Mnk6 show very low RMSDs, meaning for Mnk1, Mnk2 and Mnk4 a dramatic stabilization of the structures when copper is added. High RMSDs of 2.5 Å remain for Mnk3 and Mnk5, in agreement for the latter with the experimental RMSD between models of 2.4 Å and also with recent simulations of the fifth metal binding domain of the Wilson protein [180].

5.3.3.2 Following sessions: Apo

The large RMSDs from initial NMR structures obtained for some Mnk systems motivated us to run other simulations with increased number of restraints.

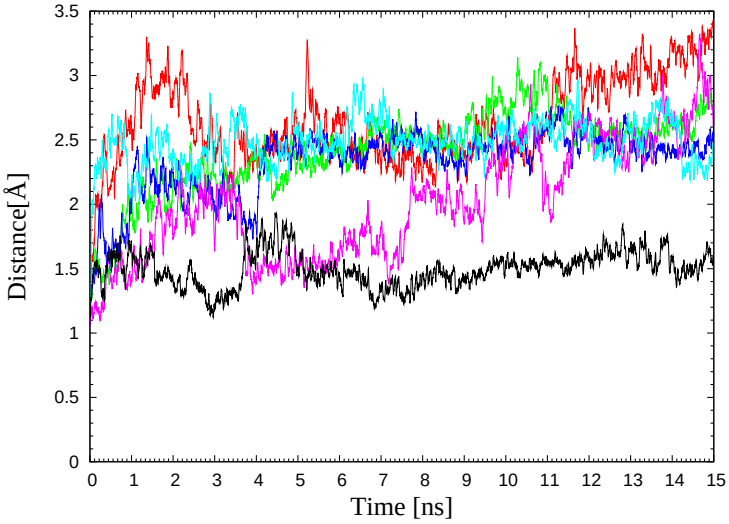

Figure 5.6: Backbone atom coordinate RMSD from initial NMR structures during MD simulations for Apo Mnk (session 1); Red: Mnk1; Green: Mnk2; Blue: Mnk3; Pink: Mnk4; Cyan: Mnk5; Black: Mnk6

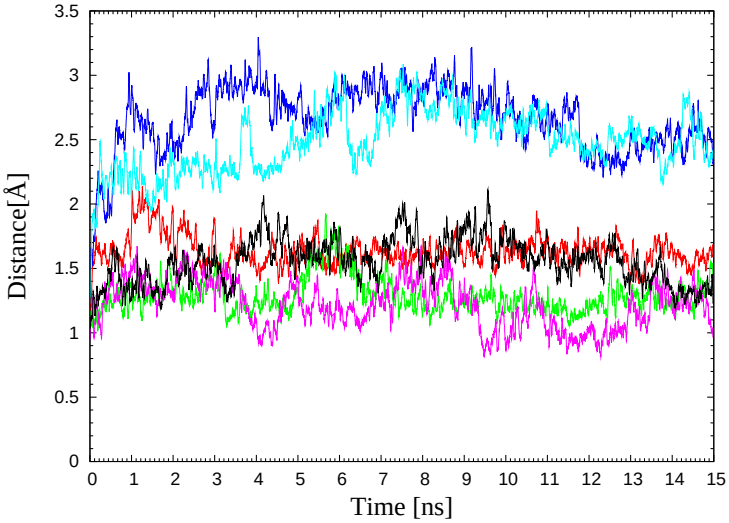

Figure 5.7: Backbone atom coordinate RMSD from initial NMR structures during MD simulations for Holo Mnk (session 1); Red: Mnk1; Green: Mnk2; Blue: Mnk3; Pink: Mnk4; Cyan: Mnk5; Black: Mnk6

System	<RMSD-bb> (σ) (Å)	<RMSD-Tot.> (σ) (Å)
Mnk1	2.9 (0.28)	4.2 (0.24)
Mnk2	2.6 (0.16)	3.9 (0.13)
Mnk3	2.4 (0.12)	3.5 (0.11)
Mnk4	2.5 (0.25)	3.6 (0.28)
Mnk5	2.5 (0.17)	3.7 (0.15)
Mnk6	1.5 (0.10)	2.9 (0.12)

Table 5.5: Average RMSD (standard deviation in parentheses) of backbone (bb) or all (Tot.) atom coordinates from their initial values in the NMR structures. Averages were calculated over the last 5 ns of the 15 ns MD runs and for Mnk Apo systems (session 1).

System	<RMSD-bb> (σ) (Å)	<RMSD-Tot.> (σ) (Å)
Mnk1	1.6 (0.10)	2.8 (0.09)
Mnk2	1.2 (0.11)	2.5 (0.09)
Mnk3	2.5 (0.17)	3.5 (0.17)
Mnk4	1.1 (0.17)	2.1 (0.17)
Mnk5	2.5 (0.14)	3.7 (0.13)
Mnk6	1.4 (0.15)	2.8 (0.12)

Table 5.6: Average RMSD (standard deviation in parentheses) of backbone (bb) or all atom (Tot.) coordinates from their initial values in the NMR structures. Averages were calculated over the last 5 ns of the 15 ns MD runs and for Mnk Holo systems (session 1).

System	<RMSD-bb> (σ) (Å)	<RMSD-Tot.> (σ) (Å)
Mnk1	1.9 (0.10)	3.0 (0.07)
Mnk2	1.2 (0.08)	2.6 (0.06)
Mnk3	1.6 (0.12)	2.8 (0.12)
Mnk4	1.8 (0.11)	2.9 (0.10)
Mnk5	2.4 (0.08)	3.5 (0.09)
Mnk6	1.4 (0.13)	2.8 (0.12)

Table 5.7: Average RMSD (standard deviation in parentheses) of backbone (bb) or all (Tot.) atom coordinates from their initial values in the NMR structures. Averages were calculated over the last ns of the 5 ns MD runs and for Mnk Apo systems in the presence of all H-bond restraints (session 3).

A summary of average RMSDs calculated from the last ns analysis of 5 ns MDs in the presence of all H-bond restraints (session 3) is shown in Table 5.7. A decrease of the RMSDs from initial NMR structures is clearly visible compared to results of session 1 (See Table 5.5). This is especially true for Mnk1 from 2.9 to 1.9 Å (see Fig. 5.8) and Mnk2 from 2.6 to 1.2 Å (see Fig. 5.9). The RMSDs have also been calculated with respect to the centroid of the 10 NMR models of each structure. The result is shown in black for Mnk1 in Fig. 5.8. The centroid xas computed with the program *kclust* of the MMTSB tools [181]. No significant difference was observed compared to RMSDs calculated from NMR model 1. Session 2 where we introduced 6 distance restraints instead of 4 was a complete failure with a still larger increase of the RMSD, and was abandoned. Noticeably, a large drift of the structure of Mnk5 is still visible (See Fig. 5.10), even after all H-bonds are maintained through NOE-like restraints.

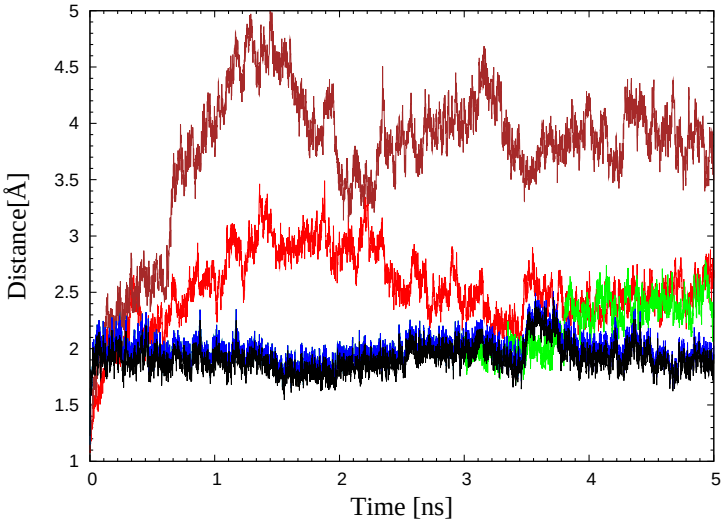

Figure 5.8: Backbone atom coordinate RMSD from initial NMR structures during MD simulations for Apo Mnk1: Red: Session 1 with 4 restraints only; Brown: Session 2 with 6 restraints; Blue: Session 3 where all possible H-bond restraints are introduced during the 5 ns MD run; Green: Session 4 with all restraints during the first 3 ns MD and released during the last 2 ns; Black: same simulation but RMSDs are calculated from the centroid of the NMR structures instead of the first model.

Running the same simulations of Apo Mnk systems but after release of the restraints during the last 2 ns (session 4) gives results summarized in Table 5.8. Mnk2, 3, 4 and 6 seem to have reached their stable conformation. A significant increase of RMSD after removing the restraints is visible in the case of Mnk1. No big change is visible for Mnk5 which either is not yet equilibrated or tries to reach an equilibrium far from the NMR structure.

5.3.3.3 Following sessions: Holo

Additional MD sessions were run for Mnk3 and Mnk5, only. All other Mnk systems were correctly equilibrated after session 1 in their Holo form. In session 6, a clear decrease of the RMSDs from 2.5 to 1.6 Å (with respect to session 1) is visible for Mnk3 and Mnk5 in Table 5.9. This is illustrated in Fig. 5.11 showing the backbone atom coordinate RMSDs of Holo-Mnk5 during the different sessions. Thus, in their Holo form, Mnk3 and Mnk5 could be stabilized by enforcing that H-bonds defining

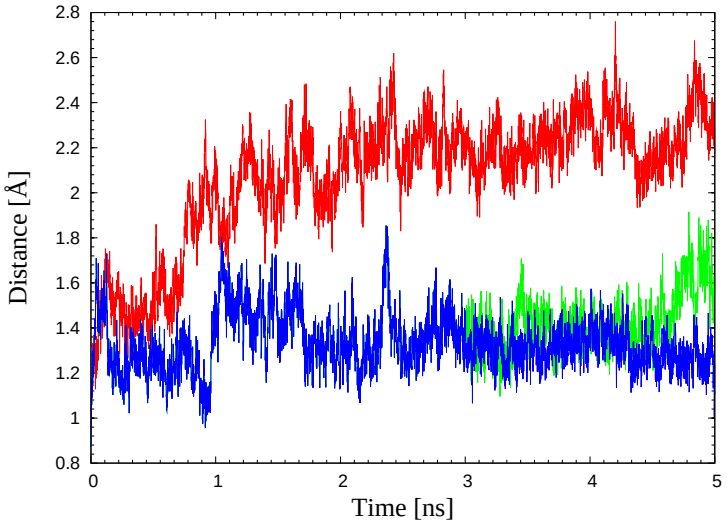

Figure 5.9: Backbone atom coordinate RMSD from initial NMR structures during MD simulations for Apo Mnk2: Red : Session 1 with 4 restraints only; Blue : Session 3 where all possible H-bond restraints are introduced during the 5 ns MD run; Green : Session 4 with all restraints during the first 3 ns MD and released during the last 2 ns.

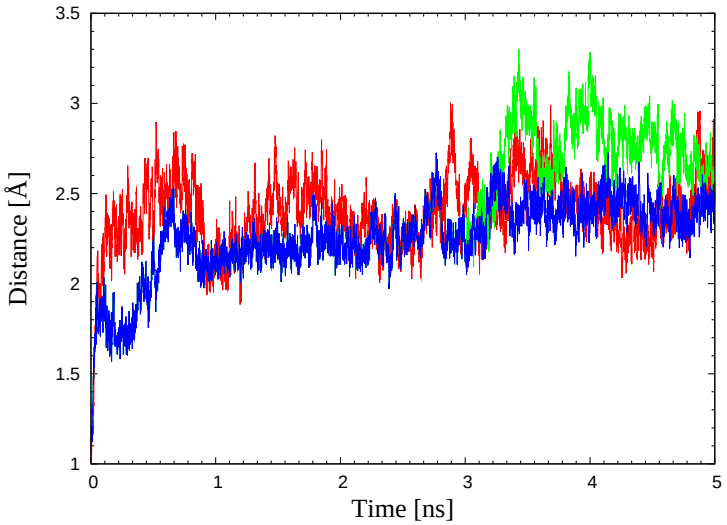

Figure 5.10: Backbone atom coordinate RMSD from initial NMR structures during MD simulations for Apo Mnk5: Red : Session 1 with 4 restraints only; Blue : Session 3 where all possible H-bond restraints are introduced during the 5 ns MD run; Green : Session 4 with all restraints during the first 3 ns MD and released during the last 2 ns.

System	<RMSD-bb> (σ) (Å)	<RMSD-Tot.> (σ) (Å)
Mnk1	2.3 (0.13)	3.4 (0.12)
Mnk2	1.4 (0.13)	2.8 (0.12)
Mnk3	1.6 (0.12)	2.8 (0.12)
Mnk4	1.5 (0.12)	2.8 (0.12)
Mnk5	2.7 (0.14)	3.9 (0.11)
Mnk6	1.2 (0.11)	2.6 (0.10)

Table 5.8: Average RMSD (standard deviation in parentheses) of backbone (bb) or all (Tot.) atom coordinates from their initial values in the NMR structures. Averages were calculated over the last ns of the 5 ns MD runs and for Mnk Apo systems. Restraints were set during 3 ns and released for the last 2 ns (session 4).

secondary structures be conserved during the simulations.

System	<RMSD-bb> (σ) (Å)	<RMSD-Tot.> (σ) (Å)
Mnk3	1.6 (0.11)	2.5 (0.09)
Mnk5	1.6 (0.15)	2.9 (0.09)

Table 5.9: Average RMSD (standard deviation in parentheses) of backbone (bb) or all (Tot.) atom coordinates from their initial values in the NMR structures for Mnk3 and Mnk5. Averages were calculated over the last ns of the 5 ns MD runs and for Mnk Holo systems in the presence of all H-bond restraints (session 6).

5.3.4 Conservation of secondary structure

We just proved that enforcing H-bonds defining secondary structure elements of the MBDs could have a large influence on their stabilization, especially in their Apo form. We will now see the direct effect on these secondary structure elements.

5.3.4.1 Session 1

Apo

A secondary structure analysis with stride [167] reflects the results concerning RMSDs presented above. Graphs of the evolution of secondary structure elements of Mnk1, Mnk5 and Mnk6, in their Apo form, versus time, are shown in Fig. 5.12, 5.13 and 5.14, respectively. Interesting information can be drawn from these plots:

- As already mentioned, Mnk6 presents a nice conservation of all its secondary structure elements at all time.

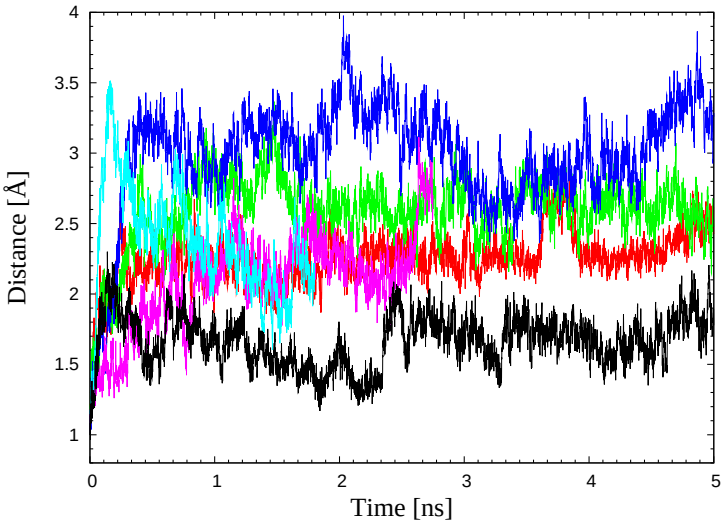

Figure 5.11: Backbone atom coordinate RMSD from initial NMR structures during MD simulations for Holo Mnk5: Red : Session 1; Green : Session 2; Blue: Session 3; Pink: Session 4; Cyan: Session 5 and Black: Session 6

- Alpha-helix α-1 shows instabilities in both Mnk1 and Mnk5. The α-helix changes for π- or 3-10-helices or even β-bridges. Helix α-2, on the other hand, is very stable in all Mnk systems.

- Globally, β-strands are conserved even though they are shortened most of the time.

Quantitative details about the conservation of the secondary structures for all Menkes MBDs in their Apo form are given in Table 5.10. A diminution of the number of residues in α-helix from the NMR initial structure to the structure obtained after MD is clear for Mnk1, 2 and 3 and compensated by an equal increase of turns (hydrogen bonded turn), or the formation of a π-helix in the case of Mnk1. Mnk4 and Mnk6 appear very stable in terms of secondary structure. Mnk5 undergoes dramatic disappearance of β strands turning into coils or β bridges (single pair of β-sheet hydrogen bond formation). 3-10 helices (3-turn helices) when present in the NMR structures disappear after MD and minimization.

Sec. Struct.	Mnk1 (74)		Mnk2 (72)		Mnk3 (74)		Mnk4 (72)		Mnk5 (73)		Mnk6 (71)	
	NMR	MD	NMR	MD	NMR	MD	NMR	MD	NMR	MD	NMR	MD
α-helix	24	20	18	16	23	20	24	24	22	22	23	24
3-10-helix	3		3	3	-		-		3		3	
β-strand	20	15	21	22	24	24	25	24	25	17	23	23
Turn	7	20	19	19	18	20	15	14	15	14	13	12
Coil	20	19	11	12	9	10	8	10	8	16	9	12

Table 5.10: Number of residues in each type of secondary structure for Apo Mnk. Results of stride are given for the initial PDB and for the coordinates obtained after 15 ns MD and minimization with CHARMM (session 1). After MD, 5 residues forming a π-helix (5-turn helix) appear in Mnk1 and 4 Bridges appear in Mnk5.

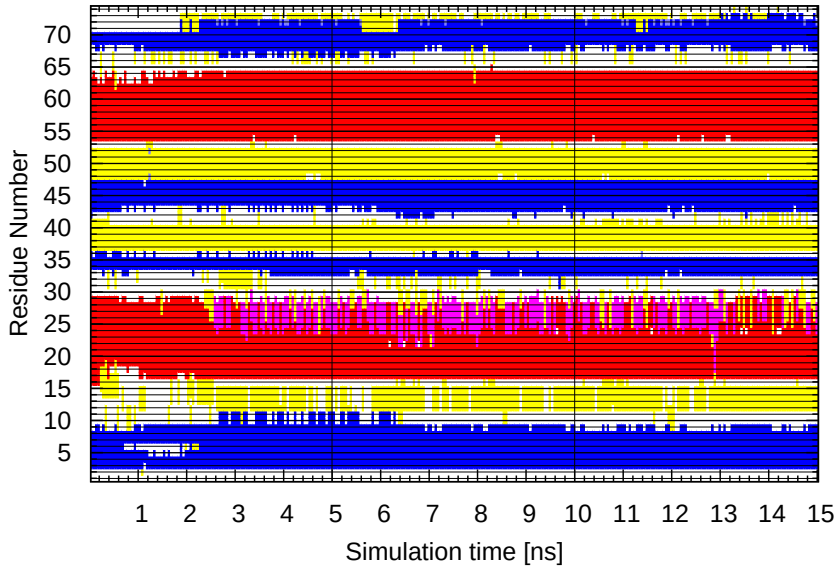

Figure 5.12: Evolution of secondary structure elements of Mnk1 Apo with time; color code is red for α-helices, blue for β-sheet, yellow for hydrogen-bonded turns, light red for 3-10 helices, magenta for π-helices, light blue for β-bridges and white for coils. The $\beta\alpha\beta\beta\alpha\beta$ fold is clearly visible.

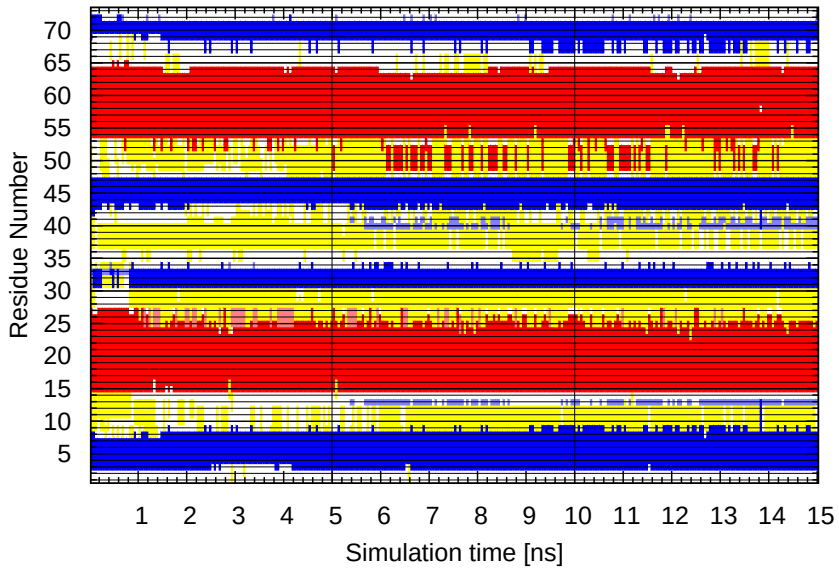

Figure 5.13: Evolution of secondary structure elements of Mnk5 Apo with time; Same color code as in Fig. 5.12

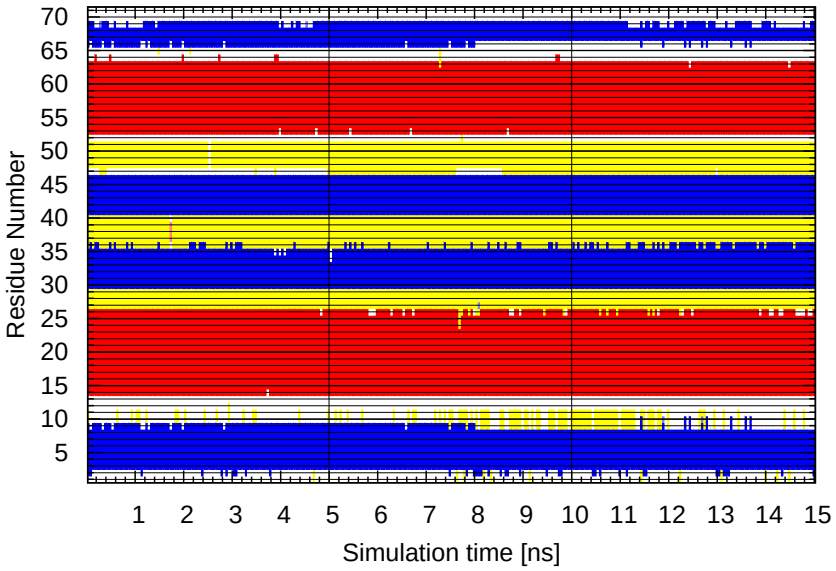

Figure 5.14: Evolution of secondary structure elements of Mnk6 Apo with time; Same color code as in Fig. 5.12

Holo

In their Holo forms, Mnk1, 2, 4 and 6 are stable in terms of secondary structure. As an example, a graph of the evolution of Mnk2 Holo with time is shown in Fig. 5.15.

Noticeably, for Mnk3, an increase of the number of amino acids in α-helix and β-strands is observed after MD at the expense of turns (see Table 5.11 and Fig. 5.16). Logically from the RMSD results, loss of structuration for Holo Mnk5 is observed both in Table 5.11 and in Fig. 5.17. Number of amino acids in α-helix, β-strands and turns, all decrease significantly and are replaced by coils. Strand β-2 almost vanishes.

Sec. Struct.	Mnk1 (74)		Mnk2 (72)		Mnk3 (74)		Mnk4 (72)		Mnk5 (73)		Mnk6 (71)	
	NMR	MD	NMR	MD	NMR	MD	NMR	MD	NMR	MD	NMR	MD
α-helix	24	24	24	24	22	24	24	24	23	21	24	24
β-strand	25	21	22	22	18	23	25	25	20	15	24	23
Turn	12	16	16	16	26	15	14	14	21	16	16	13
Coil	13	13	10	10	7	12	9	9	9	19	7	11

Table 5.11: Number of residues in each type of secondary structure for Holo Mnk. Results of stride are given for the initial PDB and for the coordinates obtained after 15 ns MD and minimization with CHARMM (session 1). 1 additional bridge was found in the NMR structure of Mnk3 and Mnk5 while 2 Bridges remain in Mnk5 after MD.

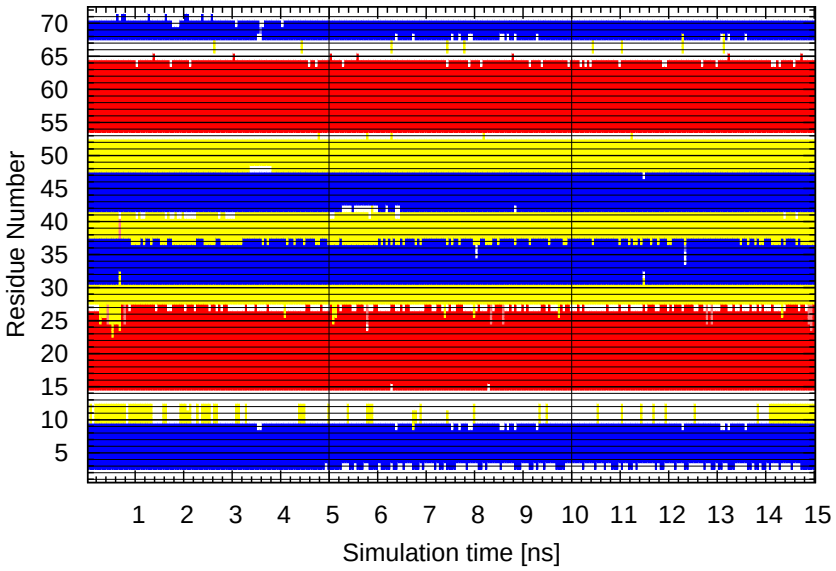

Figure 5.15: Evolution of secondary structure elements of Mnk2 Holo with time; Same color code as in Fig. 5.12

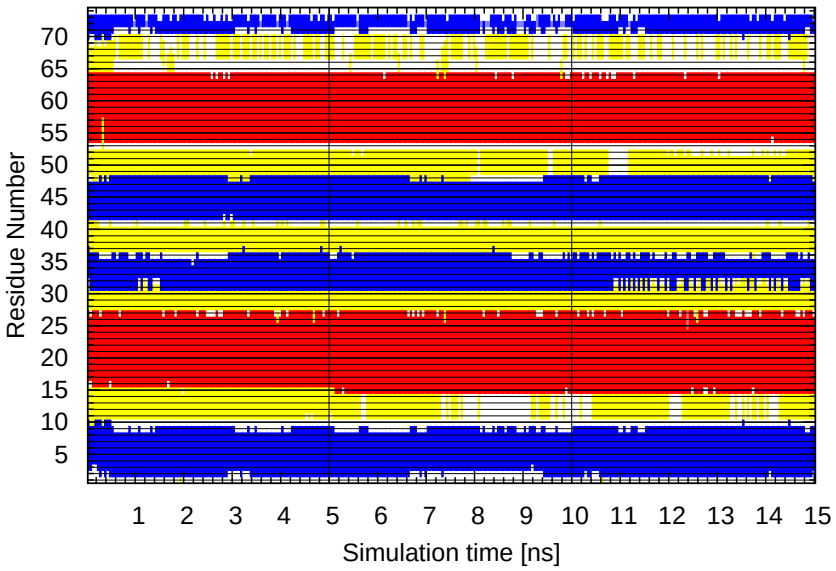

Figure 5.16: Evolution of secondary structure elements of Mnk3 Holo with time; Same color code as in Fig. 5.12

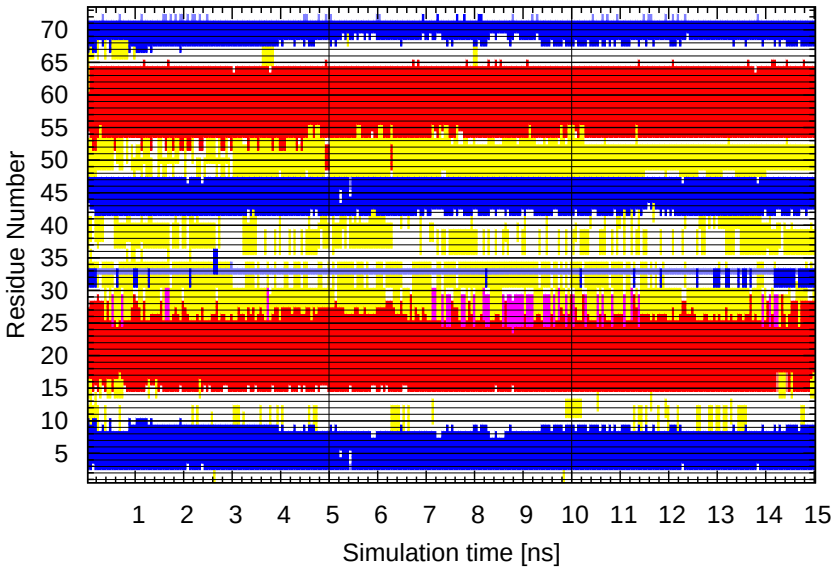

Figure 5.17: Evolution of secondary structure elements of Mnk5 Holo with time; Same color code as in Fig. 5.12

5.3.4.2 Following sessions

After session 4 (Apo) and 6 (Holo) a summary of secondary structure elements is shown in Table 5.12. Mnk1 and Mnk5-Apo have regained stability of their α-1 helix and β-strands. Consistently with the introduction of restraints on all H-bonds, most residues in Mnk3 and Mnk5 in their Holo form belong to structured motif, eben more than in the original NMR structures.

Sec. Struct.	Mnk1-Apo	Mnk5-Apo	Mnk3-Holo	Mnk5-Holo
α-helix	26	21	23	22
β-strand	18	23	25	22
Turn	10	16	21	20
Coil	20	11	5	8

Table 5.12: Number of Residues present in different secondary structures for some Mnk systems after CHARMM MD and minimisation: Mnk1-Apo after 5 ns MD, session 4; Mnk5-Apo after 5 ns MD, session 4; Mnk3-Holo after 9 ns MD, session 6; Mnk5-Holo after 5 ns MD, session 6 has 1 additional bridge

5.3.5 Root Mean Square Fluctuations

Root Mean Square Fluctuations (RMSF) of atom coordinates correspond to standard deviations of these coordinates around their average value in time. (RMSDs are averaged over several atoms or

structures at a given time; RMSFs are averaged over time for a given atom or structure). Average RMSF values in the MD simulation are usually considered as the criterion for overall flexibility of the system. The C_α atom coordinate RMSFs averaged over the last 5 ns of 15 ns MD runs (session 1) are plotted in Figures 5.18 to 5.23 for Mnk1 to 6, respectively. Logically, the residues in the α and β-folded regions are the least mobile, with mean RMSF around 0.5 Å in both Apo and Holo forms. More importantly, except for Mnk3 and Mnk5, a global reduced mobility of the residues in the Holo form is clearly visible in the loops compared to the Apo form, notably for residues in and around the Metal Binding Site (MBS), in agreement with NMR results, with a maximum reduction of the fluctuations around 2 Å for Mnk1.

Mnk3 and Mnk5 show a greater mobility in the presence of Cu, with average RMSFs 0.5 Å larger than in the Apo form. Mnk6 is the MBD with lowest average fluctuations of its backbone atoms with the exception of its loop L2 which, in its Holo form, exhibits RMSFs larger than 2 Å. This agrees with recent simulations of the sixth metal binding domain of the Wilson protein [180]. In general, in the Apo forms, large fluctuations are observed in loop L1 between β_1 and α_1, in agreement with the frequent absence of experimental NOEs in this region. Although NMR measurements and MD simulations generally reflect dynamical properties of proteins in the same picoseconds to sub-microsecond timescales, the line broadening observed experimentally, in the region of residues 11-16, is very probably due to conformational averaging occurring on time scales much longer than nanoseconds and thus difficult to completely reproduce using MD simulations.

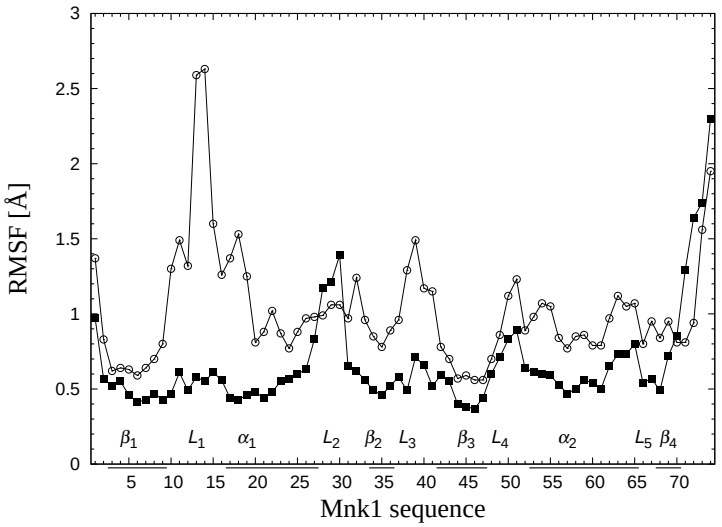

Figure 5.18: Root mean square fluctuations of the Cα atom positions of Mnk1 averaged over the 5 last ns of a 15 ns MD simulation. Empty circles for the Apo form and filled squares for the Cu^+ Holo form. Secondary structure elements are highlighted.

An analysis of the RMSFs of the two cysteines (noted Cys1 and Cys2) involved in matal binding is shown in Table 5.13. For all MBDs, in both their Apo and Holo forms, Cys1 which always belongs to loop L1 has a higher RMSF than Cys2. This is in complete agreement with NMR results for Mnk MBDs and also for the metallochaperones ApoCopZ, CuCopZ, Atx1 and Hah1, where the first cysteine residue was found more mobile than the second one. In general, both cysteines exhibit a reduction of their fluctuations when Cu^+ is bound. Mnk5 is an exception with an increase of the

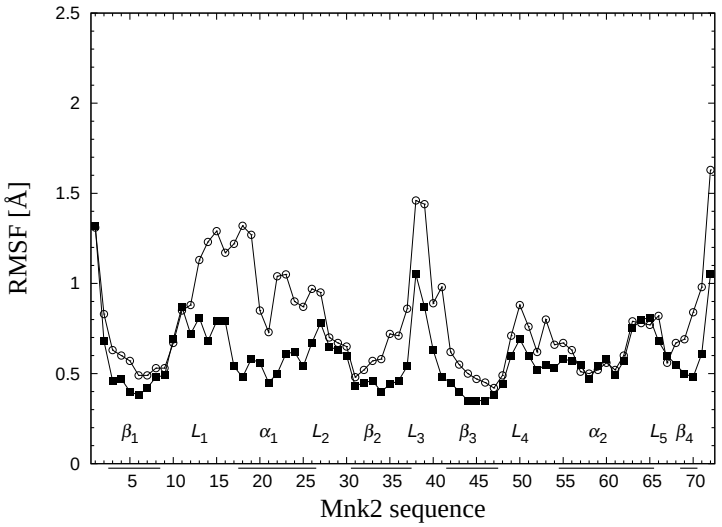

Figure 5.19: Root mean square fluctuations of the Cα atom positions of Mnk2 averaged over the 5 last ns of a 15 ns MD simulation. (Same legend as in Fig. 5.18)

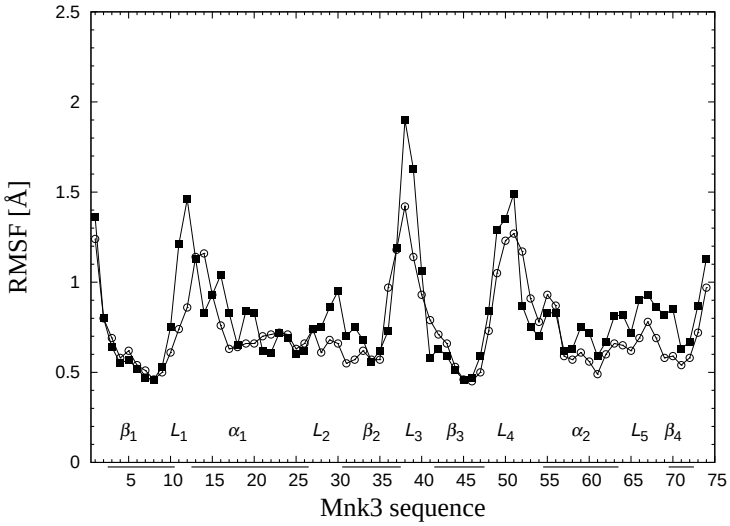

Figure 5.20: Root mean square fluctuations of the Cα atom positions of Mnk3 averaged over the 5 last ns of a 15 ns MD simulation. (Same legend as in Fig. 5.18)

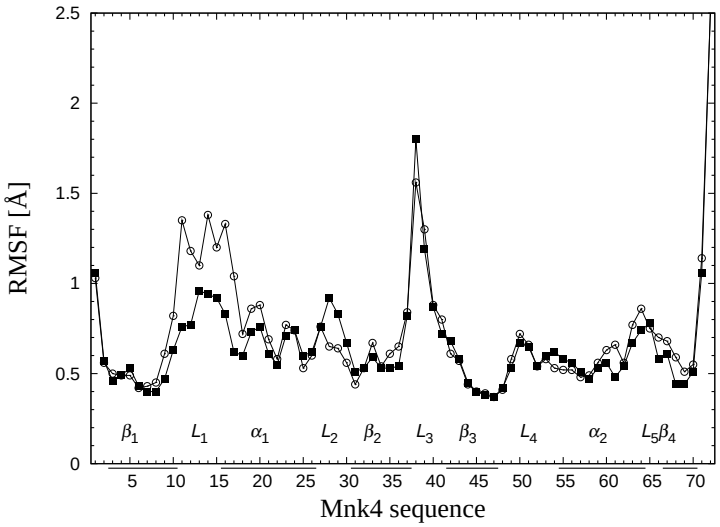

Figure 5.21: Root mean square fluctuations of the Cα atom positions of Mnk4 averaged over the 5 last ns of a 15 ns MD simulation. (Same legend as in Fig. 5.18)

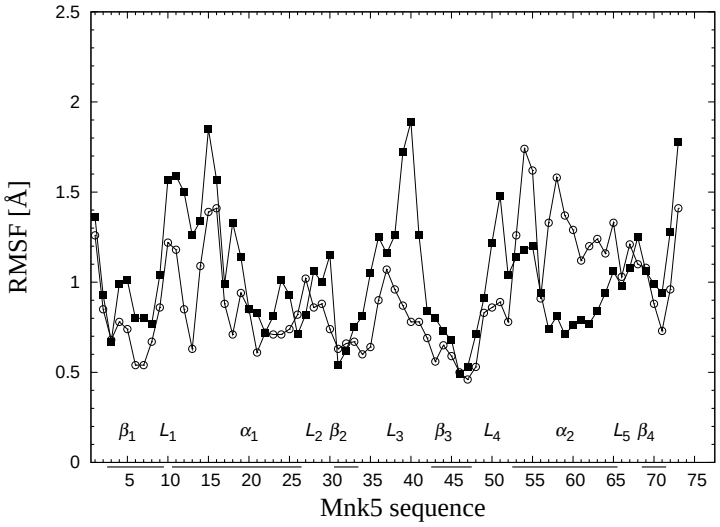

Figure 5.22: Root mean square fluctuations of the Cα atom positions of Mnk5 averaged over the 5 last ns of a 15 ns MD simulation. (Same legend as in Fig. 5.18)

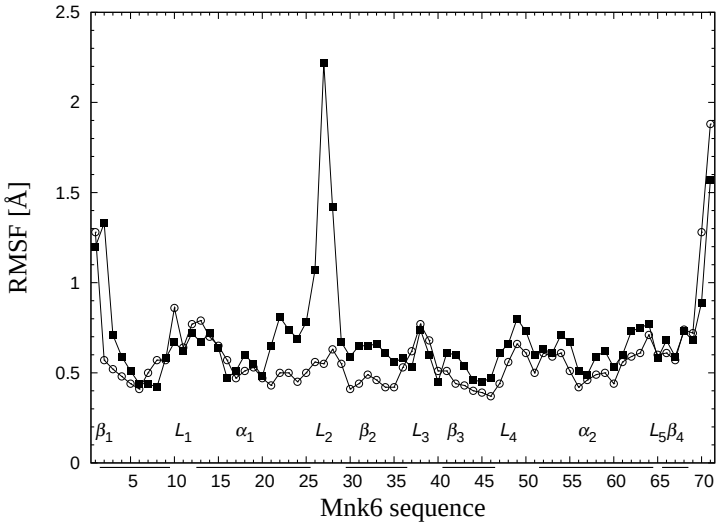

Figure 5.23: Root mean square fluctuations of the Cα atom positions of Mnk6 averaged over the 5 last ns of a 15 ns MD simulation. (Same legend as in Fig. 5.18)

fluctuations of these cysteines upon metal binding. Mnk1 is outstanding in terms of dynamics of the cysteines of its MBS: both cysteines show the largest fluctuations of all MBDs in their Apo form and the largest reduction leading to the smallest fluctuations of all MBDs in their Holo form.

Cys	Mnk1		Mnk2		Mnk3		Mnk4		Mnk5		Mnk6	
	Holo	Apo	Holo	Apo	Holo	Apo	Holo	Apo	Holo	Apo	Holo	Apo
Cys1	0.55	2.63	0.68	1.23	0.83	1.16	0.94	1.38	1.34	1.09	0.67	0.79
Cys2	0.44	1.37	0.54	1.22	0.83	0.63	0.62	1.04	0.99	0.88	0.47	0.57

Table 5.13: RMSFs of the Cα atom positions of the two metal binding cysteine residues of the CxxC motif. RMSFs are averaged over the 5 last ns of a 15 ns MD simulation.

5.3.6 Radial Distribution Function of water around Copper

A radial distribution function (RDF), g(r), describes the probability: g(r).δr of finding an atom (oxygen of water molecule) at a distance between r and $r+\delta r$ from another atom (Cu^+) as a function of the atom separation r. RDF has been widely used to study the dynamical structural modification of dense systems, the dynamical properties of metal ion hydration, and to search metal binding sites in proteins during MD simulations. The exact nature of the copper coordination sphere in Menkes MBDs has remained for years the subject of several pungent questions. Results of the radial distribution function of water oxygen around Cu^+ are presented in Figure 5.24 for all Menkes MBDs and a summary of the results is shown in Table 5.14. From the first peak in the radial distribution function, we observe that a first "shell" of water molecules is located between 3.1 and 3.2 Å (x_H) away from the copper ion. A notably low number of two water molecules (N_W) hydrate the Cu^+ ion in Mnk1, 2, 4 and 6. This number becomes 3 or 4, reliably and significantly higher in the case of Mnk3 and Mnk5 for which the copper then appears more solvent-exposed than for the other MBDs. There appears to be some kind

of relation between the RMSDs of the MBDs with respect to their initial structure (See Table 5.6) and the number of bound waters in proximity to the copper ion: the more mobile the structure is, the more solvent exposed is the Cu^+ ion. Note also that 7.5 Å away from Cu^+, in Mnk2 Mnk4 and Mnk6, the distribution of water molecules around the ion is already that of the bulk.

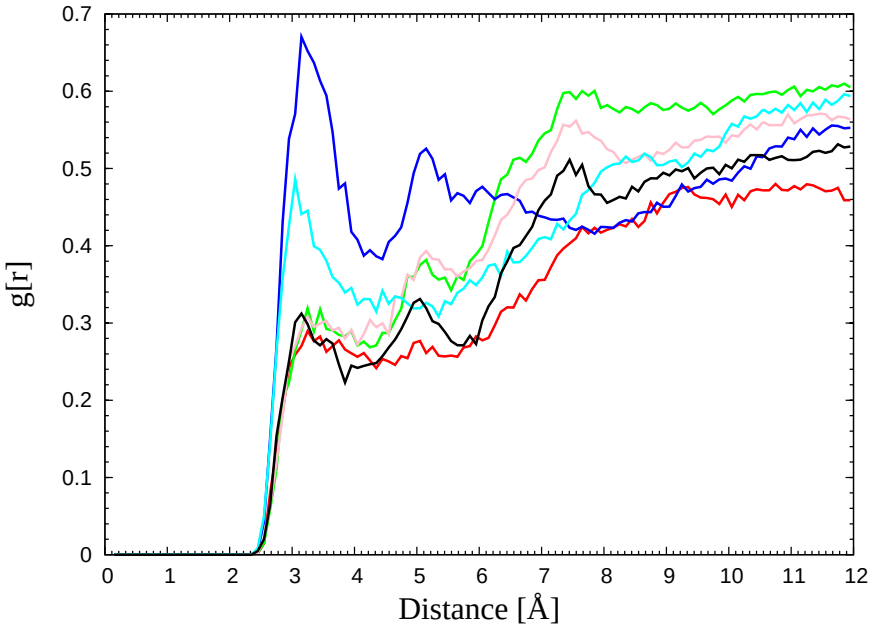

Figure 5.24: Radial distribution function for all HoloMnk systems from 5 last ns of 15 ns MD runs (session 1): Red : Mnk1; Green : Mnk2; Blue : Mnk3; Pink : Mnk4; Cyan : Mnk5; Black : Mnk6

g(r) data	Mnk1	Mnk2	Mnk3	Mnk4	Mnk5	Mnk6
x_H	3.25	3.25	3.15	3.25	3.05	3.15
g_H	0.28	0.31	0.67	0.30	0.48	0.31
x_L	4.35	4.25	4.15	4.05	4.35	3.85
g_L	0.24	0.26	0.38	0.27	0.31	0.22
$< g >$	0.47	0.59	0.53	0.56	0.57	0.51
N_W	2.4	2.3	3.9	1.9	3.4	1.5

Table 5.14: Analysis of radial distribution functions of the Holo Mnk simulations (from the last 5 ns of 15 ns MD runs, session 1): x_H is the distance between water oxygen and Cu^+ corresponding to the first peak in g(r):g_H; x_L and g_L are the distance and g values corresponding to the following minimum of g(r). The position of this minimum is used as higher bound for the integration of the first peak yielding the average number of water molecules in the first hydration sphere of Cu^+: N_W; Finally, $< g >$ is the average value (for r between 10 and 12 Å) of the plateau reached by g(r) for large r.

I also calculated the radial distribution function of water oxygens around Cu^+ for Mnk3 and Mnk5 in the MD simulation run in the presence of all H-bond restraints, corresponding to session 6. The results compared to those obtained in session 1 are shown in Figure 5.15 and summarized in Table 5.25. Significant variations exist between the results in the two sessions meaning that either MD runs in

session 6 were too short or that restraints on the secondary structure of the MBDs has a non negligible influence on the distribution of waters around Cu^{+}. However, the main result of a higher hydration of Mnk3 and Mnk5 by 3 to 4 water moleculs in first sphere is still valid.

Information	Mnk3	Mnk5
x_{H}	3.15	3.05
g_{H}	0.69	0.63
x_{L}	4.65	4.15
g_{L}	0.24	0.38
$< g >$	0.58	0.63
N_{W}	4.43	3.84

Table 5.15: Analysis of radial distribution functions of Holo Mnk3 and Mnk5: session 6 after 5 ns MD with full restraints. Same legend as for Table 5.14.

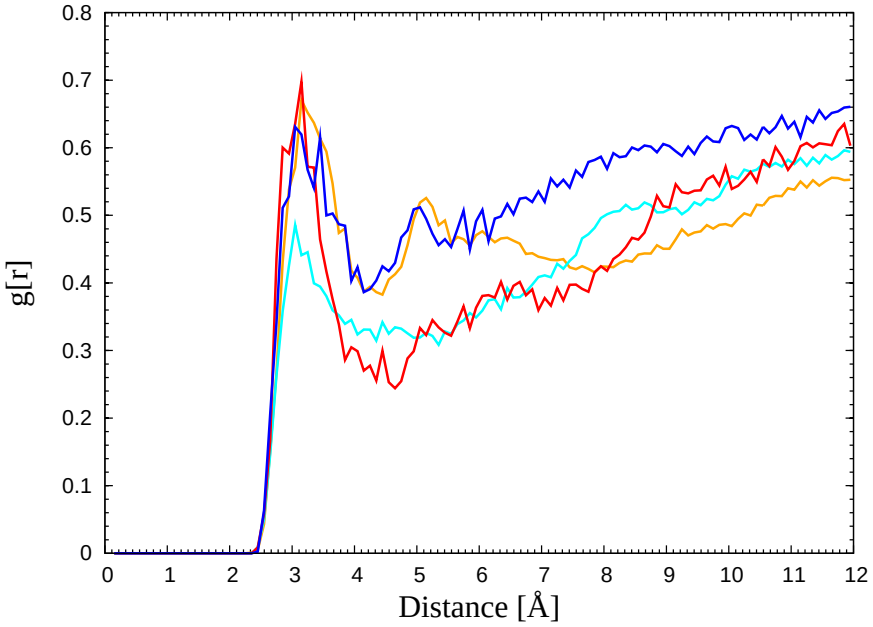

Figure 5.25: Influence of the number of restraints on radial distribution functions for Holo Mnk3 and Mnk5 (from 1 ns of MD runs in session 6). Orange: Mnk3 session 1; Cyan: Mnk5 session 1; Red: Mnk3 session 6; Blue: Mnk5 session 6.

5.3.7 Average structures

Finally, to illustrate the previous results, I show in Tables 5.16 and 5.17 average structures of the 6 Mnk systems calculated over the last 5 ns of the 15 ns MD runs (session 1) for Apo and Holo Mnk, respectively. The figures nicely summarize and complete results obtained from the analysis of RMSFs. We observe:

- Large fluctuations of loops L1 and L3 and beginning of helix α-1 for all Mnk MBDs in their Apo form, except Mnk6.

- High stabilization of L1 and α-1 in Mnk1 to Mnk4 upon Cu^+ binding.

- Large movement of helix α-2 in Mnk5 Apo stabilized upon Cu^+ binding.

- Large fluctuations in general at the C-terminal.

- Large movement of loop L2 in Mnk6 Holo.

Globally, Mnk6 Apo and Mnk5 with bound Cu^+ appear as the most and least stable structures, respectively.

5.4 Summary, discussion and perspectives

We have carefully simulated and analyzed the dynamics of the six MBDs of the Menkes protein starting from the coordinates of their solution structure. All the simulations were run on isolated MBDs whereas they are probably tightly packed in the real protein. This could explain why some of them were unstable showing large coordinate RMSDs from the initial structure; we corrected this problem by introducing NOE-like distance restraints maintaining the secondary structure elements of the proteins through their H-bonds. We noticed, in some cases, like for Mnk1 (session 2), that a slight increase in the number of restraints certainly lowered the probability of intra-helix or intra-strand fraying but, at the same time, introduced dramatic loss of inter-secondary structure element interactions.

From the published structures of Mnk MBDs [183] and from our simulations, each domain consists of four β-strands and two α-helices folded into a stable $\beta\alpha\beta\beta\alpha\beta$ "ferredoxin-like" structure. The structures of the individual domains are quite similar. The third metal binding domain, Mnk3, which has the greatest amount of sequence variation (as shown in section 5.3.1), has some notable differences in the metal binding region, but retains the same general fold. In all MBDs, the conserved GM(T/H)CxxC copper binding motif lies at the surface of the domain in the β_1-α_1 loop, with the second copper-coordinating cysteine at the beginning of the α_1 helix. Apo-structures of MBDs show that this loop is flexible and relatively unstructured. Except for Mnk3 and Mnk5, upon binding of Cu(I), the loop acquires rigidity with the copper atom bound between the two cysteine residues in a linear coordinate environment. The S-Cu(I)-S linear geometry has been suggested by X-ray absorption spectroscopy with S-Cu(I) bond lengths of 2.2-2.3 Å [184, 185]. In Mnk1, the experimental average S-Cu-S bond angle for the calculated structures of Mnk1 is 170 degrees [171]. Other experiments on Mnk4 show that Cu^+ is strongly coordinated to the two cysteines and is weakly bound to a third sulfur atom of an exogenous thiol or a Met side chain [174]. In addition to Cys residues directly involved in copper coordination, several other amino acids contribute to copper binding environment within individual MBDs. Mutations of the "X" residues in the CxxC motif have been shown to alter the flexibility of the metal binding loop [171]. The loop is situated near a core of hydrophobic residues, which provide a stable association between the β_1 strand and the α_1 helix regardless of whether or not copper is bound. Four conserved residues contribute to this core - Met12 in the β_1- α_1 loop, Ile21 in the α_1 helix, Leu38 in the β_2- β_3 loop, and Phe66 in the α_2- β_4 loop (see Fig. 5.4). In all N-terminal MBDs, except Mnk3, Phe66 is conserved.

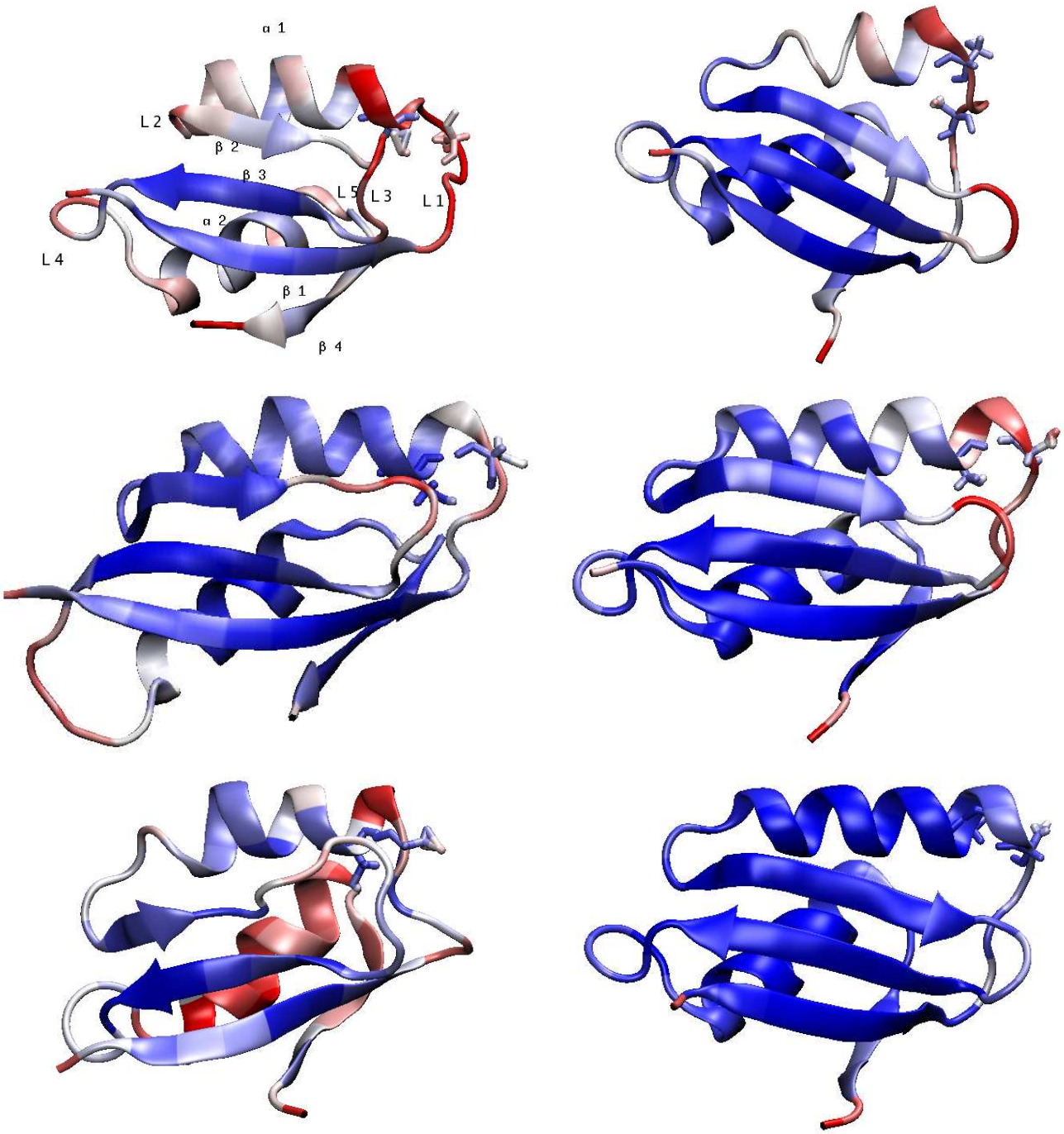

Table 5.16: Average structures of the 6 Mnk systems, in their Apo form, calculated over the last 5 ns of the 15 ns MD runs (session 1). Mnk1 to 6 from top to bottom and left to right. Only secondary structure elements and the two Cys residues forming the MBS are represented. Atoms are color coded from blue to red corresponding to backbone atom coordinate RMSDs from the initial NMR structures from 0.5 to 1.5 Å. Figures have been drawn with program VMD [182].

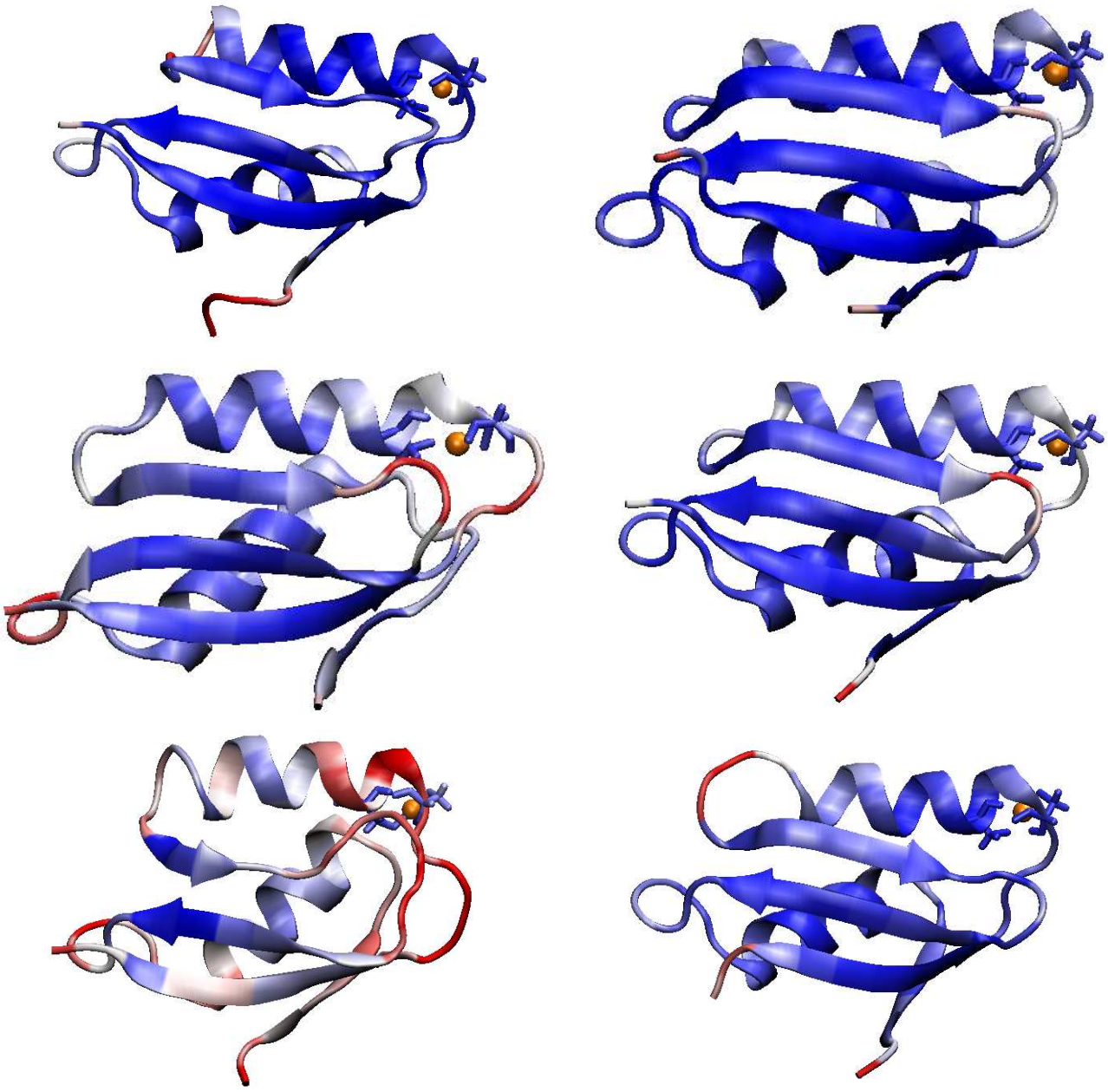

Table 5.17: Average structures of the 6 Mnk systems, in their Holo form, calculated over the last 5 ns of the 15 ns MD runs (session 1). Mnk1 to 6 from top to bottom and left to right. Only secondary structure elements, the two Cys residues and the Cu^{+} ion forming the MBS are represented. Color legend as in Table 5.16.

The third metal binding domain has the most unique sequence of the six domains, as well as having lower affinity for copper in solution [183]. In Mnk3 the residue Phe66 is substituted by a Pro residue. however the hydrophobic pocket is maintained by the presence of a Tyr at position 69. The α_2 helix and the β_4 strand are shorter, generating an extended α_2-β_4 loop that is unique to MBD3. Also unlike other domains, the Apo-form of Mnk3 features an extended α_1 helix which incorporates both metal coordinating cysteines. To bind copper, the helix must be unwound in order for the coordinating sulfur groups to form a linear geometry with the bound Cu(I) molecule. As the structures of individual MBDs were solved in absence of the rest of the protein, it is possible that in the full-length Cu-ATPase such unwinding occurs as the result of interactions of Mnk3 with other MBDs (or perhaps with the copper chaperone), making this domain more suitable for copper binding. From our studies, Mnk3 appears the MBD for which the binding of Cu^+ has the least influence on the dynamics and fluctuations of the domain compared to the Apo form.

As already mentioned, although XAS of individual MBDs shows a dominant two-coordinate copper binding environment, a third ligand can contribute to copper coordination [186]. In solution, such ligands could be provided by reducing reagents present in the buffers, including DTT, glutathione, or TCEP, Researchers in our team have recently developped parameters for Cu(I) bound to 3 sulfur atoms and it might be interesting to run new simulations with a trigonal geometry around the ion for the Holo Mnk MBDs, by adding a glutathione molecule to the system, for example. The structural rigidity of individual MBDs, emerging from NMR data and our analyses of secondary structures, implies that significant conformational changes observed in the entire N-terminal region of Cu-ATPases upon copper binding [184] take place in the loops connecting these individual domains.

The loops connecting N-terminal MBD vary in length (Fig. 5.26). In addition, the N-terminus of human ATP7B has an extension of 63 amino acid residues, which is critical for proper trafficking of ATP7B in polarized hepatocytes [187]. ATP7A lacks the very N-terminal extension, but has a long (91 amino-acid residues) insert between Mnk1 and Mnk2, whereas ATP7B has a much shorter MBD1,2 linker (13 residues). Secondary structure predictions for both the N-terminal extension of ATP7B and the ATP7A MBD1,2 insert suggest that these regions are folded and thus may represent small recognition domains for protein-protein interactions that are unique for the corresponding Cu-ATPases.

Although specific length of connecting loops varies, the N-terminal domains of ATP7A and ATP7B share a common motif where the linker between MBDs 4 and 5 is much longer than the one between MBD5 and 6 (Fig. 5.26). This observation and the presence of only one or two MBDs in bacterial Cu-ATPase suggest that MBD5,6 may represent an autonomous sub-domain within the N-terminus of Cu-ATPases. Interesting simulations could be run from the solution structure of these two domains of the Wilson protein, the only structure where one MBD is not completely isolated from the others [162].

As described in section 1.5.2, Atox1 is the copper metallochaperone associated to ATP7A and ATP7B in humans. In vitro, Cu-Atox1 has been demonstrated to transfer copper directly to the N-terminal domain of ATP7B in a dose dependent fashion [188]. All six sites can be filled in this manner, though a significant excess of Atox1 (5-50 fold over the N-terminal domain of ATP7B) is needed to do so.

Individually, the N-terminal MBDs have similar copper association constants that are equivalent or lower than that of Atox1. However, the ability of MBDs to accept copper from Atox1 in the

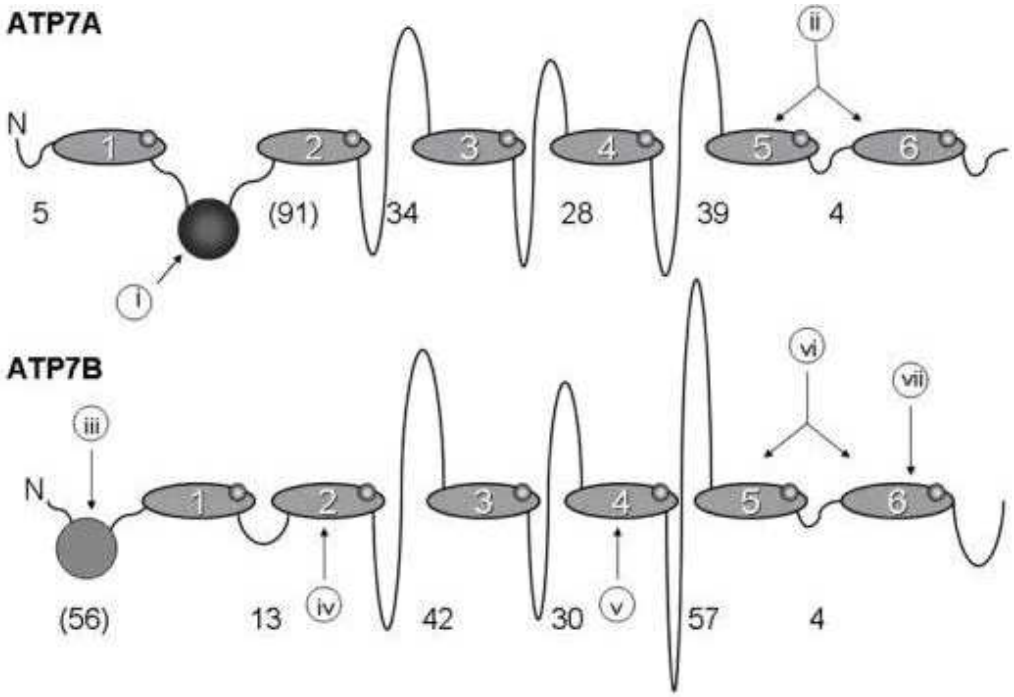

Figure 5.26: Schematic of N-Terminal domains of ATP7A (A) and ATP7B (B) The length of inter-domain loops (in amino acid residues) is indicated. Loop is defined as the segment between folded 72-aa MBDs. (i) The loop between metal binding sites 1 and 2 of ATP7A is predicted to have a structure similar to the other metal binding sites, but lacks copper coordinating residues. (ii) Metal binding sites 5 and 6 have been shown to be necessary for proper trafficking of ATP7A. (iii) The very N-Terminal region of ATP7A is also essential for proper trafficking and membrane targeting. (iv) MBS 2 has been shown to be the primary site of copper transfer from Atox1. (v) MBS 4 has a higher affinity for Atox1 than does MBS2, and also has been shown capable of transferring copper to MBSs 5 and 6. (vi) MBDs 5 and 6 are essential for copper transport, and have been shown to regulate affinity of the Intra-membrane copper binding site. (vii) Location of the G591D mutation in ATP7B (eqv. G30 in Mnk6, see Fig. 5.4), a disease causing mutation that disrupts Atox1 interactions with the N-Terminal domain.

presence of other domains differs [189]. When Cu-Atox1 is presented to the N-terminal domain of ATP7B in vitro, the copper is preferentially transferred to the MBD2. It is currently not clear whether better exposure or other characteristics of MBD2 make it a site of preferential copper transfer from Atox1. When presented as individual domains, MBD4 has a stronger affinity for Cu-Atox1 than does MBD2 [162], but in the fully folded protein MBD4 may not be sufficiently exposed and may only become available after transfer of the first copper to MBD2. It has been also proposed that the charged surface exposed residues may contribute to chaperone-target recognition. Atox1 has a considerable positively charged surface region featuring a number of lysine residues in proximity to the metal binding loop. Negatively charged regions of several MBDs may be complimentary [185], however data from the laboratory [190] and others does not support the idea of surface complimentarity being a driving force for selective Atox1-MBD recognition. Rather, the complementary surface charge distribution may be an important factor in packing of the MBDs within the N-terminal domain. The observation that the distribution of charges on the surface of MBDs of ATP7A and ATP7B appear to follow the same pattern [191] is consistent with this idea.

For ATP7A, little is known about the copper transfer process, though themes similar to those obtained for Atox1-ATP7B interactions are emerging. When presented to the entire N-terminal tail (Mnk1-6), Cu-Atox1 showed, in solution, the formation of detectable amounts of a macromolecular complex with both domains 1 and 4 whereas domain Mnk6 was able to remove Cu(I) from the metallochaperone [192]. Presented with MBDs 4-6 of ATP7A, only, Atox1 demonstrated a preference for donating copper to Mnk4 [193], although Mnk5 and 6 of ATP7A can also accept copper from Cu-Atox1 at variance with ATP7B. From our calculations on isolated domains, of the fluctuations of the MBS in Mnk1, the large entropy loss of loop L1 upon metal binding could be a reason for proposing an unfavorable transfer of the metal to the first MBD compared to the five others. Mnk3 of ATP7A is poorly metallated by Cu-Atox1 in solution, likely due to the unusual organization of its metal-binding site. From our results, Mnk2 and Mnk4 show a bulk-like exposure of the metal after 7.5 Å distance from the copper ion and this exposure is maximum for Mnk2. Mnk3 is clearly the MBD for which the most stable "shell" of water molecules assemble around the copper. Could it be that this shell "protects" Mnk3 from Cu^+ transfer from Atox1? (Researchers in the team have shown from quantum chemical calculations that Cu^+ presents unfavourable interactions with water oxygen [59].) Could it be that a good transfer of Cu^+ from the chaperone to the MBDs means a low density of waters around the metal?

Banci2007

Ongoing projects in the team aim at calculating binding affinities of all MBDs for Cu^+ using Free Energy Perturbation (FEP) techniques. These calculations are expensive in terms of computer time for a proper sampling of the configurations of loop L1, very flexible in the Apo form of the MBDs. Other projects involve using more sophisticated QM/MM- type methods to study the transfer of the metal from chaperone to ATPase.

Chapter 6

Conclusion

The sheer complexity of a protein system might make it virtually intractable for conventional scientific approaches, in the sense that the intricate relationships between all constituent parts of the protein are in themselves so complex and convoluted that any simplification or division into separate entities, as is generally done in science, must be artificial and arbitrary and worse, only valid for the level at which this simplification or subdivision was made. This means that no general rules derived from any simplified model of a protein will hold for the real protein system. This statement is bold, since it implies that only a model that describes the whole protein (and possibly a substantial part of its surroundings) in virtually exact detail, will ever be able to describe the protein behaviour in such a way that any conclusions can be derived from it. It also implies that, if we happen to stumble on only part of the "correct" description, it might be difficult, if not impossible, to tell that this description was in fact correct. In other words, it becomes very difficult to gauge if a certain direction of research on proteins is "on track", i.e. likely to provide useful results in the long run. It would be like a blind man stumbling around in the dark, not knowing where he is or where he is going, only to be able to "feel" the destination when he is right at it (the worst case being when the destination cannot even be "felt", making it no more efficient than random diffusion is as a means of finding your way home after a wild party).

In the years spent in Dr. Serge's team I have been able to explore many aspects of P1-type ATPases structure and function. However, the objective of thesis is not only answering a number of questions regarding a certain topic, but also to open new lines of research to explore. In this regard, future research on the function and structure function of P1-type ATPases should address answers to many key questions. Membrane protein structure prediction is still a difficult problem, the more new structures became available, the more I realized that we did not know much, even it gets worst if there is no homology structure available.

In this thesis, I have predicted the TMHs region of Cadmium ATPase. For predicting these TMHs, I have used standard programs like MODELLER, XPLOR, CHARMM, and participated in the development of our own programs in collaboration with INRIA (AMD). First we have refined the topology of the TMHs in CADA (where do they begin and end, in sequence). Then, we have defined models corresponding to all possible arrangements of the TM bundle (initial positioning of one helix with respect to all others). In the next stage, we have built 3D structure determination (all coordinates) of these models using XPLOR and a procedure similar to structure determination from

NMR experiments with distance restraints imposed by the topology or experimental findings. Then, we have used Adaptive Molecular Dynamics (AMD) program, to interactively check the models for bad loop positioning or "knots" in the structure. Finally, we have refined these models using CHARMM and MD simulations in the presence of an implicit membrane. An MD simulation has been performed with the aim to obtain the optimized structure. It is shown that MD is able to provide structural and dynamical information about the membrane proteins not accessible from experimental methods.

Next, I have studied N-terminal conserved GxTCxxC metal binding domain(s) (MBDS) of P1 type ATPase. In the case of the human Cu+ transporting Menkes ATPase, there are 6 MBDs, each of them being able to bind Cu^{+}. The structure of each MBD separately is known by NMR spectroscopy but the structure of the assembly is unknown. Using MD simulations, we have studied the dynamics of each MBD in the presence or absence of metal with the final goal of understanding how the metal is transferred from the metallochaperone to the MBD and why the presence of 6 MBDs is needed for the correct functioning of the pump. These studies take advantages of our recent work in the field of parameterization of metal ions for molecular mechanics force fields and MD studies on metallochaperones. We show that fast approximate in silico models of metal ions can help understand metal binding and transport in metalloproteins or ATPases, which are known to be major pharmaceutical targets.

Chapter 7

Supplementary Material

7.1 Program for finding topological models of CadA: ”buildtopo”

7.2 First page of output of ”buildtopo”

7.3 All 2D-grid models of CADA

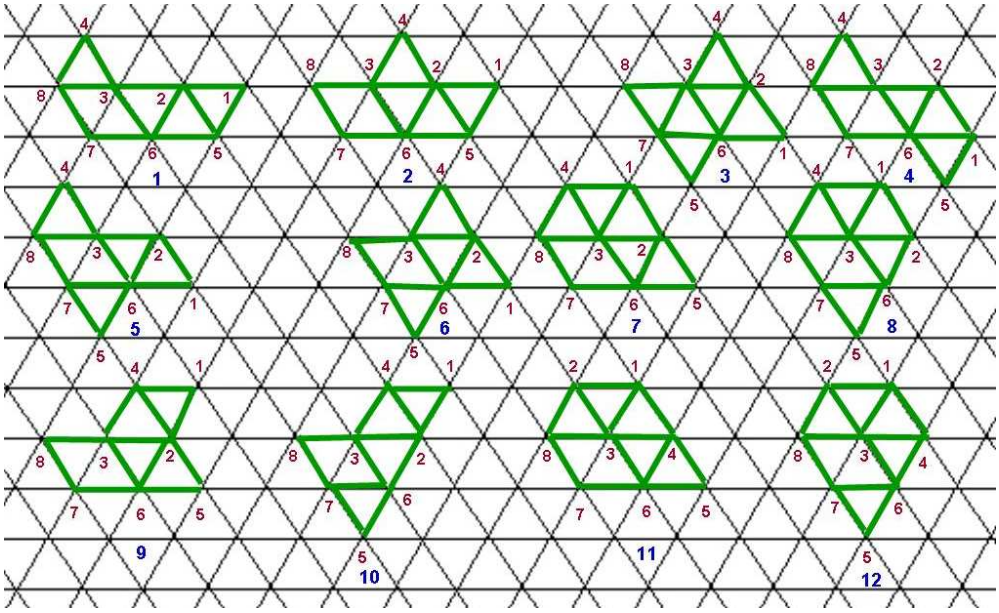

Figure 7.1: 2D-grid models 1 to 12; The average radius of Models from 1 to 12 are 1.07, 1.02, 1.04, 1.01, 1.01, 1.00, 0.97, 0.97, 1.01, 1.04, 0.97 and 0.97, respectively.

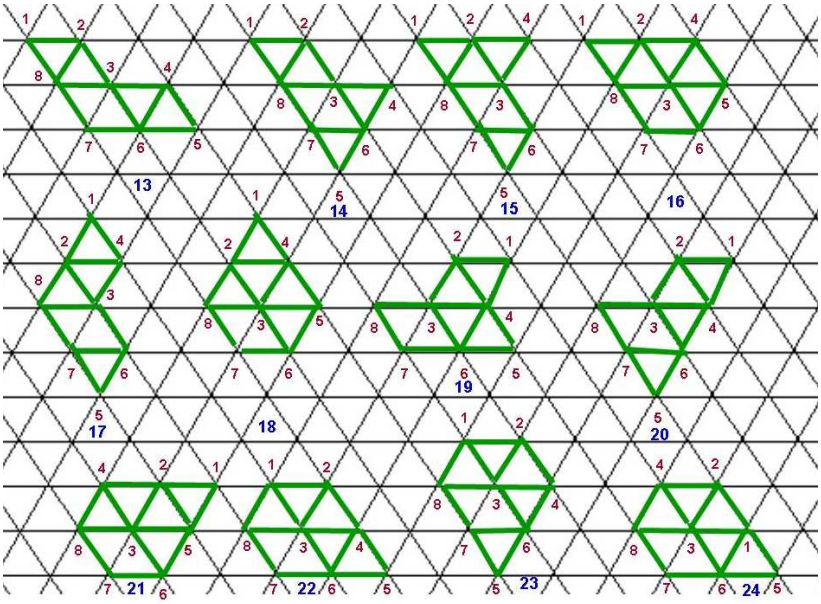

Figure 7.2: 2D-grid models 13 to 24;The average radius of Models from 13 to 24 are 1.06, 1.04, 1.04, 0.97, 1.06, 0.97, 1.01, 1.04, 0.97, 0.97, 0.97 and 0.97, respectively.

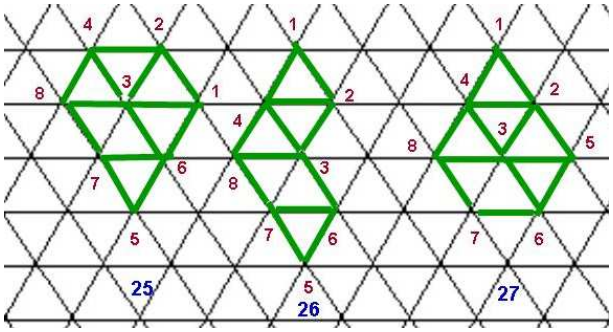

Figure 7.3: 2D-grid models 25 to 27; The average radius of Models from 25 to 27 are 0.97, 1.06 and 0.97, respectively.

7.4 Abstracts of published articles

Bibliography

[1] Watson J.D. and Crick F.H. Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid. *Nature* 171, 737–738 (1953).

[2] Nirenberg M. Historical review: Deciphering the genetic code–a personal account. *Trends Biochem Sci* 29, 46–54 (2004).

[3] Anfinsen C.B. Principles that govern the folding of protein chains. *Science* 181, 223–230 (1973).

[4] Levinthal C. Are there pathways for protein folding? *Journal de Chimie Physique et de Physico-Chimie Biologique* 65, 44–45 (1968).

[5] Singer S.J. and Nicolson G.L. Fluid Mosaic Model of Structure of Cell-Membranes. *Science* 175, 720–731 (1972).

[6] White S.H. and Wimley W.C. Membraneprotein folding and stability:Physical principles. *Annu. Rev. Bioph. Biom.* 28, 319–365 (1999).

[7] Lee A.G. Lipid-protein interactionsin biological membranes: a structural perspective. *Biochim. Biophys. Acta* 1612, 1–40 (2003).

[8] Engelman D.M. Membranes are more mosaic than fluid. *Nature* 438, 578–580 (2005).

[9] Guidotti G. Membrane Proteins. *Annu. Rev. Biochem.* 41, 731 (1972).

[10] Tanford. The Hydrophobic Effect. *Wiley* (1973).

[11] Simons K. and Ikonen E. Functional rafts in cell membranes. *Nature* 387, 569–572. (1997).

[12] Bachar M., Brunelle P., Tieleman D.P., and Rauk A. Molecular dynamics simulation of a polyunsaturated lipid bilayer susceptible to lipid peroxidation. *J. Phys. Chem.* 108, 7170–7179. (2004).

[13] Margulies M., Egholm M., Altman W.E., Attiya S., Bader J.S., Bemben L.A., Berka J., Braverman M.S., Chen Y.J., and Chen Z. Genome sequenc-ing in microfabricated high-density picolitre reactors. *Nature* 437, 376–380 (2005).

[14] Russell R.B. and Eggleston D.S. New roles for structure in biology and drug discovery. *Nat Struct Biol* 7 Suppl:, 928–930 (2000).

[15] Hopkins A.L. and Groom C.R. The druggable genome. *Nat Rev Drug Discov* 1, 727–730 (2002).

[16] Krogh A., Larsson B., von Heijne G., and Sonnhammer E.L. Predicting transmembrane topology with a hidden Markov model: application to complete genomes. *J. Mol. Biol.* 305, 567–580 (2001).

[17] Dong C., Beis K., Nesper J., Brunkan-Lamontagne A.L., Clarke B.R., Whitfield C., and Naismith J.H. Wza the translocon for E. coli capsular polysaccha-rides defines a new class of membrane protein. *Nature* 444, 226–229 (2006).

[18] Wimley W.C. Toward genomic identification of beta-barrel membrane pro-teins: composition and architecture of known structures. *Protein Sci* 11, 301–312 (2002).

[19] Garrow A.G., Agnew A., and Westhead D.R. TMB-Hunt: an amino acid composition based method to screen proteomes for beta-barrel transmembrane proteins. *BMC Bioinformatics* 6, 56 (2005).

[20] Picot D., Loll P.J., and Garavito R.M. The X-ray crystal structure of the membrane protein prostaglandin H2 synthase-1. *Nature* 367, 243–249 (1994).

[21] Gupta K., Selinsky B.S., Kaub C.J., Katz A.K., and Loll P.J. The 2.0 A resolution crystal structure of prostaglandin H2 synthase-1: structural insights into an unusual peroxidase. *J. Mol. Biol.* 335, 503–518 (2004).

[22] Frausto da Silva J.J.R. and Williams R.J.P. The biological chemistry of the elements. *Oxford University Press, New York.* (2001).

[23] Horton H.R., Moran L.A., Ochs R.S., Rawn J.D., and Gray-Scrimgeour K. . *Principles of Biochemistry* 3rd ed. (2002).

[24] Hughes M. N. The inorganic chemistry of biological processes. *N Z Vet J* (1981).

[25] Outten C. E. and OHalloran T.V. Femtomolar Sensitivity of Metalloregulatory Proteins Controlling Zinc Homeostasis. *Science* 292, 2488–2492 (2001).

[26] Gatti D., Mitra B., and Rosen B.P. Escherichia coli Soft Metal Ion-translocating ATPases. *J. Biol. Chem.* 275, 34009–34012 (2000).

[27] Fraga C.G. Relevance, essentiality and toxicity of trace elements in human health. *Mol Aspects Med* 26, 235–244 (2005).

[28] Roberts V.A., Fisher C.L., Redford S.M., McRee D.E., Parge H.E., Getzoff E.D., and Tainer J.A. Mechanism and atomic structure of superoxide dismutase. *Free Radic Res Commun* 12-13 Pt 1:, 269–278 (1991).

[29] Tapiero H. and Tew K.D. Trace elements in human physiology and pathology: zinc and metallothioneins. *Biomed Pharmacother* 57, 399–411 (2003).

[30] He Z.L., Yang X.E., and Stoffella P.J. Trace elements in agroecosystems and impacts on the environment. *J Trace Elem Med Biol* 19, 125–140 (2005).

[31] Balamurugan K. and Schaffner W. Copper homeostasis in eukaryotes: teetering on a tightrope. *Biochim Biophys Acta* 1763, 737–746 (2006).

[32] Changela A., Chen K., Xue Y., Holschen J., Outten C.E., O'Halloran T.V., and Mondragon A. Molecular basis of metal-ion selectivity and zeptomolar sensitivity by CueR. *Science* 301, 1383–1387 (2003).

[33] Callahan D.L., Baker A.J., Kolev S.D., and Wedd A.G. Metal ion ligands in hyperaccumulating plants. *J Biol Inorg Chem* 11, 2–12 (2006).

[34] Clemens S. Molecular mechanisms of plant metal tolerance and homeostasis. *Planta* 212, 475–486 (2001).

[35] Rae T.D., Schmidt P.J., Pufahl R.A., Culotta V.C., and Ohalloran T.V. Undetectable intracellular free copper: the requirement of a copper chaperone for superoxide dismutase. *Science* 284, 805–808 (1999).

[36] Hamza I., Faisst A., Prohaska J., Chen J., Gruss P., and Gitlin J.D. The metallochaperone Atox1 plays a critical role in perinatal copper homeostasis. *98* Proc. Natl. Acad. Sci. U. S. A., 6848–6852 (2001).

[37] Pufahl R.A., Singer C.P., Peariso K.L., Lin S.J., Schmidt P.J., Fahrni C.J., Culotta V.C., Penner-Hahn J.E., and OHalloran T.V. Metal ion chaperone function of the soluble Cu(I) receptor Atx1. *Science* 278, :853–856 (1997).

[38] Wernimont A.K., Huffman D.L., Lamb A.L., OHalloran T.V., and Rosenzweig A.C. Structural basis for copper transfer by the metallochaperone for the Menkes/Wilson disease proteins. *Nat Struct Biol* 7, 766–771 (2000).

[39] Poger D., Fillaux C., Miras R., Crouzy S., Delangle P., Mintz E., Den Auwer C., and Ferrand M. Interplay between glutathione, Atx1 and copper: X-ray absorption spectroscopy determination of Cu(I) environment in an Atx1 dimer. *J. Biol. Inorg. Chem.* 13(8), 1239–1248 (2008).

[40] Banci L., Bertini I., Cantini F., Felli I.C., Gonnelli L., Hadjiliadis N., Pierattelli R., Rosato A., and Voulgaris P. The Atx1-Ccc2 complex is a metal-mediated protein-protein interaction. *Nat. Chem. Biol.* 2, 367–368 (2005).

[41] Brünger A.T. *X-PLOR version 3.1: A system for X-ray crystallography and NMR*. Yale University Press, New Haven, CT, (1992).

[42] Crippen G.M. and Havel T.F. *Distance geometry and molecular conformation*. Research Studies Press (Wiley), New York, (1988).

[43] Nilges M., Clore G.M., and Gronenborn A.M. Determination of three-dimensional structures of proteins from interproton distance data by dynamical simulated annealing from a random array of atoms. *FEBS Lett.* 239, 129–136 (1988).

[44] Sander C. and Schneider R. Database of homology-derived protein structures and the structural meaning of sequence alignment. *Proteins* 9(1), 56–68 (1991).

[45] Sali A. and Blundell T.L. Comparative protein modelling by satisfaction of spatial restraints. . *J. Mol. Biol.* 234(3), 779–815 (1993).

[46] Head-Gordon T. and Brooks C. Virtual rigid body dynamics. *Biopolymers* 31, 77–100 (1991).

[47] Herzyk P and Hubbard R E. A reduced representation of proteins for use in restraint satisfaction calculations. *Proteins* 17(3), 310–324 (1993).

[48] Brooks B.R.and Bruccoleri R.E., Olafson B.D., States D.J. andSwaminathan S., and Karplus M. CHARMM: a program for macromolecular energy, minimization, and dynamics calculations. . *J. Comput. Chem.* 4, 187–217 (1983).

[49] Leach A.R. Molecular Modelling. *Pearson Prentice Hall, Harlow,GB* (2001).

[50] Diehl M. and Fischer T. Skript zur Mathematik fuer die Molekulare Biotechnologie. *Universitaet Heidelberg,* 3.8 (2003).

[51] Redon S. and Lin M.C. An efficient, error-bounded approximation algorithm for simulating quasi-statics of complex linkages. *Comput. Aided Design* 38, 300–314 (2006).

[52] R. Featherstone. A divide-and-Conquer Articulated-Body Algorithm for Parrallel O(log(n)) Calculation of Rigid-Body Dynamics. Part1: Basic Algorithm. *Int. J. Robot. Res.* 18(9), 867–875 (1999).

[53] Rossi R., Isorce M., Morin S., Flocard J., Arumugam K., Crouzy S., Vivaudou M., and Redon S. Adaptive torsion-angle quasi-statics: a general simulation method with applications to protein structure analysis and design. *Bioinformatics.* 23, 408–17 (2007).

[54] McQuaerrie D.A. Statistical Mechanics. *University Science Books* Sansalito (2000).

[55] Born M. and Oppenheimer R. Zur Quantentheorie der Molekeln. *Annalen der Physik* 84, 457 (1927).

[56] Allinger N.L. MM2: A Hydrocarbon Force Field Utilizing V1 and V2 Torsional Terms. *j.am.chem.soc* 97, 8127 (1977).

[57] Allinger N.L., Li F., and Yan L. The MM3 Force Field for Alkenes. *J. Comput. Chem.* 11, 848 (1990).

[58] Allinger N.L., Chen K., and Lii J.-H. An Improved Force Field (MM4) for Saturated Hydrocarbons. *J. Comput. Chem.* 17, 642 (1996).

[59] J.-F. Fuchs, H. Nedev, D. Poger, M. Ferrand, V. Brenner, J.-P. Dognon, and S. Crouzy. New model potentials for sulfur-copper(I) and sulfur-mercury(II) interactions in proteins : from *ab initio* to molecular dynamics. *J. Comput. Chem.* 27(7), 837–856 (2006).

[60] Vollhardt K.P.C. and Schore N.E. . *Organic Chemistry* (1999).

[61] Cox S.R. and Williams D.E. Representation of Molecular Electrostatic Potential by a New Atomic Charge Model. *J. Comput. Chem.* 2, 304 (1981).

[62] Fletcher R. and Reeves C.M. . *The Computer Journal* 7, 149 (1964).

[63] Marsh D. Polarity and permeation profiles in lipid membranes. *Proc. Natl. Acad. Sci. U. S. A.* 98, 7777–7782 (2001).

[64] Richards F.M. Areas, volumes, packing, and protein structure. *Ann. Rev. Biophys. Bioeng.* 6, 151–176 (1977).

[65] Roux B. and Simonson T. Implicit solvent models. *Biophys Chem* 78, 1–20. (1999).

[66] Zhou R. Free energy landscape of protein folding in water: explicit vs. implicit solvent. *Proteins* 53, 148–61 (2003).

[67] Holtzer A. The cratic correction and related fallacies. *Biopolymers* 35, 595–602 (1984).

[68] Ooi T., Oobatake M., Némethy G., and Scheraga H.A. Accessible surface area as a measure of the thermodynamic parameters of hydration of peptides. *Proc. Natl. Acad. Sci. U. S. A.* 84, 3086–3090 (1987).

[69] Eisenberg D. and McLachlan A.D. Solvation energy in protein folding and binding. *Nature* 319, 199–203 (1986).

[70] Fogolari F., Brigo A., and Molinari H. The Poisson-Boltzmann equation for biomolecular electrostatics: a tool for structural biology. *J Mol Recognit* 15, 377–92 (2002).

[71] Shestakov A.I., Milovich J.L., and Noy A. Solution of the nonlinear Poisson-Boltzmann equation using pseudo-transient continuation and the finite element method. *Colloid Interface Sci* 247, 62–79 (2002).

[72] Lu B., Zhang D., and McCammon J.A. Computation of electrostatic forces between solvated molecules determined by the Poisson-Boltzmann equation using a boundary element method. *J Chem Phys* 122, 214102 (2005).

[73] Koehl P. Electrostatics calculations: latest methodological advances. *Curr Opin Struct Biol* 16, 142–51 (2006).

[74] Still W.C., Tempczyk A., Hawley R.C., and Hendrickson T. Semianalytical treatment of solvation for molecular mechanics and dynamics. . *J. Am. Chem. Soc.* 112, 6127–6129 (1990).

[75] Onufriev A., Bashford D., and Case D.A. . *J Comp Chem* 23, 1297–1304 (2002).

[76] Ho B.K. and Dill K.A. Folding very short peptides using molecular dynamics. *PLoS Comput Biol* 2, e27 (2006).

[77] Im W., Feig M., and Brooks C.L. An implicit membrane generalized born theory for the study of structure, stability, and interactions of membrane proteins. *Biophys J* 85, 2900–18 (2003).

[78] Eisenberg D.T. and Wesson L. Atomic solvation parameters applied to molecular dynamics of proteins in solution. *Protein Sci* 1, 227235 (1992).

[79] Lazaridis T. and Karplus M. Effective Energy Function for Proteins in Solution. *Proteins: Structure, Function, and Genetics* 35, 133–152 (1999).

[80] Ferrara P., Apostolakis J., and Caflisch A. Evaluation of a fast implicit solvent model for molecular dynamics simulations. *Proteins* 46, 24–33 (2002).

[81] Lee, Michael S., Freddie R. Salsbury Jr., and Mark A. Olson. An Efficient Hybrid Explicit/Implicit Solvent Method for Biomolecular Simulations. *j.comp.chem.* 25, 1967–1978 (2004).

[82] Sharp K.A., Nicholls A., Fine R.F., and Honig B. Reconciling the magnitude of the microscopic and macroscopic hydrophobic effects. *Science* 252, 106–109 (1991).

[83] Tamar Schlick. Molecular Modeling and Simulation: An Interdisciplinary Guide Interdisciplinary Applied Mathematics: Mathematical Biology. *Springer-Verlag New York, NY* (2002).

[84] Moreau C.J., Dupuis J.P., Revilloud J., Arumugam K., and Vivaudou M. Coupling ion channels to receptors for biomolecule sensing. *Nat Nanotechnol.* 10, 620–5 (2008).

[85] Goetz B.A., Perozo E., and Locher K.P. Distinct gate conformations of the ABC transporter BtuCD revealed by electron spin resonance spectroscopy and chemical cross-linking. *FEBS Lett.* 583, 266–70 (2009).

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

[87] Rensing C., Ghosh M., and Rosen B.P. Families of soft-metal-ion transporting ATPases. *J Bacteriol* 181, 5891–5897 (1999).

[88] Lutsenko S and Kaplan J.H. Organization of P-type ATPases: significance of structural diversity. *Biochemistry* 34, 15607–15613 (1995).

[89] Axelsen K.B. and Palmgren M.G. Evolution of substrate specificities in the P-type ATPase superfamily. *J Mol Evol* 46, 84–101 (1998).

[90] Post R.L., Hegyvary C., and Kume S. Activation by adenosine triphosphate in the phosphorylation kinetics of sodium and potassium ion transport adenosine triphosphatase. *J. Biol. Chem.* 247, 6530–6540 (1972).

[91] Eren E. and Arguello J.M. Arabidopsis HMA2, a divalent heavy metaltransporting P1B-type ATPase, is involved in cytoplasmic Zn+2 homeostasis. *Plant Physiol* 136, 3712–3723 (2004).

[92] Fan B. and Rosen B.P. Biochemical characterization of CopA, the Escherichia coli Cu(I)-translocating P-type ATPase. *J. Biol. Chem.* 277, 46987–46992 (2002).

[93] Toyoshima C, Nomura H, and Tsuda T. Lumenal gating mechanism revealed in calcium pump crystal structures with phosphate analogues. *Nature* 432(7015), 361–368 (2004).

[94] Pedersen B.P., Buch-Pedersen M.J., Morth J.P., Palmgren M.G., and Nissen P. Crystal structure of the plasma membrane proton pump. *Nature* 450, 1111–4. (2007).

[95] Moller J.V., Juul B., and le Maire M. Structural organization, ion transport and energy transduction of P-type ATPases. *Biochim. Biophys. Acta* 1286, 1–51 (1996).

[96] Arguello J.M. Identification of ion selectivity determinants in heavy metal transport P1B-type ATPases. *J Membr Biol* 195, 93–108 (2003).

[97] Melchers K, Weitzenegger T, Buhmann A, Steinhilber W, Sachs G, and Schafer KP. Cloning and membrane topology of a P type ATPase from Helicobacter pylori. *Biol. Chem.* 271, 446–457 (1996).

[98] Tsai K.J., Lin Y.F., Wong M.D., Yang H.H., Fu H.L., and Rosen B.P. Membrane topology of the p1258 CadA Cd(II)/Pb(II)/Zn(II)-translocating P-type ATPase. *J. Bioenerg. Biomembr.* 34, 147–156 (2002).

[99] Mandal A.K., Yang Y., Kertesz T.M., and Arguello J.M. Identification of the transmembrane metal binding site in Cu+-transporting P1B-type ATPases. *J. Biol. Chem.* 279, 54802–54807 (2004).

[100] Sazinsky M.H., Mandal A.K., Arguello J.M., and Rosenzweig A.C. Structure of the ATP binding domain from the Archaeoglobus fulgidus Cu+-ATPase. *J. Biol. Chem.* 281, 11161–11166 (2006).

[101] Sazinsky M.H., Agarwal S., Arguello J.M., and Rosenzweig A.C. Structure of the actuator domain from the Archaeoglobus fulgidus Cu(+)-ATPase. *Biochemistry* 45, :9949–9955 (2006).

[102] Banci L., Bertini I., Ciofi-Baffoni S., Miras R., Bal N., Mintz E., Catty P., Shokes J. E., and Scott R. A. Structural Basis for Metal Binding Specificity: the N-terminal Cadmium Binding Domain of the P1-type ATPase CadA. *J. Mol. Biol.* (2005).

[103] Arguello J.M., Eren E., and González-Guerrero M. The structure and function of heavy metal transport P 1B-ATPases . *BioMetals* 20, 233–248 (2007).

[104] Olesen C., Sorensen T.L., Nielsen R.C., Møller J.V., and Nissen P. Dephosphorylation of the calcium pump coupled to counterion occlusion. *Science* 306, 2251–2255 (2004).

[105] Banci L., Bertini, Ciofi-Baffoni S., Finney L.A., Outten C.E., and O'Halloran T.V. A new Zinc-protein coordination site in intracellular metal trafficking: solution structure of the apo and Zn(II) forms of ZntA. *J. Mol. Biol.* 323, 883–897 (2002).

[106] DiDonato M., Narindrasorasak S., J.R. Forbes, Cox D.W., and Sarkar B. Expression, purification, and metal binding properties of the N-terminal domain from the Wilson disease putative copper-transporting ATPase (ATP7B). *J. Biol. Chem.* 272, 33279–33282 (1997).

[107] Axelsen K.B. and Palmgren M.G. Inventory of the superfamily of P-type ion pumps in Arabidopsis. *Plant Physiol.* 126, 696–706 (2001).

[108] Nucifora G., Chu L., Misra T.K., and Silver S. Cadmium resistance from Staphylococcus aureus plasmid pI258 cadA gene results from a cadmium-efflux ATPase. *Proc. Natl. Acad. Sci. U.S.A.* 86, 3544–3548 (1989).

[109] Kahn D., David M., Domergue O., Daveran M.L., Ghai J., Hirsch P.R., and Batut J. Rhizobium meliloti fixGHI sequence predicts involvement of a specific cation pump in symbiotic nitrogen fixation. *J. Bacteriol.* 171, 929–939 (1989).

[110] Rensing C., Fan B., Sharma R., Mitra B., and Rosen B.P. CopA: An Escherichia coli Cu(I)-translocating P-type ATPase. *Proc. Natl. Acad. Sci. U. S. A.* 97, 652–656 (2000).

[111] Odermatt A., Suter H., Krapf R., and Solioz M. Primary structure of two P-type ATPases involved in copper homeostasis in Enterococcus hirae. *J. Biol. Chem.* 268, 12775–12779 (1993).

[112] Tottey S., Rich P.R., Rondet S.A., and Robinson N.J. Two Menkes-type atpases supply copper for photosynthesis in Synechocystis PCC 6803. *J. Biol. Chem.* 276, 19999–20004 (2001).

[113] Phung L.T., Ajlani G., and Haselkorn R. P-type ATPase from the cyanobacterium Synechococcus 7942 related to the human Menkes and Wilson disease gene products. *Proc. Natl. Acad. Sci. U. S. A.* 91, 9651–9654 (1994).

[114] Sharma R., Rensing C., Rosen B.P., and Mitra B. The ATP hydrolytic activity of purified ZntA, a Pb(II)/Cd(II)/Zn(II)-translocating ATPase from Escherichia coli. *J. Biol. Chem.* 275, 3873–3878 (2000).

[115] Mana-Capelli S., Mandal A.K., and Arguello J.M. Archaeoglobus fulgidus CopB is a thermophilic Cu2+-ATPase: Functional role of its histidine-rich N-terminal metal binding domain. *J. Biol. Chem.* 278, 40534–40541 (2003).

[116] Yuan D.S., Dancis A., and Klausner R.D. Restriction of copper export in Saccharomyces cerevisiae to a late Golgi or post-Golgi compartment in the secretory pathway. *J. Biol. Chem.* 272, 25787–25793 (1995).

[117] Petris M.J., Mercer J.F., Culvenor J.G., Lockhart P., Gleeson P.A., and Camakaris J. Ligand-regulated transport of the Menkes copper P-type ATPase efflux pump from the Golgi apparatus to the plasma membrane: a novel mechanism of regulated trafficking. *Embo J* 15, 6084–6095 (1996).

[118] Schaefer M, Hopkins R.G., Failla M.L., and Gitlin J.D. Hepatocyte-specific localization and copper-dependent trafficking of the Wilson's disease protein in the liver. *Am J Physiol* 276, G639–G646 (1999).

[119] Liu J, Dutta S.J., Stemmler A.J., and Mitra B. Metal-binding affinity of the transmembrane site in ZntA: Implications for metal selectivity. *Biochemistry* 45, 763–772 (2006).

[120] Arguello J.M., Whitis J., Cheung M.C., and Lingrel J.B. Functional role of oxygen containing residues in the fifth transmembrane segment of the Na,K-ATPase alpha subunit. *Arch. Biochem. Biophys.* 364, 254–263 (1999).

[121] Toyoshima C., Nakasako M., Nomura H., and Ogawa H. Crystal structure of the calcium pump of sarcoplasmic reticulum at 2.6 A resolution. *Nature* 405(6787), 647–655 (2000).

[122] Yoshimizu T, Omote H, Wakabayashi T, Sambongi Y, and Futai M. Essential Cys-Pro-Cys motif of Caenorhabditis elegans copper transport ATPase. *Biosci Biotechnol Biochem* 62, 1258–1260 (1998).

[123] Lowe J, Vieyra A, Catty P, Guillain F, Mintz E, and Cuillel M. A mutational study in the transmembrane domain of Ccc2p, the yeast Cu(I)-ATPase, shows different roles for each Cys-Pro-Cys cysteine. *J. Biol. Chem.* 279, 25986–25994 (2004).

[124] Bull P.C. and Cox D.V. Wilson disease and Menkes disease: new handles on heavy-metal transport. *Trends Genet* 10, 246–252 (1994).

[125] Petrukhin K, Lutsenko S, Chernov I, Ross BM, Kaplan JH, and Gilliam TC. Characterization of the Wilson disease gene encoding a P-type copper transporting ATPase: genomic organization, alternative splicing, and structure/function predictions. *Hum Mol Genet* 3, 1647–1656 (1994).

[126] Hirayama T., Kieber J.J., Hirayama N., Kogan M., Guzman P., Nourizadeh S., Alonso J.M., Dailey W.P., Dancis A., and Ecker J.R. RESPONSIVE-TO-ANTAGONIST1, a Menkes/Wilson disease-related copper transporter, is required for ethylene signaling in Arabidopsis. *Cell* 97, 383–393 (1999).

[127] Woeste K.E. and Kieber J.J. A strong loss-of-function mutation in RAN1 results in constitutive activation of the ethylene response pathway as well as a rosette-lethal phenotype. *Plant Cell* 12, 443–455 (2000).

[128] Dutta S.J., Liu J., Hou Z., and Mitra B. Conserved aspartic acid 714 in transmembrane segment 8 of the ZntA subgroup of P1B-type ATPases is a metal-binding residue. *Biochemistry* 45, 5923–5931 (2006).

[129] Bissig K.D., Wunderli-Ye H., Duda P.W., and Solioz M. Structure-function analysis of purified Enterococcus hirae CopB copper ATPase: effect of Menkes/Wilson disease mutation homologues. *Biochem. J.* 357, 217–223 (2001).

[130] Rutherford J.C., Cavet J.S., and Robinson N.J. Cobalt-dependent transcriptional switching by a dual-effector MerR-like protein regulates a cobalt-exporting variant CPxtype ATPase. *J. Biol. Chem.* 274, 25827–25832 (1999).

[131] Seigneurin-Berny D, Gravot A, Auroy P, Mazard C, Kraut A, Finazzi G, Grunwald D, Rappaport F, Vavasseur A, Joyard J, Richaud P, and Rolland N. HMA1, a new Cu-ATPase of the chloroplast envelope, is essential for growth under adverse light conditions. *J. Biol. Chem.* 281, 2882–2892 (2006).

[132] Morin I., Gudin S., Mintz E., and Cuillel M. Dissecting the role of the N-terminal metal-binding domains in activating the yeast copper ATPase in vivo. *FEBS J.* 276(16), 4483–4495 (2009).

[133] Tsivkovskii R, MacArthurs B., and Lutsenko S. The Lys(1010)-Lys(1325) fragment of the Wilson's disease protein binds nucleotides and interacts with the N-terminal domain of this protein in a copper-dependent manner. *J. Biol. Chem.* 276, 2234–2242 (2001).

[134] Dmitriev O., Tsivkovskii R., Abildgaard F., Morgan C.T., Markley J.L., and Lutsenko S. Solution structure of the N-domain of Wilson disease protein: distinct nucleotidebinding environment and effects of disease mutations. *Proc. Natl. Acad. Sci. U. S. A.* 103, 5302–5307 (2006).

[135] Petris M.J., Strausak D., and Mercer J.F. The Menkes copper transporter is required for the activation of tyrosinase. *Hum. Mol. Genet.* 9, 2845–2851 (2000).

[136] Steveson TC, Ciccotosto GD, Ma XM, Mueller GP, Mains RE, and Eipper BA. Menkes protein contributes to the function of peptidylglycine alpha-amidating monooxygenase. *Endocrinology* 144, 188–200 (2003).

[137] Tchaparian E.H., Uriu-Adams J.Y., Keen C.L., Mitchell A.E., and Rucker R.B. Lysyl oxidase and P-ATPase-7A expression during embryonic development in the rat. *Arch. Biochem. Biophys.* 379, 71–77 (2000).

[138] Qin Z., Itoh S., Jeney V., Ushio-Fukai M., and Fukai T. Essential role for the Menkes ATPase in activation of extracellular superoxide dismutase: implication for vascular oxidative stress. *FASEB J.* 20, 334–6 (2006).

[139] Terada K., Nakako T., Yang X.L., Iida M., Aiba N., Minamiya Y., Nakai M., Sakaki T., Miura N., and Sugiyama T. Restoration of holoceruloplasmin synthesis in LEC rat after infusion of recombinant adenovirus bearing WND cDNA. *J. Biol. Chem.* 273, 1815–1820 (1998).

[140] Ke B.X., Llanos R.M., Wright M., Deal Y., and Mercer J.F. Alteration of copper physiology in mice overexpressing the human Menkes protein ATP7A. *J. Physiol. Regul. Integr. Comp. Physiol.* 290, R1460–1467 (2006).

[141] Kaler S.G. Am. Diagnosis and therapy of Menkes syndrome, a genetic form of copper deficiency. *J. Clin. Nutr.* 67, 1029S–1034S (1998).

[142] Tao T.Y. and Gitlin J.D. Hepatic copper metabolism: insights from genetic disease. *Hepatology* 37, 1241–1247 (2003).

[143] Das S.K. and Ray K. Wilson's disease: an update. *Nat. Clin. Pract. Neurol.* 9, 482–493 (2006).

[144] Royce P.M. and Steinmann B. Markedly reduced activity of lysyl oxidase in skin and aorta from a patient with Menkes' disease showing unusually severe connective tissue manifestations. *Pediatr. Res.* 28, 137–141 (1990).

[145] Lutsenko S, LeShane ES, and Shinde U. Biochemical basis of regulation of human copper-transporting ATPases. *Arch. Biochem. Biophys.* 463, 134–48 (2007).

[146] DAmico F., Skarmoutsou E., Sanfilippo S., and Camakaris J. Menkes protein localization in rat parotid acinar cells. *J. Acta Histochem.* 107, 373–8 (2005).

[147] Kelleher S.L. and Lonnerdal B. Mammary gland copper transport is stimulated by prolactin through alterations in Ctr1 and Atp7A localization. *Am. J. Physiol. Regul. Integr. Comp. Physiol.* 291, R1181–1191 (2005).

[148] Niciu M.J., Ma X.M., E.l. Meskini R., Ronnett G.V., Mains R.E., and Eipper B.A. Developmental changes in the expression of ATP7A during a critical period in postnatal neurodevelopment. *Neuroscience* 139, 947–964 (2006).

[149] Tsivkovskii R., Eisses J.F., Kaplan J.H., and Lutsenko S. Functional properties of the copper-transporting ATPase ATP7B (the Wilson's disease protein) expressed in insect cells. *J. Biol. Chem.* 277, 976–983 (2002).

[150] Tsivkovskii R., Efremov R.G., and Lutsenko S. The role of the invariant His-1069 in folding and function of the Wilson's disease protein, the human copper-transporting ATPase ATP7B. *J. Biol. Chem.* 278, 13302–13308 (2003).

[151] Voskoboinik I., Mar J., Strausak D., and Camakaris J. The regulation of catalytic activity of the Menkes copper-translocating P-type ATPase. Role of high affinity copper-binding sites. *J. Biol. Chem.* 276, 28620–28627 (2001).

[152] Barnes N., Tsivkovskii R., Tsivkovskaia N., and Lutsenko S. The copper-transporting ATPases, Menkes and Wilson disease proteins, have distinct roles in adult and developing cerebellum. *J. Biol. Chem.* 280, 9640–9645 (2005).

[153] El Meskini R., Culotta V.C., Mains R.E., and Eipper B.A. Supplying copper to the cuproenzyme peptidylglycine alpha-amidating monooxygenase. *J. Biol. Chem.* 278, 12278–12284 (2003).

[154] Petris M.J., Voskoboinik I., Cater M., Smith K., Kim B.E., Llanos R.M., Strausak D., Camakaris J., and Mercer J.F. Copper-regulated trafficking of the Menkes disease copper ATPase is associated with formation of a phosphorylated catalytic intermediate. *J. Biol. Chem.* 277, 46736–46742 (2002).

[155] Cater M.A., La Fontaine S., and Mercer J.F. Copper binding to the N-terminal metal-binding sites or the CPC motif is not essential for copper-induced trafficking of the human Wilson protein (ATP7B). *Biochem J* 401, 143–153 (2007).

[156] Dmitriev O., Tsivkovskii R., Abildgaard F., Morgan C.T., Markley J.L., and Lutsenko S. Solution structure of the N-domain of Wilson disease protein: distinct nucleotide-binding environment and effects of disease mutations. *Proc. Natl. Acad. Sci. U. S. A.* 103, 5302–5307 (2006).

[157] Voskoboinik I., Mar J., and Camakaris J. Structure of the actuator domain from the Archaeoglobus fulgidus Cu(+)-ATPase. *Biochem Biophys Res Commun* 301, 488–494 (2003).

[158] Rice W.J., Kovalishin A., and Stokes D.L. Role of metal-binding domains of the copper pump from Archaeoglobus fulgidus. *Biochem Biophys Res Commun* 348, 124–131 (2006).

[159] Huster D. and Lutsenko S. The distinct roles of the N-terminal copper-binding sites in regulation of catalytic activity of the Wilson's disease protein. *J. Biol. Chem.* 278, 32212–32218 (2003).

[160] Coronado V., Nanji M., and Cox D.W. The Jackson toxic milk mouse as a model for copper loading. *Mamm Genome* 12, 793–795 (2001).

[161] Payne A.S. and Gitlin J.D. Functional expression of the Menkes disease protein reveals common biochemical mechanisms among the copper-transporting P-type ATPases. *J. Biol. Chem.* 273, 3765–3770 (1998).

[162] Achila D., Banci L., Bertini I., Bunce J., Ciofi-Baffoni S., and Huffman D. L. Structure of human Wilson protein domains 5 and 6 and their interplay with domain 4 and the copper chaperone HAH1 in copper uptake. *Proc. Natl. Acad. Sci. U. S. A.* 103(15), 5729–5734 (2006).

[163] Wu C.C., Rice W.J., and Stokes D.L. Structure of a copper pump suggests a regulatory role for its metal-binding domain. *Structure* 16(6), 976–985 (2008).

[164] Wu C.C., Gardarin A., Martel A., Mintz E., Guillain F., and Catty P. The cadmium transport sites of CadA, the Cd2+-ATPase from Listeria monocytogenes. *J. Biol. Chem.* 281, 29533–41 (2006).

[165] Lübben M., Portmann R., Kock G., Stoll R., Young M.M., and Solioz M. Structural model of the CopA copper ATPase of *Enterococcus hirae* based on chemical cross-linking. *Biometals* (2008).

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

[167] Frishman D. and Argos P. Knowledge-based protein secondary structure assignment. *Proteins: Struct. Funct. Gen.* 23, 566–579 (1995).

[168] Takahashi M., Kondou Y., and Toyoshima C. Interdomain communication in calcium pump as revealed in the crystal structures with transmembrane inhibitors. *Proc. Natl. Acad. Sci. U.S.A.* 104(14), 5800–5805 (2007).

[169] Okkeri J. and T. Haltia. The metal-binding sites of the zinc-transporting P-type ATPase of Escherichia coli. Lys693 and Asp714 in the seventh and eighth transmembrane segments of ZntA contribute to the coupling of metal binding and ATPase activity. *Biochimica et Biophysica Acta (BBA) - Bioenergetics* 1757, 1485–1495 (2006).

[170] Feller S.E. and MacKerell A.D.Jr. An improved empirical potential energy function for molecular simulations of phospholipids. *J. Phys. Chem. B* 104(31), 7510–7515 (2000).

[171] De Silva T.M, Veglia G, and Opella S.J. Solution Structure of reduced and Cu(I) bound forms of the first metal binding sequence of ATP7A associated with Menkes disease. *Proteins* 61(4), 1038–1049 (2005).

[172] Banci L., Bertini I., Del Conte R., D'Onofrio M., and Rosato A. Solution structure and backbone dynamics of the Cu(I) and apo forms of the second metal-binding domain of the Menkes protein ATP7A. *Biochemistry* 43, 3396–3403 (2004).

[173] Banci L., Bertini I., Cantini F., DellaMalva N., Herrmann T., Rosato A., and Wüthrich K. Solution structure and intermolecular interactions of the third metal-binding domain of ATP7A, the Menkes disease protein. *J. Biol. Chem.* 281(39), 29141–29147 (2006).

[174] Gitschier J, Moffat B, Reilly D, Wood WI, and Fairbrother WJ. Solution structure of the fourth metal-binding domain from the Menkes copper-transporting ATPase. *Nat. Struct. Biol.* 5, 47–54 (1998).

[175] Banci L., Bertini I., Ciofi-Baffoni S., Chasapis C.T., Hadjiliadis N., and Rosato A. An NMR study of the interaction between the human copper(I) chaperone and the second and fifth metal-binding domains of the Menkes protein. *FEBS J.* 272(3), 865–871 (2005).

[176] Banci L., Bertini I., Cantini F., Migliardi M., Rosato A., and Wang S. An atomic-level investigation of the disease-causing A629P mutant of the menkes protein, ATP7A. *J. Mol. Biol.* 352, 409–417 (2005).

[177] Jorgensen W.L., Chandrasekhar J., Madura J.D., Impey R.W., and Klein M.L. Comparison of simple potential functions for simulating liquid water. . *J. Chem. Phys.* 79, 926–935 (1983).

[178] Darden T., York D., and Pedersen L.G. Particle mesh ewald: An n-log(n) method for Ewald sums in large systems. *J. Chem. Phys.* 98(12), 10089–10092 (1993).

[179] Ryckaert J-P., Ciccotti G., and Berendsen H.J.C. Numerical integration of the cartesian equations of motion of a system with constraints: molecular dynamics of n-alkanes. *J. Comput. Phys.* 23(3), 327–341 (1977).

[180] Rodriguez-Granillo A., Crespo A., and Wittung-Stafshede P. Conformational dynamics of metal-binding domains in Wilson disease protein: molecular insights into selective copper transfer. *Biochemistry* 48(25)), 5849–5863 (2009).

[181] Feig M., Karanicolas J., and Brooks C.L. MMTSB tool set: enhanced sampling and multiscale modeling methods for applications in structural biology. *J. Mol. Graph. Model.* 22, 377–395 (2004).

[182] Humphrey W, Dalke A, and Schulten K. VMD - Visual Molecular Dynamics. *J. Molec. Graphics* 14, 33–38 (1996).

[183] Banci L, Bertini I, Cantini F, DellaMalva N, Herrmann T, Rosato A, and Wuthrich K. Solution structure and intermolecular interactions of the third metal-binding domain of ATP7A, the Menkes disease protein. *J. Biol. Chem.* 281, 29141–29147 (2006).

[184] DiDonato M., Hsu H.F., Narindrasorasak S., Que L.Jr., and Sarkar B. Copper-induced conformational changes in the N-terminal domain of the Wilson disease copper-transporting ATPase. *Biochemistry* 39, 1890–1896 (2000).

[185] Walker J.M., Huster D., Ralle M., Morgan C.T., Blackburn N.J., and Lutsenko S. The N-terminal metal-binding site 2 of the Wilson's Disease Protein plays a key role in the transfer of copper from Atox1. *J. Biol. Chem.* 279, 15376–15384 (2004).

[186] Cobine P.A., George G.N., Winzor D.J., Harrison M.D., Mogahaddas S., and Dameron C.T. Stoichiometry of complex formation between Copper(I) and the N-terminal domain of the Menkes protein. *Biochemistry* 39, 6857–6863 (2000).

[187] Guo Y., Nyasae L., Braiterman L.T., and Hubbard A.L. NH2-terminal signals in ATP7B Cu-ATPase mediate its Cu-dependent anterograde traffic in polarized hepatic cells. *Am. J. Physiol. Gastrointest. Liver Physiol.* 289, 904–916 (2005).

[188] Walker J.M., Tsivkovskii R., and Lutsenko S. Metallochaperone Atox1 transfers copper to the NH2-terminal domain of the Wilson's disease protein and regulates its catalytic activity. *J. Biol. Chem.* 277, 27953–27959 (2002).

[189] Wernimont A.K., Yatsunyk L.A., and Rosenzweig A.C. Binding of copper(I) by the Wilson disease protein and its copper chaperone. *J. Biol. Chem.* 279, 12269–12276. (2004).

[190] Morin I., Cuillel M., Lowe J., Crouzy S., Guillain F., and Mintz E. Cd^{2+} - or Hg^{2+} - binding proteins can replace the Cu^{+} - chaperone Atx1 in delivering Cu^{+} to the secretory pathway in yeast. *FEBS Lett.* 579, 1117–1123 (2005).

[191] Huffman D.L. and OHalloran T.V. Function, structure, and mechanism of intracellular copper trafficking proteins. *Annu Rev Biochem* 70, :677–701 (2001).

[192] Banci L., Bertini I., Cantini F., Della-Malva N., Migliardi M., and Rosato A. The different intermolecular interactions of the soluble copper-binding domains of the Menkes protein, ATP7A. *J. Biol. Chem.* 282(32), 23140–23146 (2007).

[193] Banci L., Bertini I., Cantini F., Chasapis C.T., Hadjiliadis N., and Rosato A. A NMR study of the interaction of a three-domain construct of ATP7A with copper(I) and copper(I)-HAH1: the interplay of domains. *J. Biol. Chem.* 280(46), 38259–38263 (2005).