University of Florence

International Doctorate in Structural Biology

Cycle XXII (2007 – 2009)

Title

Bioinformatics of metal binding proteins and genome wide analysis

Ph.D. thesis of

Shailesh Sharma

Tutor

Prof. Antonio Rosato

Coordinator

Prof. Claudio Luchinat

This thesis has been approved by the Universities of Florence , the University of Farnkfurt and the Utrecht University .

S.S.D. CHIM/03

I hereby declare that this PhD thesis is based on the information extracted from the cited references and computational experimental work, which I have performed at Magnetic Resonance Center (CERM), Florence .

December 31 2009 .....

Dedicated to my parents and to my teachers....

Acknowledgments

First and foremost, I am deeply indebted to my advisor Prof Antonio Rosato, whose excellent guidance, stimulating suggestions and encouragement helped me in all time of research .

I express my profound sense of gratitude to Prof. Ivano Bertini, Director, Magnetic Resonance Center (CERM) and Prof Girjesh Govil, Department of Chemical Sciences, Tata Institute of Fundamental Research (TIFR) for giving me an opportunity to conduct my doctoral research work in excellent research infrastructure , CERM .You both will always remembered for giving me exceptional ideas, constructive criticism and full support .

I take this opportunity to sincerely thank Dr. Gabriele Cavallaro , Dr. Anusarka Bhaumik , Dr. Ravi Sekhar Gadepalli , Dr. Rahul Jaiswal , Dr. Basir Ahmad and Dr. Ravi Krishnan Elangovan for there invaluable suggestions, help and hints .

I have furthermore all my colleagues and friends Maxime Melikian ,Chiara Massagni , Stefano Cacciatore , Valentina Borsi etc .

Life in Florence would have been less pleasurable and productive too without the company of my friends Paola , Anna , Isabella , Nicola , Chiara , Deepa , Malini , Vaishali , Rajesh , Hemram , Shashank , Vasantha and Soumyasri . I really enjoyed working with all CERMians and will ever cherish the time spent at CERM.

I express my thanks through this to the family of Mr .and Mrs. Mario and Lucia Mughini for their support and care during the most difficult time of my stay in Florence .

I am deeply indebted to my parents Mr. Tapesh Kumar Sharma and Mrs. Brijesh Sharma and my brother Mr. Rajat Sharma whose constant encouragement and support always inspired me to execute my scientific tasks.

My heartfelt thanks are extended to Mrs Lucia Mughini for providing me a homely environment , support , which definitely enhanced my scientific output. Your generosity will be ever remembered .

TABLE OF CONTENTS

List of Publications

(1) Sharma S, Rosato A. Role of the N-terminal tail of metal-transporting P(1B)-type ATPases from genome-wide analysis and molecular dynamics simulations, J. Chem. Inf. Model. 49:76-83, 2009

(2) Sharma S, Cavallaro G, Rosato A. A systematic investigation of multi-heme c-type cytochromes in prokaryotes ; J.Biol.Inorg.Chem (in press)

(3) Sharma S, Rosato A Benchmarking protocols for structure determination of proteins from chemical shift data (in preparation)

Name: Shailesh Sharma, M. Sc. ,Master 1st Level in Bioinformatics .

Office Address : Via L.Sacconi, 6 50019 Sesto Fiorentino (FI)-Tel:+39-055-4574245 Fax: +39-055-4574253 e-mail: haitoshailesh@gmail.com

Education:

2007 – 2009 **International Doctorate in Structural Biology** at the **Magnetic Resonance Centre (CERM),University of Florence.** This doctorate course is recognized by the Universities of Frankfurt (Germany) and Utrecht (Netherlands).

2005 – 2006 Master 1st Level in Bioinformatics
University of Torino.

2002 – 2004 M.Sc. in Bioinformatics
University of Allahabad.

1999 – 2002 B.Sc in Biology
M.D.S. University ,Ajmer.

1998 - 1999 All India Senior School Certificate Examination, 1999. The examination was conducted by Central Board of Secondary Education.

1995 – 1996 All India Secondary School Examination 1996. The examination was conducted by Central Board of Secondary Education.

Scholarship and Awards :

Selected for **Italian Government scholarship** in Biotechnology for the year 2005/2006 processed by Ministry H.R.D. , **Government of India** and completed First Level Master in Bioinformatics from University of Turin and Biotechnology Foundation of Turin, obtaining 102/110 marks.

Obtained highest marks and selected for fellowship from **EU** and University of Florence for Doctorate at the Magnetic Resonance Centre.

Selected for research as a project trainee at National High field NMR facility for the partial fulfillment of M. Sc degree. The project was titled “The application of molecular electrostatic potential in drug design” and was under the supervision of Prof. Girjesh Govil, Department of Chemical Sciences, **Tata Institute of Fundamental Research,** Homi -Bhabha Road, Colaba, Mumbai 400 005 India (govil@tifr.res.in).

Selected for the research project entitled “Mapping all Exon Microarray Probes over the human

Transcriptome" under the supervision of Prof. Raffaele Calogero, **Genomic - Bioinformatic Unit** ,Department of Clinical Science and Biology,Hospital San Luigi – University of Torino. This project was funded by Italian government and was in partial fulfilment of Master 1st Level in Bioinformatics degree.

Selected for 10 week Japan visit funded by National Institute of basic biology (NIBB) , Japan. This visit was for investigation and development of computational tools at Laboratory of Genome Informatics in NIBB and especially focused on microbial genomes . Fellowship was provided by NIBB to cover travelling , living expenses lodging at he institute's guest house and overseas travel insurance .

Original article:

***Sharma S**, Rosato A. Role of the N-terminal tail of metal-transporting P(1B)-type ATPases from genome-wide analysis and molecular dynamics simulations, J. Chem. Inf. Model. 49:76-83, 2009*

A systematic investigation of multi-heme c-type cytochromes in prokaryotes ***Shailesh Sharma** , Gabriele Cavallaro , Antonio Rosato ; J.Biol.Inorg.Chem*

***Sharma S** , Rosato A Benchmarking protocols for structural determination of proteins from chemical shift data .(in preparation)*

Current projects:

At present I am involved in two on going research projects at CERM: First is an approach which exploit metal binding patterns (MBP's) of metalloproteins to search genomes for new metalloproteins . This approach is applied to multi heme binding Cytochrome c proteins. CxxCH is a motif which is responsible for the binding of a single heme ligand by forming two covalent bonds with two S atoms present in two Cystine amino acids. Ensemble of sequences of the whole PDB is used to assess the potentiality and limits of the methods and to identify the level of confidence for the predictions output by the search. We identified 1629 multi heme binding Cytochrome C proteins in 594 distinct bacterial genomes.

Second is the protein structure calculation by NMR chemical shifts .It involves homology modeling based on the sequence-structure alignment, SPARTA , Cheshire ,CS - ROSETTA, cs23d and PSVS web servers complemented by locally written programs. We are focusing on different biological systems on ATPases and SCO proteins.

Workshops and Meetings in India:

Winner of "**Science Society Quiz Award 2004**", Post Graduate level, Awarded by Prof. G. K. Mehta ,Vice Ch., University of Allahabad.

Certified by **Department of Biochemical Engineering and Biotechnology, I.I.T., Delhi** for the training in Bioinformatics: Tools and applications by Prof. V. K. Srivastava, Dean Industrial Research Development (IRD) and By Prof. Saroj Mishra coordinator and Head of the Department of Biochemical Engineering and Biotechnology.

Certified by **I.I.T. Kharagpur** in its Continuing Education Program by participating in the short term course on Bioinformatics in Genomics And Proteomics on 24-25 Sept'04. The certificate is from Prof. S. C. Kundu and Prof. S. Dey Coordinators Department of Biotechnology and Bani Chatterjee Dean ,Continuing Education.

Certified by Bioinformatics center, **Birla Institute of Scientific Research, BISR, Jaipur** for the Training program in Bioinformatics: Genomics and Proteomics .

Paper presented in C.O.N.I.A.P.S. International Conference held in University of Allahabad in 2004. Topic of the paper was "Protein Structure Prediction."

International meetings and workshops:

1 A j**oint meeting of (I3) EU-NMR and (CA) NMR-LIFE** Florence, January 18-20, 2007
2 **Biobanks and biomedical research** (March 22, 2007)
3 A joint workshop of (I3) **EU-NMR and (CA) NMR-LIFE Florence**, April 10-12, 2007
4 **Symposium on: New Challenges in the Life Sciences** - Prioritizing European Research in Molecular Systems Biology (October 18-19, 2007)
5 11th Chianti **Workshop on Magnetic Resonance METHODS FOR BIOMOLECULAR MAGNETIC RESONANCE** Vallombrosa (Florence), Italy June 3 - 8, 2007
6 Magnetic Resonance in the Life Sciences: What's New NMR TO LAY THE BRICKS FOR MOLECULAR SYSTEMS BIOLOGY December 14th - 18th, 2008 - Montecatini Terme, Italy.
7 Structuring a Pan – European Bio – NMR Community June 15 – 18 , Sesto Fiorentino , Italy

Extra Curricular activities:

Certified in " SCAMPI farmers meet cum awareness program" organized by Marine products Export Development Authority, MPEDA in association with satellite Scampi Farming Resource Center on 22nd July 2003.

Certified on being successful in "Applied Fisheries Orientation Program" organized by Fish Farmer's Development Agency, FFDA and University of Allahabad on 4th Feb. 2003.

Having two fluency certificates in Italian language from both University of Turin and University of Florence .

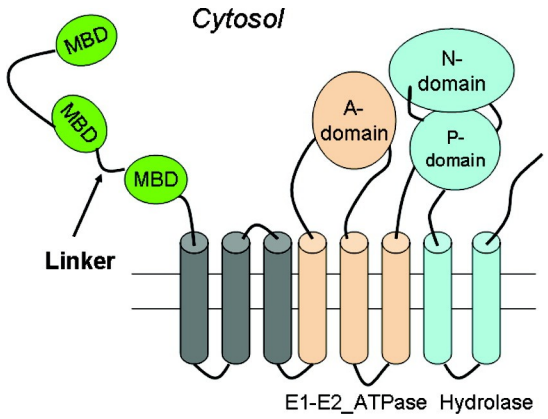
CHAPTER 1

Introduction :

Historically, the main focus of bioinformatics has been on computational analysis of biological macromolecules, i.e. proteins and nucleic acids. Advent of high-throughput sequencing methods provides bioinformaticians with more and more raw sequence data to analyze. Since proteins are biochemical entities, the lack of specific biochemical data results in the immense information gap between protein structure and function.. Computational protein structure analysis is one of the cornerstones of bioinformatics. It seems strange how little attention the bioinformatics community has paid to metalloproteins and other complex proteins. (To get an idea, try PubMed search with the combination 'bioinformatics' + 'biological inorganic' or 'computational' + 'bioinorganic'.) This is particularly striking considering the remarkable efforts and progress made in computational inorganic chemistry in the last few years .

In this PhD research work we aimed to further understand the biochemical properties of selected families of metalloproteins through bioinformatics . In the first project we studied copper-binding proteins . In eukaryotes, these copper-binding proteins are localized in various cellular compartments (such as the cytosol or the mitochondrion), or can be extracellular . In prokaryotic organisms, copper-binding proteins are mainly periplasmic, in Gram-negative organisms, or associated to the plasma membrane in Gram-positive organism. Copper is crucial for the correct functioning of cells, but it can also be potentially toxic. Copper toxicity is indeed at the basis of e.g. the use of this metal as a parasiticide in agriculture. These features made it necessary for living organisms to develop mechanisms that take care of copper uptake and transport to the appropriate sub cellular locations as well as of removal of excess intracellular copper (collectively called copper homeostasis). For the many prokaryotic organisms that do not use copper within their cellular processes, only copper removal is a relevant aspect. The regulation of copper homeostasis occurs principally at the transcriptional level in prokaryotes, whereas it is mainly dealt with at the post-translational level in mammals.

As a part of copper homeostasis mechanisms, one pathway that is particularly widespread involves the use of two protein partners, a soluble small (ca. 70 amino acids) copper(I) binding protein (called a metallochaperone) and an ATPase that can transport copper(I) ions across membranes at the expenses of ATP hydrolysis . Copper(I)-transporting ATPases are of the so-called P-type, i.e. they catalyze reactions proceeding through a covalent phosphorylated "P" intermediate . Based on their structural organization, and in particular on the number and position of transmembrane segments, P-type ATPases can be further separated into subgroups, with proteins of the P_1-subgroup being responsible for the transport of heavy metals, such as Cd^{2+}, Zn^{2+}, Pb^{2+}, Co^{2+}, Cu^{2+}, Ag^{+}, Cu^{+} [7].



Scheme of the architecture of P1B-type ATPases and correspondence to the Pfam domains used for complete proteome analysis. An ATPase with three metal-binding domains (MBDs) is shown. Each MBD corresponds to a single HMA Pfam domain (green). The transmembrane and cytosolic regions map to two different Pfam domains: the Pfam domain E1-E2_ATPase (light orange) contains three transmembrane helices and the Actuator (A-) soluble domain; the Pfam domain Hydrolase (light blue) contains the ATP-binding site (in the N-domain) and the phosphorylation site (in the P-domain) as well as the two most C-terminal helices. The gray transmembrane helices do not belong to a Pfam domain, and their position is different in different ATPase classes. The regions corresponding to the E1-E2_ATPase and Hydrolase Pfam domains are present in all P-type ATPases. The linker region on which the present study focuses is highlighted by an arrow. Relative region sizes are not to scale.

The H+,K+ ATPase is instead an example of P2-type ATPase. Phylogenetic analyses have shown that the P1-subgroup encompasses also some relatively uncommon bacterial ATPases that feature an organization in multiple protein subunits and are involved in potassium transport . In this work, we investigated the occurrence and properties of the P1B-ATPases and, partly, of their partner metallochaperones. We found that the latter proteins are typically encoded in organisms containing also ATPases of the subtypes 1B-1 or 1B-2. These subtypes have a characteristically extended N-terminal cytoplasmic tail that contains multiple metal-binding domains (MBDs), which can receive the metal ion from the metallochaperone. Therefore, we investigated the impact of the linker region connecting two of the cytoplasmatic metal-binding domains on their reciprocal dynamics and possible interaction with other the domains of the enzyme. For ATPases containing three or more MBDs, the two MBDs closest to the transmembrane part of the enzyme were focused upon, as an extensive body of literature indicates that they play a different biochemical role than the others. We observed a significant

variability in the number and spacing in sequence of the MBDs. On the basis of molecular dynamics simulations, we proposed that the MBDs could be quite free to reorient with respect to one another. The relative conformational freedom increased rapidly with the length of the linker between the MBDs. Also based on available experimental studies, these data suggested that the reciprocal mobility of MBDs is instrumental to permit the tuning of the selectivity and/or affinity of the ATPase for the substrate as well as to modulate the enzymatic activity of the system. We additionally detected a small but significant number of instances in which a metallochaperone is likely to interact directly with the transmembrane domain of P-type ATPases lacking cytoplasmic MBDs .

My second project was on c – type cytochromes who are ubiquitous in nearly all living organisms, where they play vital roles in mediating electron transfer (ET) reactions associated with respiration. Although their amino acid sequences differ greatly, all c-Cyts possess at least one haem that is covalently bound through amino acid side-chains of the proteins to position and orient the haem moiety and thereby facilitate efficient reactions. The haem moieties are commonly co-ordinated through two thioester bonds to proximal cysteines in the protein, where the signature motif of most c-Cyts is CX2CH (other common motifs include CX3−4CH, CX2CK and A/FX2CH). These motifs with covalently bound haems are the key components used to constitute the haem-containing domains whose diverse functions range from binding of O2 and catalysis to electron transfer and accumulation . c-Cyts have been extensively investigated, and several excellent reviews have been dedicated to the structures, chemistry and biogenesis of c-Cyts . This review focuses on the unique features of bacterial c-Cyts with multiple haems and their roles in bacteria-mediated dissimilatory reduction of solid metal (hydr)oxides . The c-Cyts are essential for the versatile anaerobic respiration capabilities . c-Cyt maturation system are unable to produce functional c-Cyts and consequently fail to grow when fumarate, dimethyl sulphoxide (DMSO) or trimethylamine N-oxide (TMAO) is used as the terminal electron acceptor . Genome sequence analysis has also revealed that most of these c-Cyts polypeptides found in DMRB possess more than one CX2CH motif, and that one of these putative c-Cyts in G. sulfurreducens has as many as 27 CX2CH motifs, in sharp contrast to the c-Cyts found in eukaryotes, which typically contain only one haem . Some multihaem c-Cyts found in DMRB are located in the outer membrane, where they are positioned to interact with extracellular substrates, whereas most membrane c-Cyts found in other bacteria, including multihaem c-Cyts in sulphate respiring bacteria such as Desulfovibrio, are associated with the cytoplasmic or inner membranes .

Although their overall three-dimensional (3-D) structures vary considerably, one of the unique features found in most bacterial multihaem c-Cyts whose 3-D structures have been solved is the arrangement of haem groups. In these multihaem c-Cyts, all haem groups are positioned in such way that

each is in close proximity to at least one of the other haems, and the porphyrin rings of two adjacent haems are positioned either parallel or perpendicular to each other. These arrangements are thought to facilitate rapid ET with considerable specificity among the haem groups that form a continuous ‘electric wire’ . When protein complexes are formed among multihaem c-Cyts, at least one haem group in one c-Cyt subunit is usually positioned close to a haem group in another c-Cyt subunit, again permitting rapid and specific inter-ET between the proximal subunits . Formation of protein complexes among multihaem c-Cyts and the close arrangement of haem groups within and between multihaem c-Cyts make it possible to transfer electrons rapidly over relatively long distances. The c-Cyts quinol dehydrogenase (NrfH)/nitrite reductase (NrfA) complex of Desulfovibrio vulgaris consists of two NrfHs and four NrfAs with a total of 28 haems that are used to form the entire ET network of the NrfH/NrfA complex. The longest distance that electrons could flow from the haem possibly used for quinol oxidation in one of the NrfH subunits to the haem for nitrite reduction in an NrfA subunit along the haem network (centre-to-centre) is ~98 Å (or 9.8 nm), in which 10 haems are involved .

It has long been recognized that the 3D structure of a protein is directly related to its amino acid sequence . *De novo* structure predictions from solely the sequence thus provide another pathway to generate protein structural models. Among those, ROSETTA is one of the most successful programs for obtaining atomic level 3D structures of small proteins. For each small segment of the query protein, ROSETTA selects two hundred fragments from the crystallographic structural database that are similar in amino acid sequence and hence representative of the conformations the peptide segment is likely to sample during folding. A Monte Carlo based assembly process then uses these fragments to search for compact, low energy folds. The ROSETTA full atom refinement protocol, which employs Monte Carlo minimization coupled with a detailed all-atom force field, is then used to search for low energy structures with close complementary side chain packing in the vicinity of the starting model . Adding the structural information contained in experimentally determined NMR chemical shifts holds promise to greatly improve the structural accuracy of selected fragments, and thereby to improve ROSETTA performance without any significant change in the basic structure or functioning of this well established program.

My third project was on protein structure determination . The important practical result achieved by the present work is that researchers in the field of protein structure determination may take advantage of sequence analysis , molecular dynamics and all computationally analysed data produced by us . The strategies we have developed allow one to obtain optimum conditions for the concerted use of predictions of different nature in calculations, as well as to analyse and compare the properties and the impact of each class of data. The recent gain in popularity of structural information suggests that

the number of users of these tools should become larger and larger in the next future.

The development of detailed protocols to be used in calculations also relates to the efforts which are currently aimed at increasing the degree of automation within NMR protein structure determination. The establishment of computational methods to replace the conventional manual approaches is a major future challenge: this is driven by the need of speeding up the progress of structural genomics projects, which are intended to provide structural information on a genome-wide scale. The achievement of automated structure determination through NMR demands the methods for the assessment of structure quality to be likewise improved, in order to ensure that the reliability and the robustness of the conventional procedure is not compromised.

Even more than NMR structure determination of individual proteins, the methodology for the structural characterization of protein-protein and protein-nucleic acid adducts, as well as of protein-ligand complexes, demands faster protocols to be established. Therefore, the development of suitable approaches to this task, which deals with a huge variety of potentially interacting systems, stands out as an active field of research.

CHAPTER 2

The role of the N-terminal tail of metal-transporting P_{1B}-type ATPases from genome-wide analysis and molecular dynamics simulations

Introduction :

For many organisms copper is an essential metal, because of its role as a cofactor in a variety of enzymes and electron carriers [1]. In eukaryotes, these copper-binding proteins are localized in various cellular compartments (such as the cytosol or the mitochondrion), or can be extracellular [1,2]. In prokaryotic organisms, copper-binding proteins are mainly periplasmic, in Gram-negative organisms, or associated to the plasma membrane [1,2]. A notable exception is observed in the case of photosynthetic prokaryotes that contain copper proteins in thylakoids [3]. Notwithstanding its crucial role for the correct functioning of cells, copper can be potentially toxic *in vivo* [1]. Copper toxicity is indeed at the basis of e.g. the use of this metal as a parasiticide in agriculture. These features made it necessary for living organisms to develop mechanisms that take care of copper uptake and transport to the appropriate sub cellular locations as well as of removal of excess intracellular copper [4]. For the many prokaryotic organisms that do not use copper within their cellular processes, only copper removal is a relevant aspect [2].

Among the various biochemical solutions occurring in Nature to address the above-mentioned needs, one that is particularly widespread involves the use of two protein partners, a soluble small (ca. 70 amino acids) copper(I) binding protein (called a metallochaperone) and an ATPase that can transport copper(I) ions across membranes at the expenses of ATP hydrolysis [4-6]. Copper(I)-transporting ATPases are of the so-called P-type, i.e. they catalyze reactions proceeding through a covalent phosphorylated "P" intermediate [6]. Based on their structural organization, and in particular on the number and position of transmembrane segments, P-type ATPases can be further separated into subgroups, with proteins of the P_1-subgroup being responsible for the transport of heavy metals, such as Cd^{2+}, Zn^{2+}, Pb^{2+}, Co^{2+}, Cu^{2+}, Ag^{+}, Cu^{+} [7]. The H^{+},K^{+} ATPase is instead an example of P_2-type ATPase. Phylogenetic analyses have shown that the P_1-subgroup encompasses also some relatively uncommon bacterial ATPases that feature an organization in multiple protein subunits and are involved in potassium transport [8]. The latter form the so-called P_{1A} sub-subgroup, whereas the ATPases transporting heavy metals are dubbed P_{1B} [9].

The aforementioned combination of a small, soluble copper(I)-transporter, operating in the cell cytosol, and of an enzyme that actively catalyses the translocation of the metal ions allows cells to remove copper(I) ions from the cytosol and either pump them outside the cell or into intracellular organelles, depending on the localization of the ATPase [6,10,11].

All P-type ATPases share a basic "core" architecture [8], comprising a hydrophilic region protruding into the cytosol, which contains the phosphorylation and ATP-binding sites and a smaller cytosolic region (sometimes called the A-domain), which has a regulatory function and is required for the phosphatase step of the catalytic cycle (dephosphorylation of the intermediate formed during ATP hydrolysis [12]). These cytosolic parts of the polypeptide chain are connected by a number of transmembrane helices, which are involved in the formation of an intramembranous channel, and whose organization, as mentioned, leads to the definition of P_1 and P_2 subgroups [7]. In addition to the core structure, the distinguishing feature of P_{1B}-type copper(I)-transporting ATPases, which are the focus of this work, is their long N-terminal tail, which contains a variable (between one and six) number of 70-aa independently folded domains [5]. Each domain harbours a conserved sequence motif CXXC, through which it can bind one equivalent of copper(I) [13]. The motif is often preceded by a methionine in position –2 (i.e. MXCXXC), which however is not involved in copper(I) coordination. In humans, there are two relevant ATPases, namely ATP7A and ATP7B, also known as the Menkes (MNK) and Wilson (WND) disease proteins, respectively. Many studies are available for these two systems that demonstrate that, even *in vivo*, the presence of either the intact fifth or intact sixth metal-binding domain (i.e. the two closest to the transmembrane domain) is sufficient to support the activity of the protein, including intracellular trafficking, at levels normal or close to normal [14-18].

The stretches of aminoacidic sequence linking the folded domains in the N-terminal tail are poorly structured [19,20]. Notably, the length of such linker regions is very variable [5], both for linkers connecting different domains within the same protein or for linkers between corresponding domains in different proteins, and ranges from three to several tens of amino acids. In the systems for which detailed experimental studies on the structure at the atomic level of the N-terminal tail are available [20-24], the last two domains are connected by a relatively small number of amino acids (less than ten, to be compared to a few tens of amino acids linking the most N-terminal domains). The features of the flexible N-terminal tail and its structural plasticity are important for the understanding of the overall functioning of P_{1B} ATPases. Indeed, there is a substantial body of evidence that the interaction between the various regions of the enzyme as well as with the metallochaperone affect significantly its activity and, for the mammalian enzymes, the balance between onward and backward protein trafficking [11,25]. A possible role of the N-terminal tail could be that of modulating the ATPase activity through metal-

dependent interactions with the ATP-binding domain and the A domain of the enzyme [26].

In the present work, we aimed at furthering our understanding of the role of the linker region through molecular dynamics simulations of systems having different linker length and through a bioinformatic analysis of the spacing occurring between corresponding domains in various ATPases with multidomain cytoplasmic tails.

Methods

Sequence analysis

We used the SMART (http://smart.embl-heidelberg.de/) database [27] to identify proteins having a domain architecture similar to that of the yeast copper-transporting ATPase Ccc2, i.e. containing at least two soluble metal-binding domains in addition to all other domains characteristic of P_{1B}-type ATPases. Incomplete (i.e. containing only protein fragments) sequences were discarded. The SMART database contains protein sequences from Swiss-Prot and spTrembl databases as well as Ensembl proteomes [27]. Archaeal, bacterial, and eukaryotic organisms were investigated. We retrieved more than 400 proteins and extracted from the latter the sequences of the two metal-binding domains closest to the transmembrane region. These sequences were aligned to check the conservation of the metal-binding CXXC motif. The proteins in which one or both of the two domains lacked the motif were removed. Then we separated the domain pairs on the basis of the length of the polypeptide region linking the two domains by a locally written program, which exploited the definition of domain boundaries of SMART. We selected a few different representative systems and performed molecular dynamics simulations on their apo and holo-forms: *Bacillus subtilis*; the human Menkes's disease protein (MNK); the human Wilson's disease protein (WND); *Deinococcus geothermalis* and *Brucella abortus* domains whose interdomain linker regions comprise three, seven, seven, eleven and thirty five amino acids respectively.

Among these five, only for *B. subtilis* and WND the structure of the two-domain construct is in the PDB (entries 1P6T [23] and 2EW9 [22], respectively). Structural models for the other sequences were thus built using the program Modeller with standard parameters on the basis of these two available structures. Because of the choice of the template structures, in all MD calculations the two domains were close in space at the beginning of the simulation. Several models were generated for each sequence, and each of them was then visually inspected. Models without apparent defects were ranked on the basis of their stereochemical quality and energy; the best model was used as input for molecular dynamics simulations. For each sequence, only the model with the best stability in the first hundreds of

picoseconds of the trajectory (after equilibration) was retained.

We used the Pfam [28] domains HMA, E1_E2_ATPase and Hydrolase, which are contained in P-type ATPases and metallochaperones (HMA only), to investigate the occurrence of these proteins in completely sequenced prokaryotic proteomes. Note that the E1-E2 ATPase domain is specific of P-type ATPases,whereas the Hydrolase domain is not specific but is required for phosphatase activity. The program HMMER [29] was used for this purpose, with a E-value threshold of 10^{-5} (i.e. only domains with an E-value better than 10^{-5} were retained for analysis). Proteome sequences were retrieved from the Ref_Seq database [30]. We used this approach, which is similar to what done in other studies by our and other laboratories [31-34], to obtain a more detailed view in a dataset of protein sequences more restricted than SMART. The present dataset however had the advantage of including only complete proteomes and therefore allowed us to perform meaningful comparisons among the ATPase and metallochaperone content of different organisms.

Molecular Dynamic Simulations (MD)

AMBER 8.0 [35] was used to make individual simulations both in apo and in holo forms. Holo forms were built from the corresponding apo forms by adding a copper(I) ion in between the two S atoms of the cysteines of the motif CXXC. To do so, the side chains of the cysteins were properly pre-oriented by restraining the distance between the S atoms and minimizing the structure of the apo-protein. After insertion of the copper(I) ion, the S-Cu-S angle was loosely restrained at 180 degrees for a short time during the MD. Parameters for the metal site were taken from [36,37]. The SHAKE algorithm was used to maintain bond lengths fixed, permitting the use of a time-step of 1.5 fs. The protein was solvated using TIP3P water and a ten Å buffering distance between the edges of the box and the protein. Initially, we minimized the energy of each system in two stages. In the first stage, we minimized the water molecules while holding the protein and counterions fixed, in order to relax solute-solvent contacts. In the second stage, we minimized the complete system. MD were then performed using periodic boundary conditions, at constant pressure (1 atm) and temperature (298 K) for six nanoseconds. For each trajectory, we analyzed a portion of four nanoseconds after the system had equilibrated. RMS versus time graphs are shown in Figure S1.

Results

Separation of MBDs and its effect on the system properties

We retrieved the sequences of P_{1B}-type ATPases that contained at least two metal-binding domains (MBDs) from both the SMART database, which contains sequences from organisms in all domains of Life, and completely sequenced prokaryotic proteomes, in the latter case using the releant

HMM's provided by the Pfam database. This resulted in, respectively, 320 and 290 sequences. The length of the linker between the two MBDs closest to the trans-membrane region was determined using the domain definitions of either the SMART or the Pfam database. A summary of the data is given in Table XX. As shown in Figure 1, the computed linker lengths featured a significant variability. The most likely separation between the two MBDs taken into account was three aminoacids, which was equal or very close to the first quartile of the distribution (Table 1). Nevertheless, the third quartile was t 16-17 amino acids, implying that one quarter of the sequences (i.e. 70-80) in the ensembles examined had linker regions with a length exceeding this value. The computed separation did not depend on the total number of domains in the ATPase (not shown).

Table 1

	Pfam	SMART
Mean	15 ± 22	16 ± 24
First quartile	4	3
Third quartile	17	16
95th percentile	49	70
Minimum value	1	3
Maximum value	144	179

We then verified whether, independently of their separation, two consecutive MBDs formed a tight unit thanks to energetically favorable inter-domain interactions. To evaluate this hypothesis we built structural models of two-MBD units with different linker lengths (Figure XX), and subjected them to molecular dynamics simulations. During MD, individual MBDs remained stable with backbone rms deviations, after equilibration, in the range 1-2 Å. On the other hand, when considering their relative position, we observed larger rearrangements with respect to the initial orientation the longer the linker length . These rearrangements did not depend on the modeling procedure, as the MD trajectories starting from the model structure of the MNK protein feature RMSD values as large as (for the apo-protein) or even smaller than (for the copper-protein) those observed for the simulation of the WLN protein, which has the same linker length of MNK and started from an experimental structure . We evaluated the energies of inter-domain interaction along the various trajectories and computed for each system the average energy of the apo- and copper-protein in the 4 nanosecond production trajectory. The results are shown in Figure 2XX as a function of the separation in sequence of the two MBDs. It can be readily observed that the average inter-domain interaction energy falls sharply with increasing linker length.

Distribution of metallochaperones and ATPases in prokaryotic organisms

In this work, we investigated selected members of a sub-class of P-type ATPases, namely

copper(I)-transporting P_{1B}-type ATPases. These ATPases catalyze the transport of copper(I) ions across biomembranes at the expenses of ATP hydrolysis. The catalyitic cycle involves the formation of a covalent phosphorylated "P" intermediate, hence the label "P-type" [6]. P_1-type ATPases are characterized by a common organization of their transmembrane segments that is distinct from that of P_2-type ATPases [7]; and the P_{1B} sub-type is in particular responsible for the transport of heavy metals (Cd^{2+}, Zn^{2+}, Pb^{2+}, Co^{2+}, Cu^{2+}, Ag^{+}, Cu^{+}) [8,9]. The entire group of P_{1B}-type ATPases can be further split in various subgroups, which have specific structural and sequence features throughout the entire polypeptide that are linked to their metal specificity [38,39]. The presence of (most frequently) N-terminal, cytoplasmic MBDs typically harboring a CXXC metal binding pattern is common to P_{1B}-type ATPases transporting Cu^{+} (subgroup 1B-1), Zn^{2+}, Cd^{2+}, Pb^{2+} (subgroup 1B-2) [38-40]. The MBDs themselves contain sequence features that help discriminating the above metals [41]. ATPases containing multiple MBDs normally transport copper(I) and exceptionally transport zinc(II) [39].

It is known that copper(I)-transporting ATPases from eukaryotes tend to have two or more MBDs [5]. We investigated how common these are in prokaryotic systems, by scanning the complete proteome sequences of as many as 594 organisms using the Pfam HMM's representing the HMA (corresponding to the soluble MBD), Hydrolase and E1_E2_ATPase domains. P-type ATPases were identified by the simultaneous presence of these two domains (which define their common basic "core" architecture [8]); proteins were instead classified as P_{1B-1} or P_{1B-2} if they additionally contained one or more HMA domains. At the present level of investigation, the P_{1B-1} and P_{1B-2} sub-types cannot be distinguished, therefore from now on we will refer to the ATpases of P_{1B-1} or P_{1B-2} type as $P_{1B-1,2}$-type. Hereafter, the figures referring to P-type ATPases exclude $P_{1B-1,2}$-type ATPases. The results obtained are summarized in Table 2.

65% of the prokaryotic organisms analyzed simultaneously contained one or more $P_{1B-1,2}$-type ATPases and one or more other P-type ATPases. In addition, 82 (14%) organisms encoded only $P_{1B-1,2}$-type ATPases, yielding a total of 468 (79%) organisms that encoded at least one $P_{1B-1,2}$-type ATPase. In all the organisms analyzed, we detected only three instances of proteins containing the E1-E2 ATPase but not the Hydrolase domain, which is not specific of P-type ATPases but is nevertheless required for function. When counting individual proteins, we detected as many as 826 $P_{1B-1,2}$-type ATPases. Most of these organisms encoded one or two $P_{1B-1,2}$ ATPases; three was also relatively common. *Haloarcula marismortui* encoded nine different such ATPases, which is probably related to its halophilic lifestyle. Bacterial $P_{1B-1,2}$-type ATPases had between one and four MBDs (Table 3). 14 organisms encoded a single ATPase with four MBDs, with the separation between the two MBDs closest to the transmbrane domain ranging between less than 10 and 30 amino acids. In *Ralstonia metallidurans* and *P.*

lavamentivorans, the four MBDs did not contain the canonical CXXC metal-binding pattern, which was replaced by CXXEE. For *R. metallidurans* this was implied with the substrate being lead(II) rather than copper(I), and the gene was called PbrA [42]. In the proteins with three MBDs, all of them had the canonical CXXC pattern and featured a quite variable spacing. For example, in the ATPases from *Yersinia* species the first and second MBDs were closely spaced whereas the third MBD was relatively distant (in sequence) from the second, as well as from the transmembrane domain. Finally, 244 bacterial $P_{1B-1,2}$-type ATPases had two MBDs. *R. metallidurans* had two such ATPases, one with the canonical patterns and the other with the CXXEE patterns in the both MBDs. Proteins similar to the latter, i.e. having two MBDs with CXXEE pattern, are found in a variety of organisms (e.g. *Klebsiella pneumoniae*, *Shewanella frigidimarina*, etc.). Instances of proteins in which the two MBDs have different patterns also exist, not only in *R. metallidurans* [42], but also in distant organisms such as *Aeropyrum pernix*.

We analyzed also the distribution of the proteic partners of copper(I)-transporting ATPses, i.e. metallochaperones. We detected putative metallochaperones, assigned as such on the basis of their sequence containing only MBDs and no other known domains, in 335 organisms (Table 2). Of these, only 20 did not contain any $P_{1B-1,2}$-type ATPase. On the other hand, the majority of the organisms lacking a metallochaperone encoded one or more $P_{1B-1,2}$-type ATPases. Of the metallochaperones potentially without a partner $P_{1B-1,2}$-type ATPase, 10 were from as many *Lactobacilli* species and were all adjacent in the genome to a P-type ATPase lacking any MBD. The same was true for *Leuconostoc mesenteroides*. In *Thermoplasma acidophilum* and *Thermoplasma volcanium* the only metallochaperone encoded was instead next to a mercuric reductase (MerA), and thus presumably had Hg^{2+} as its target (i.e. was part of an operon containing the *merA* and *merP* genes [43]; MerP features a single HMA domain). In *Thermofilum pendens* the gene encoding the metallochaperone was next to one encoding a protein containing a rubrerythrin domain, which binds iron. Other instances of organism containing a metallochaperone but not $P_{1B-1,2}$-type ATPase appeared as sequencing errors, where the ATPase sequence was interrupted or abnormal in some way.

Discussion

P_{1B}-type ATPases of subgroups 1 and 2, which transport respectively Cu^+ or Zn^{2+}, Cd^{2+}, Pb^{2+} ion, contain between one and six MBD's [5]. Prokaryotic ATPases contain up to four MBD's (Table 3). The separation in protein sequence between the two cytoplasmic metal-binding domains (MBDs) is highly variable (Figure 1). In structurally characterized two-domain systems (*B. subtilis* CopA [23,24] and the human protein WND [22]), it has been proposed that the two MBDs that are closest to the transmembrane region domains form relatively tight units, possibly owing to their short linkers. In more complex

multi-domain constructs, such as a three-domain construct from the MNK protein, the two terminal MBDs (domains 5 and 6) are more rigidly connected to one another than domain 5 is connected to the preceding MBD (domain 4) [20]. A study of the entire N-terminal tail of MNK is also available, where it is readily observed that within any pair of consecutive domains besides 5 and 6 the two MBDs are nearly completely free to reorient with respect to one another [21].

The presence of a relatively short linker region between the two MBDs closest to the trans-membrane part of the enzyme was proposed to be instrumental to maintain a fixed relative orientation of their respective metal-binding sites [21]. This would avoid interactions between them in the presence of the metal substrate. In *Anabaena* AztA the interaction between a pair of MBDs can result in enhanced selectivity in a zinc(II)-transport system, where other divalent cations such as lead(II) or cadmium(II) form a stable metal-bridged intermediate involving the two MBDs that inhibits the ATPase [44]. Within this frame it is notable that a small number of copper-transporting ATPases feature one MBD at the N-terminal part of the trans-membrane region and one at the C-terminal part, which are therefore truly independent units. These are the ATPases from *Archeoglobus fulgidus* [45], *Bacteroides fragilis* and *Troponema denticola*. A model of the structure of the *A. fulgidus* ATPase is available [26], where however the C-terminal MBD has been removed and thus no information on its possible interaction with the N-terminal MBD.

In principle, the length of the linker separating two consecutive MBDs is not sufficient to establish whether they form a tight unit or not. Indeed, there could exist inter-domain interactions that are sufficiently stable (from the thermodynamic point of view) to hold the two together regardless of the length of the linker. However, MD simulations showed that the two domains display increasing freedom of relative reorientation with increasing separation in sequence and, in parallel, less significant energies of inter-domain interaction (Figure 2). Therefore, the linker effectively uncouples the two MBDs. These conclusions applied to both the apo and holo forms, possibly to a larger extent in the latter than in the former even though metallation did not trigger significant changes within a single MBD, besides the region of the metal-binding site [46]. Indeed, the regions of interdomain contact at the surfaces of the two MBDs appear in general to be poorly optimized for interaction. For example, Figure 3 shows the electrostatic potential at the interfaces of the two MBDs of the WND protein, based on the experimental 2EW9 structure: it can be noted that regions with similar electrostatic potential are in proximity (at the top and in the center of the interfacial regions).

In agreement with the above MD data, experimental NMR relaxation data for the two-domain systems from *B. subtilis* CopA and from WND show that the linker region is essentially as rigid as the rest of the protein in the former [23] but more mobile in the latter [22]. This confirms that even the short

increase in linker length from the CopA to the WND protein (from three to five residues) is sufficient to appreciably increase the relative freedom of the two MBDs. Analogously, relaxation data for the three-domain construct of MNK already mentioned, which contains the fourth, fifth and sixth MBDs, demonstrate the flexibility of all linker regions [20]. The linker between the fourth and fifth MBDs is nearly 50 amino acids in length and displays significantly higher flexibility than the linker between the fifth and sixth MBDs, which comprises only seven amino acids. The order parameters for bond vectors in the linker residues extracted from our simulations fell sharply from around 0.9 to around 0.5 when going from the shortest to the longest linker. In summary, the sequence analysis data (Table 1 and Figure 1) and the MD simulations concur to establish that even the MBD pairs closest to the transmembrane domain do not need to have a fixed orientation with respect to one another. Relatively short linker lengths already permit an appreciable degree of relative conformational freedom; systems where the linker length is sufficient to allow the two MBDs to bring their metal-binding sites at short distance are fairly common (Table 1).

As mentioned, P_{1B}-type ATPases are enzymes that transport heavy metals across biomembranes play a key role in the homeostasis and the mechanisms of biotolerance of these metals [38-40]. Here we focused on $P_{1B-1,2}$-type ATPases , which feature N-terminal, cytoplasmic MBDs. ATPases containing multiple MBDs normally transport copper(I) or, less commonly, divalent cations such as zinc(II) [39]. The common metal-binding pattern for these systems is CXXC; we also detected proteins with CXXEE patterns that are presumably involved in the transport of divalent cations, such as lead(II). In some instances, proteins with two MBDs had a different metal-binding pattern in each of them. These systems were common to phylogenetically distant organisms. Notably, extensive horizontal gene transfer of $P_{1B-1,2}$-type ATPases has been recently reported among bacteria isolated from subsurface soils contaminated by metals, involving not only proteobacteria but also actinobacteria and firmicutes [47].

We carried out extensive investigation of the distribution of P-type ATPases and of their partner metallochaperones in prokaryotes. Metallochaperones always occurred as single-MBD proteins, with the five detected exceptions appearing again as sequencing errors, introducing breaks within $P_{1B-1,2}$-type ATPase sequences. They are known to be involved in the transport of either copper(I) or mercury(II) [43]. In some organisms it is actually possible to find both copper(I) chaperones and mercury(II) chaperones, having respectively a $P_{1B-1,2}$-type ATPase and a protein of the *mer* operon as their partners. Notably, MerA contains a MBD domain as well, which is the domain for interaction with the Hg^{2+} chaperone [43]. In *Streptococcus gordonii* we identified an operon with a MerA containing two MBD domains and a MerP homologue. Metallochaperones were detected only in proteomes encoding also at least one $P_{1B-1,2}$-

type ATPase, with only 22 exceptions (Table 3). These included 10 *Lactobacilli*, where the metallochaperone was in the same operon of a P-type ATPase lacking any MBD. However, in 75 organisms we detected a larger number of metallochaperones than $P_{1B-1,2}$-type ATPases. In fact, we observed the occurrence of metallochaperones in the same operon of P-type ATPases not belonging to the 1B-1 or 1B-2 subgroups also in the proteome of organisms that did encode $P_{1B-1,2}$-type ATPases. This finding suggests that the ATPase does not strictly need to receive its metal substrate through a cytosolic domain first. Indeed, for CopA from *A. fulgidus* it has been recently suggested that the metallochaperone may interact directly with the transmembrane site, without an intermediate step involving the MBDs of the ATPase [45]. In the proposed mechanism of function of the SERCA Ca^{2+}-ATPase, which is an archetype for P-type ATPases, calcium(II) is directly bound to the trans-membrane site [48]. In other instances, metallochaperones had as their protein partner a MerA homologue. Finally, the observation that a large number of organisms encode $P_{1B-1,2}$-type ATPase sequences but not metallochaperones (Table 3) strongly suggests that these enzymes can sequester their metal substrate directly in the cytoplasmic space, either from a small-molecule complex or from an unidentified metallochaperone.

Conclusion

We investigated the occurrence and properties of $P_{1B-1,2}$-type ATPases and, partly, of their partner metallochaperones. $P_{1B-1,2}$-type ATPases may contain multiple MBDs in the N-terminal cytoplasmic part of their sequence. This, together with the observation that the number of these domains tends to be higher in more complex organisms, has stimulated numerous studies of their overall properties, function and specialization. In particular, it has been shown that in the human MNK and WND proteins, which contain six MBDs, the two domains closest in sequence to the transmembrane part of the protein have a different role than the other four [10]. Here we demonstrated through sequence analysis across a large dataset of organisms and molecular dynamics that in the majority of organisms, these two domains are structurally independent and can reorient one with respect to the other. Therefore, it appears very likely that they can exert their function independently. The MBDs have a role in regulating the ATPase activity through interactions with the other protein domains [26]. The presently observed relative flexibility can thus be instrumental in optimizing the regulation of the activity in multi-MBD ATPases.

Metallochaperones are single-MBD proteins that typically deliver copper(I) ions to partner $P_{1B-1,2}$-type ATPases or mercury(II) ions in the mercury detoxification system. However, also when involved in the latter process, metallochaperones are detected nearly exclusively in organisms encoding $P_{1B-1,2}$-type ATPases as well, possibly indicating that they evolved originally to interact with the

ATPases and then adapted to scavenge also mercury(II). We also described several hints suggesting that metallochaperones can interact also with P-type ATPases lacking MBDs. Conversely, and much more commonly, there are several $P_{1B-1,2}$-type ATPases that likely function in the absence of a partner metallochaperone. This is typically the case of ATPases transporting divalent cations [39]. For these systems, in addition to modulating the overall enzymatic activity, the reciprocal mobility of MBDs could be important to tune the selectivity and/or the affinity for the substrate, which could be the metal ion complexed to either an organic molecule [49] or an unidentified cytoplasmic metallochaperone.

Figure 1. Number of proteins with a given length of the linker connecting the two MBDs closest to the transmembrane domain. Open squares: proteins from complete proteomes retrieved using the Pfam domains; Open circles: proteins from the SMART dataset. Inset: full graph. The *R* correlation coefficient between the two datasets is 0.91.

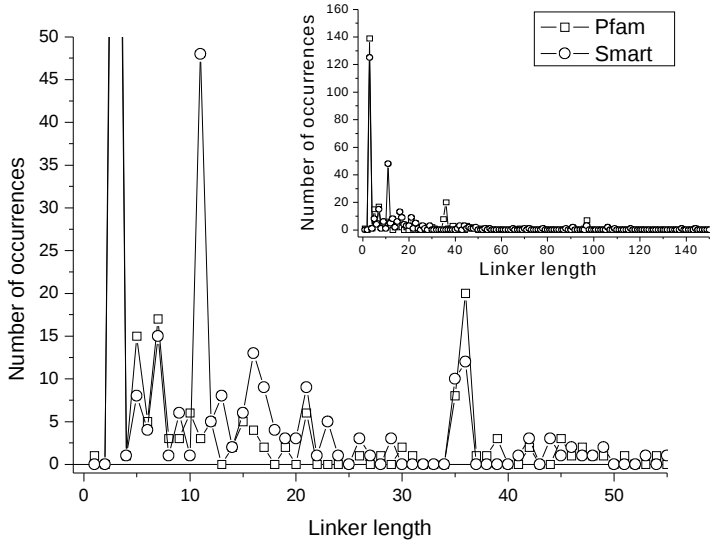


Figure 2. Interdomain interaction energy (separated into electrostatics, van der Waals and total energy, including also hydrogen bonding contributions)

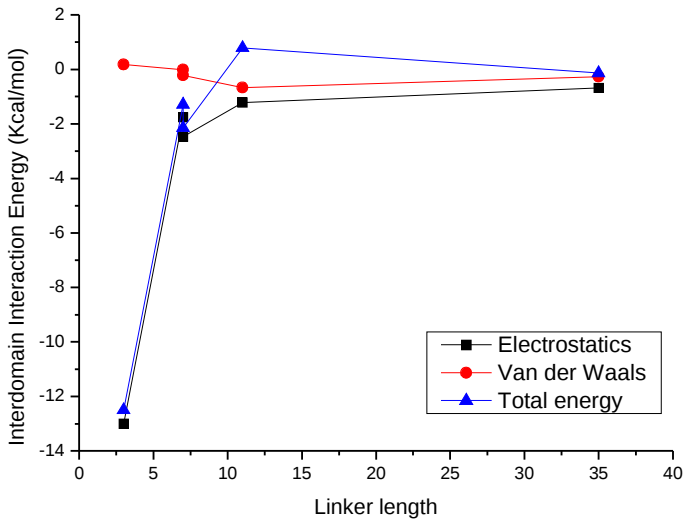


Figure 3 Surface electrostatic potential of the two MBDs of the WND protein (based on the experimental 2EW9 structure). Left: surface of the first domain show, with the second domains in ribbon representation; Right: vice versa. The two panels are interchanged by a rotation of 180∞ along the vertical axis. The electrostatic potential was calculated with the program MOLMOL [50].

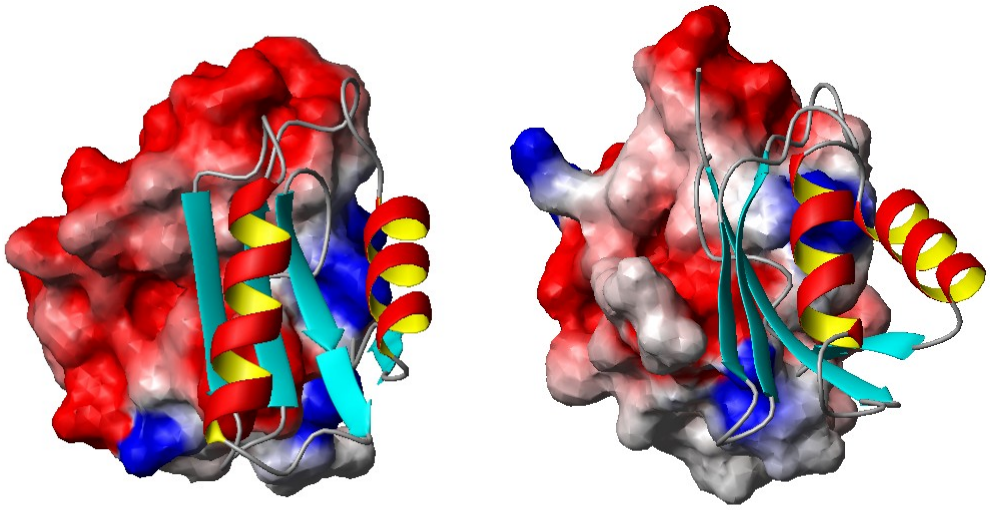


Table 2. Summary of the results of the analysis on 594 fully sequenced prokaryotic genomes using Pfam domains. $P_{1B-1,2}$ ATPase were defined as those containing at least one MBD domain (called HMA in Pfam) in addition to other domains characteristics of P-type ATPases (called Hydrolase and E1-E2_ATPase in Pfam). $P_{1B-1,2}$ ATPases were excluded from the count of P-type ATPases to avoid counting them twice. Metallochaperones were defined as proteins containing only MBD domains

	Number of Proteins	**Number of organisms (percentage of the 594 organisms analysed)**
Individual proteins		
Metallochaperone	521	335 (56.4%)
P-type ATPase (excluding $P_{1B-1,2}$ type)	1625	462 (77.8%)
$P_{1B-1,2}$ ATPase	826	468 (78.8%)
None	n.a.	50 (8.4%)
Combinations of proteins		
Metallochaperone and $P_{1B-1,2}$ATPase	1072 (499+573)	315 (53.0%)
Metallochaperone and not $P_{1B-1,2}$ATPase	22	20 (3.4%)
$P_{1B-1,2}$ATPase and not metallochaperone	253	153 (25.8%)
Not $P_{1B-1,2}$ATPase and not metallochaperone	n.a.	106 (17.8%)
P-type ATPase and $P_{1B-1,2}$ ATPase	2082	386 (65.0%)
P-type ATPase and not $P_{1B-1,2}$ ATPase	240	76 (12.8%)
$P_{1B-1,2}$ ATPase and not P-type ATPase	129	82 (13.8%)

Table 3. $P_{1B-1,2}$ ATPases with the indicated number of HMA domains.

Number of HMA domains	**Number of $P_{1B-1,2}$ ATPases**	**Percentage over the total number of $P_{1B-1,2}$ ATPases**
1	540	65.4%
2	245	29.7%
3	28	3.4%
4	13	1.6%

Reference List

(1) Linder, M. C. *Biochemistry of Copper*; Plenum Press: New York, 1991; pp 1-13.

(2) Andreini, C.; Banci, L.; Bertini, I.; Rosato, A. Occurence of copper through the three domains of life: a bioinformatic approach *J.Proteome Res.* **2008**, *1*, 209-216.

(3) Cavet, J. S.; Borrelly, G. P.; Robinson, N. J. Zn, Cu and Co in cyanobacteria: selective control of metal availability *FEMS Microbiol Rev* **2003**, *27*, 165-181.

(4) O'Halloran, T. V.; Culotta, V. C. Metallochaperones: An Intracellular Shuttle Service for Metal Ions *J.Biol.Chem.* **2000**, *275*, 25057-25060.

(5) Arnesano, F.; Banci, L.; Bertini, I.; Ciofi-Baffoni, S.; Molteni, E.; Huffman, D. L.; O'Halloran, T. V. Metallochaperones and metal transporting ATPases: a comparative analysis of sequences and structures *Genome Res.* **2002**, *12*, 255-271.

(6) Singleton, C.; Le Brun, N. E. Atx1-like chaperones and their cognate P-type ATPases: copper-binding and transfer *Biometals* **2007**, *20*, 275-289.

(7) Lutsenko, S.; Kaplan, J. H. Organization of P-type ATPases: significance of structural diversity *Biochemistry* **1995**, *34*, 15607-15613.

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

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

(10) Lutsenko, S.; Barnes, N. L.; Bartee, M. Y.; Dmitriev, O. Y. Function and regulation of human copper-transporting ATPases *Physiol Rev.* **2007**, *87*, 1011-1046.

(11) La Fontaine, S.; Mercer, J. F. Trafficking of the copper-ATPases, ATP7A and ATP7B: role in copper homeostasis *Arch.Biochem.Biophys.* **2007**, *463*, 149-167.

(12) Clausen, J. D.; Vilsen, B.; McIntosh, D. B.; Einholm, A. P.; Andersen, J. P. Glutamate-183 in the conserved TGES motif of domain A of sarcoplasmic reticulum Ca2+-ATPase assists in catalysis of E2/E2P partial reactions *Proc.Natl.Acad.Sci.U.S.A* **2004**, *101*, 2776-2781.

(13) Lutsenko, S.; Petrukhin, K.; Cooper, M. J.; Gilliam, C. T.; Kaplan, J. H. N-terminal domains of human copper-transporting adenosine triphosphatases (the Wilson's and Menkes disease proteins) bind copper selectively in vivo and in vitro with stoichiometry of one copper per metal-binding repeat *J.Biol.Chem.* **1997**, *272*, 18939-18944.

(14) Payne, A. S.; Gitlin, J. D. Functional expression of the Menkes disease protein reveals common biochemical mechanisms among the copper-transporting P-type ATPase *J.Biol.Chem.* **1998**, *273*, 3765-3770.

(15) Huster, D.; 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.* **2003**, *2789*, 32212-32218.

(16) Cater, M. A.; Forbes, J. R.; La Fontaine, S.; Cox, D.; Mercer, J. F. Intracellular trafficking of the human Wilson protein: the role of the six N-terminal metal-binding sites *Biochem J.* **2004**, *380*, 805-813.

(17) Strausak, D.; La Fontaine, S.; Hill, J.; Firth, S. D.; Lockhart, P. J.; Mercer, J. F. The role of GMXCXXC metal binding sites in the copper-induced redistribution of the Menkes protein *J.Biol.Chem.* **1999**, *274*, 11170-11177.

(18) Voskoboinik, I.; Strausak, D.; Greenough, M.; Brooks, H.; Petris, M.; Smith, S.; Mercer, J. F.; Camakaris, J. Functional analysis of the N-terminal CXXC metal-binding motifs in the human menkes copper-transporting P-type ATPase expressed in cultured mammalian cells *J.Biol.Chem.* **1999**, *274*, 22008-22012.

(19) Walker, J. M.; Huster, D.; Ralle, M.; Morgan, C. T.; Blackburn, N. J.; 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.* **2004**, *279*, 15376-15384.

(20) Banci, L.; Bertini, I.; Cantini, F.; Chasapis, C.; Hadjiliadis, N.; 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.* **2005**, *280*, 38259-38263.

(21) Banci, L.; Bertini, I.; Cantini, F.; Della Malva, N.; Migliardi, M.; Rosato, A. The different intermolecular interactions of the soluble copper-binding domains of the Menkes protein,

ATP7A *J.Biol.Chem.* **2007**, *282*, 23140-23146.

(22) Achila, D.; Banci, L.; Bertini, I.; Bunce, J.; Ciofi-Baffoni, S.; 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.USA* **2006**, *103*, 5729-5734.

(23) Banci, L.; Bertini, I.; Ciofi-Baffoni, S.; Gonnelli, L.; Su, X. C. Structural basis for the function of the N terminal domain of the ATPase CopA from *Bacillus subtilis J.Biol.Chem.* **2003**, *278*, 50506-50513.

(24) Singleton, C.; Banci, L.; Ciofi-Baffoni, S.; Tenori, L.; Kihlken, M. A.; Boetzel, R.; Le Brun, N. E. Structure an Cu(I)-binding properties of the N-terminal soluble domains of *Bacillus subtilis* CopA *Biochem J.* **2008**, *411*, 571-579.

(25) Lutsenko, S.; LeShane, E. S.; Shinde, U. Biochemical basis of regulation of human copper-transporting ATPases *Arch.Biochem.Biophys.* **2007**, *463*, 134-148.

(26) Wu, C. C.; Rice, W. J.; Stokes, D. L. Structure of a copper pump suggests a regulatory role for its metal-binding domain *Structure.* **2008**, *16*, 976-985.

(27) Letunic, I.; Copley, R. R.; Pils, B.; Pinkert, S.; Schultz, J.; Bork, P. SMART 5: domains in the context of genomes and networks *Nucleic Acids Res.* **2005**, *34*, D257-D260.

(28) Finn, R. D.; Mistry, J.; Schuster-Bockler, B.; Griffiths-Jones, S.; Hollich, V.; Lassmann, T.; Moxon, S.; Marshall, M.; Khanna, A.; Durbin, R.; Eddy, S. R.; Sonnhammer, E. L.; Bateman, A. Pfam: clans, web tools and services *Nucleic Acids Res.* **2006**, *34*, D247-D251.

(29) Eddy, S. R. Profile hidden Markov models *Bioinformatics* **1998**, *14*, 755-763.

(30) Pruitt, K. D.; Tatusova, T.; Maglott, D. R. NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins *Nucleic Acids Res.* **2007**, *35*, D61-D65.

(31) Gabaldon, T.; Huynen, M. A. Prediction of protein function and pathways in the genome era *Cell Mol.Life Sci.* **2004**, *61*, 930-944.

(32) Zambelli, B.; Musiani, F.; Savini, M.; Tucker, P.; Ciurli, S. Biochemical studies on Mycobacterium tuberculosis UreG and comparative modeling reveal structural and functional

conservation among the bacterial UreG family *Biochemistry* **2007**, *46*, 3171-3182.

(33) Lee, S. W.; Mitchell, D. A.; Markley, A. L.; Hensler, M. E.; Gonzalez, D.; Wohlrab, A.; Dorrestein, P. C.; Nizet, V.; Dixon, J. E. Discovery of a widely distributed toxin biosynthetic gene cluster *Proc.Natl.Acad.Sci.U.S.A* **2008**, *105*, 5879-5884.

(34) Bertini, I.; Cavallaro, G.; Rosato, A. Cytochrome c: occurrence and functions *Chem.Rev.* **2006**, *106*, 90-115.

(35) Case, D. A., Darden, T. A., Cheatham, T. E., Simmerling, C. L., Wang, J., Duke, R. E., Luo, R., Merz, K. M, Wang, B., Pearlman, D. A., Crowley, M., Brozell, S., Tsui, V., Gohlke, H., Mongan, J., Hornak, V., Cui, G., Beroza, P., Schafmeister, C. E., Caldwell, J. W., Ross, W. S., and Kollman, P. A. AMBER 8. (8.0). 2004. San Francisco, CA, University of California.

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

(37) Poger, D.; Fuchs, J. F.; Nedev, H.; Ferrand, M.; Crouzy, S. Molecular dynamics study of the metallochaperone Hah1 in its apo and Cu(I)-loaded states: role of the conserved residue M10 *FEBS Lett* **2005**, *579*, 5287-5292.

(38) Arg¸ello, J. M. Identification of ion-selectivity determinants in heavy-metal transport P1B-type ATPases *J Membr Biol.* **2003**, *195*, 93-108.

(39) Arg¸ello, J. M.; Eren, E.; Gonzalez-Guerrero, M. The structure and function of heavy metal transport P1B-ATPases *Biometals* **2007**, *20*, 233-248.

(40) Solioz, M.; Vulpe, C. CPx-type ATPases: a class of P-type ATPases that pump heavy metals *Trends Biochem Sci* **1996**, *21*, 237-241.

(41) Banci, L.; Bertini, I.; Ciofi-Baffoni, S.; Su, X. C.; Miras, R.; Bal, N.; Mintz, E.; Catty, P.; Shokes, J. E.; Scott, R. A. Structural basis for metal binding specificity: the N-terminal cadmium binding domain of the P1-type ATPase CadA *J.Mol.Biol.* **2006**, *356*, 638-650.

(42) Mergeay, M.; Monchy, S.; Vallaeys, T.; Auquier, V.; Benotmane, A.; Bertin, P.; Taghavi, S.; Dunn, J.; van der, L. D.; Wattiez, R. Ralstonia metallidurans, a bacterium specifically adapted to

toxic metals: towards a catalogue of metal-responsive genes *FEMS Microbiol.Rev.* **2003**, *27*, 385-410.

(43) Brown, N. L.; Shih, Y. C.; Leang, C.; Glendinning, K. J.; Hobman, J. L.; Wilson, J. R. Mercury transport and resistance *Biochem.Soc.Trans.* **2002**, *30*, 715-718.

(44) Liu, T.; Reyes-Caballero, H.; Li, C.; Scott, R. A.; Giedroc, D. P. Multiple metal binding domains enhance the Zn(II) selectivity of the divalent metal ion transporter AztA *Biochemistry* **2007**, *46*, 11057-11068.

(45) Mandal, A. K.; Arg¸ello, J. M. Functional roles of metal binding domains of the Archaeoglobus fulgidus Cu(+)-ATPase CopA *Biochemistry* **2003**, *42*, 11040-11047.

(46) Rodriguez-Granillo, A.; Wittung-Stafshede, P. Structure and dynamics of Cu(I) binding in copper chaperones Atox1 and CopZ: a computer simulation study *J.Phys.Chem.B* **2008**, *112*, 4583-4593.

(47) Martinez, R. J.; Wang, Y.; Raimondo, M. A.; Coombs, J. M.; Barkay, T.; Sobecky, P. A. Horizontal gene transfer of PIB-type ATPases among bacteria isolated from radionuclide- and metal-contaminated subsurface soils *Appl.Environ.Microbiol.* **2006**, *72*, 3111-3118.

(48) Toyoshima, C.; Nakasako, M.; Nomura, H.; Ogawa, H. Crystal structure of the calcium pump of sarcoplasmic reticulum at 2.6 A resolution. *Nature* **2000**, *405*, 647-655.

(49) Cobine, P. A.; Ojeda, L. D.; Rigby, K. M.; Winge, D. R. Yeast contain a non-proteinaceous pool of copper in the mitochondrial matrix *J.Biol.Chem.* **2004**, *279*, 14447-14455.

(50) Koradi, R.; Billeter, M.; W¸thrich, K. MOLMOL: a program for display and analysis of macromolecular structure *J.Mol.Graphics* **1996**, *14*, 51-55.

CHAPTER 3

A systematic investigation of multi-heme *c*-type cytochromes in prokaryotes

Introduction .

c-type cytochromes are widespread metalloproteins that contain one or more covalently linked heme cofactors. The covalent linkage of the cofactor to the protein typically occurs through the chemical modification of the two vinyl groups of the cofactor, which react with two cysteine side chains to form thioether bonds in a stereospecific manner [1-3]. In the majority of cases, one protein ligand to the iron ion is the side chain of a His residue immediately following the second Cys residue (often called the proximal ligand). The typical signature for c-type heme attachment is therefore a CXXCH sequence motif [1]. Some exceptions are known, both regarding the residues covalently linked to the cofactor and the proximal iron ligand. As an example of the first case, in a restricted number of c-type cytochromes the heme moiety is linked via a single thioether bond [4]. With respect to the second case, the penta-heme cytochrome *c* nitrite reductase binds one of the heme groups via an atypical CXXCK motif whose Lys side chain coordinates the iron ion [5]. In structurally characterized *c*-type cytochromes, the number of heme cofactors that are bound to a single polypeptide chain ranges from one to as many as sixteen [6]. Those harbouring a single heme cofactor normally function as electron transfer proteins within redox chains, e.g., in aerobic or anaerobic respiration [7]. Instead, multi-heme *c*-type cytochromes (MHC's hereafter) can have a larger variety of biochemical roles, including enzymatic activity next to electron transfer [8]. In the latter role, the spatial proximity of the heme cofactors in MHC's allows electrons to rapidly travel across relatively long distances through subsequent intraprotein electron transfer steps. This feature can be further enhanced by the interaction of two or more MHC's.

In the present work, we applied computational strategies to get an overview of the distrubution of MHC's in prokaryotic organisms. We focused only on proteins containing the canonical CXXCH motif, as MHC's containg only non-canonical motifs are presently not known. This work constitutes an extension of our previous investigations of mono-heme, mitochondrial-type *c*-type cytochromes [7;9], and complements our investigation of the occurrence and characteristics of prokaryotic molecular machineries for the biosynthesis and uptake from external sources of the heme cofactor. The strategy adopted is a variation on what we developed and refined over the last few years to identify

metalloproteins in genomes. From the computational point of view, additional challenges arise due to the fact that the proteic part of multi-heme c-type cytochromes often does not adopt a well-defined fold. Indeed, it can be often the case that the structure of the protein is defined by the linkage of and hydrophobic interactions around the multiple heme cofactors [10]. This may be particularly true for cytochromes containing the higher the number of cofactors bound to the protein. Thanks to the extensive dataset of MHC's gathered in this work, we can provide various hints for future experimental work in the field.

Materials and methods

We downloaded 594 completely sequenced prokaryotic genomes from the RefSeq database (http://www.ncbi.nlm.nih.gov/RefSeq/) [11], which cumulatively coded more than 1.9 million protein sequences. The list of organisms investigated is given in Supplementary Table S1. As a first filter, we removed all proteins that did not contain at least two CXXCH motifs. The resulting ensemble of protein sequences was analysed in terms of domain content, using the program HMMER (http://hmmer.janelia.org) with standard parameters [12]. In particular, the cutoff for domain assignment was set at an E-value of 10^{-5}. All Hidden Markov Models (HMM) from the Pfam [13] and Superfamily [14] databases were used. To select Pfam HMM's representing true multi-heme cytochromes c, we identified which HMM's that were detected in the protein sequences consistently contained at least two CXXCH motifs within their boundaries. For all of these HMM's, we inspected the corresponding sequence logo to verify whether the CXXCH motifs had indeed a high degree of conservation. We then inspected the available literature information, starting from the domain description provided in Pfam. For the Superfamily database, we selected HMM's on the basis of available structures of multi-heme cytochromes as described below. We extracted from the PDB all the structures of true multi-heme cytochromes c by a query for structures containing multiple heme cofactors, with a 70% sequence identity filter, followed by manual inspection of the results to remove proteins containing multiple mono-heme domains, such as the SoxA subunit in the sulfur oxidizing enzymatic machinery [15], rather than binding multiple heme cofactors in a single domain. We run the entire Pfam and Superfamily databases against this PDB-derived dataset and retained all the HMM models that were assigned to one of the proteins in the dataset and contained within their boundaries multiple CXXCH motifs (Supplementary Table S2). As a counter-check, we then used this selection of HMM's to scan with HMMER the ensemble of all the sequences of proteins in the PDB, and verified that at our level of confidence the program retrieved only MHC's. Note that for Pfam the list of HMM's obtained from this analysis of the PDB is a subset of the one obtained from literature analysis, as it contains only domains for which a representative with known structure is available. This procedure

constitutes nevertheless an independent validation of our selection of Pfam HMM models, using the PDB database as a "gold standard".

With the above lists of HMMs, we inspected the results generated by HMMER for all proteins in the selected genomes that had at least two CXXCH motifs. We retained as true MHC's all the proteins that contained at least one of the domains corresponding to one of the selected Pfam HMMs. We instead removed all proteins whose CXXCH motifs were within the boundaries of other Pfam domains. In other words, we separated the initial list of proteins containing multiple CXXCH motifs into three groups on the basis of their content of Pfam domains: those that could be assigned as true MHC's, those that could be assigned as other proteins incidentally containing the motifs or using the motifs for other purposes, such as zinc-binding, and those that could not be assigned to any Pfam domain. Proteins that fell in the second group were discarded, whereas proteins in the third group were further analyzed. This was done initially once more against Pfam using a less restrictive 10^{-3} E-value threshold, and then using the Superfamily HMM's following essentially the same procedure. Proteins that could not be assigned to any of the Superfamily HMM's representing multi-heme cytochromes c were analyzed against the entire Superfamily database to discard non-MHC proteins.

Following the methodology described above, we obtained a list of multi-heme cytochromes c, characterized by the presence of at least two CXXCH motifs within at least one MHC (represented by either a Pfam or Superfamily HMM model). In addition, we also obtained a list of protein sequences with at least two CXXCH motifs that could not be assigned to any HMM of the entire Pfam database nor of the entire Superfamily database and are thus unclassified proteins. MHCs and unclassified proteins were clustered by CLANS [16] to obtain additional biochemical insight using an E-value threshold variable between 10^{-10} and 10^{-20}. We found, also on the basis of our previous experience [7], that very restrictive E-value thresholds are needed in CLANS to obtain an useful and functionally informative clustering of proteins. The complete list of results is given in Supplementary Table S3.

Results and Discussion

Selection of MHC domains

The list of Pfam domains representing MHC's that we assembled for the present work is given in Table 1, which indicates also PDB assignments. The corresponding information for the Superfamily database is given in Table 2. 82 different structures of MHC's were selected from the PDB. Table S2 reports the PDB assignments on a per-PDB code rather than per-domain (as in Tables 1 and 2) basis. Note that protein structures containing repetitions of mono-heme cytochrome c domains were excluded as they do not really fit the concept of MHC (i.e. a single structural unit that harbours at least two *c*-type heme cofactors) even though they do contain multiple heme cofactors in a single polypeptide

chain. We addressed the systems containing two or more mono-heme cytochromes in a previous study [7]. Only for one MHC of known structure no assignment to a HMM for both the Pfam and Superfamily domain databases could be obtained. This is the case of PDB entry 2CZS, which corresponds to a protein from *Geobacter sulfurreducens* [17]. BLAST [18] searches showed that homologues of this protein are only present in other species of the *Geobacter* genus. Owing to the methodology adopted, for all Superfamily domains at least one structural representative was identified, whereas for three Pfam domains this was not the case (Table 1). Identification of MHC domains lacking structural characterization is possible in Pfam thanks to the links to the scientific literature that the database is providing. Some independent support for the inclusion of these domains in Table 1 can be obtained from the sequence logo, which shows the conservation of the CXXCH motif in the members of the corresponding protein families (not shown). The analysis of PDB assignments provided insight on some different features of the Superfamily and Pfam databases, which can in principle affect the way they should be used. All Pfam domains corresponded to at least two PDB structures, whereas a significant number of Superfamily domains corresponded to only a single structure (Table 2). On the other hand, the number of PDB structures for which we did not get an assignment to a Pfam domain with our threshold of E-value was larger than for Superfamily domain assignments (29 and 4, respectively). Overall, the examined ensemble of PDB structures was assigned to a total of five different Pfam domains and 26 Superfamily domains. We can thus conclude that the granularity of the domain distribution in Superfamily is higher than in Pfam; at the same time, the coverage of the PDB by Superfamily is superior. This latter point is likely the result of the fact that Superfamily is originally built from the SCOP database, i.e., on structural rather than sequence families. A direct consequence of the above observations is that typically a Pfam assignment corresponds to several different Superfamily assignments. In other words, the various PDB structures that contain the same Pfam MHC domain can be assigned to different Superfamily MHC domains. For example, PDB entries assigned to the Cytochrom_CIII Pfam domain span as many as eight different Superfamily domains. This finer classification may relate to Superfamily picking up small structural variations within the Cytochrom_CIII family.

The finding and considerations described in the preceding paragraph provided the rationale for how we built our workflow of sequence analysis. The list of all the protein sequences containing multiple CXXCH motifs in a proteome contains the list of all MHC's that the organism of interest can produce. On the other hand, it is almost certain to include also non-MHC proteins. The inspection of Pfam domain assignments at a relatively stringent 10^{-5} threshold provides an easy and quick way to start separating MHC's from non-MHC's. At a less stringent 10^{-3} threshold the level of separation is only slightly enhanced (within 5%). It is reasonable to perform the Pfam analysis first because of the lower num-

ber of relevant domains and consequently greater ease of analysis. Unassigned proteins are then more finely investigated with Superfamily, and additional assignments are obtained thanks to its greater coverage (note this applies only to systems similar to proteins of known structure). The protein sequences containing multiple CXXCH motifs that remained unassigned at the end of the procedure constitute potential unprecedented MHC's (although very likely not all of them will be).

Proteome-level distribution and properties of MHC's

The numerical results of our analysis of 594 proteomes are summarized in Table 3. Out of the over 1.9 million proteins that we examined, 3783 contained two or more CXXCH motifs. In 1330 proteins the motifs (all or all but one) were contained in a Pfam domain *not* in Table 1 and therefore they were removed as they could not be true MHC's, whereas 607 could be assigned as true MHC's because two or more motifs fell within the boundaries of a true Pfam MHC domain. The remaining 1846 sequences were analyzed against the Superfamily domain database, leading to an additional 794 sequences being rejected and 985 sequences being assigned as MHC's. 67 sequences remained completely unassigned, of which 7 were of homologues of the above-mentioned 2CZS PDB structure and were therefore added to the list of true MHC's. In summary, we identified 1599 true MHC's, and remained with 60 unassigned proteins. The true MHC's constituted 42% of the initial list of proteins harbouring two or more CXXC motifs. The list of MHC's is provided as Table S3. For all subsequent analyses, the true MHC's and the unassigned proteins were grouped together. We also assumed that all CXXCH motifs bound one heme cofactor, even though it is of course well possible that some of them are actually involved in the formation of disulfide bonds or binding metal ions such as zinc(II), or both.

Our results corresponded to an average of about 2.7 MHC's per organism studied. Unsurprisingly, the actual distribution of MHC's among the various organisms is highly variable. Indeed, only 258 organisms encoded at least one MHC, and therefore the majority of proteomes (56.6%) did not contain any. Table 4 reports some statistics describing the distribution of MHC's, taking into account only the organisms that do encode at least one such protein. The average number of MHC's encoded by an organism was six; however, it is important to note that half of the organisms encoded only three MHC's or less. One quarter of the organisms that do contain MHC's had more than six such proteins, and 5% of the organisms had more than 22 MHC's. *Geobacter uraniumreducens* Rf4 was the organism with the largest number of MHC's (75), corresponding to 1.7% of its proteome.

Another interesting statistics is that regarding the number of heme cofactors potentially bound to the MHC's identified in the present work. To illustrate this, Figure 1 depicts the number of MHC proteins containing a given number of CXXCH motifs were present in our final dataset. The most com-

mon number of motifs, and therefore of potentially bound heme cofactors, in a single protein sequence was four (25.0% of instances); two was nearly as common (23.1%). MHC's with five motifs were also relatively common. Penta-heme MHC's were the next most widespread type. However, this resulted from the combination of two types of proteins: true penta-heme MHC's, such as NrfB homologues [19], and tetra-heme domains fused to mono-heme cytochromes, such as TorC or TorY [20] homologues. At higher heme contents, sequences with eight and especially ten motifs were relatively common. Deca-heme MHC's could be divided in two groups, i.e., those similar to OmcA proteins, which were found mainly in *Shewanella* species [21], and members of the DmsE family, a potential anaerobic DMSO reductase, which were found in *Shewanella* and some *Geobacter* species. Among less common heme numbers, 12 (1.3%), 16 (0.7%) and 26-27 (0.4% each) stood out. The higher (25 or greater) heme numbers were found only in δ-proteobacteria of the *Geobacter* and *Anaeromyxobacter* genus. These were typically divided by our analysis in a number of smaller MHC domains. However, in the absence of structural data, it is possible that this separation resulted from an unsatisfactory description of the large MHC's in the currently available ensembles of HMM models (Tables 1 and 2).

It has been pointed out that the covalent attachment of the heme group allows smaller protein length:heme ratios, which therefore tend to be larger in mono-heme cytochromes (either of *c*- or other types) than in MHC's. The lowest ratios in mono-heme cytochromes *c* are around 60-70 for some bacterial proteins, whereas eukaryotic cytochromes feature typically 100 amino acids or more per heme. The present large dataset of MHC's allowed us to verify whether there was a trend in these ratios with the number of heme groups bound by the polypeptide. As shown in Figure 2, there was no significant difference between the ratios observed in MHC's with three to ten motifs. Also for MHC's with two motifs there is in principle no significant difference with respect to the others, because of the large standard deviation around the average . However, the situation of di-heme MHC's is clearly different, as ratios from about 70 up to 270 are observed. For example, all MHC's containing the Pfam domain DUF1111 had ratios between 230 and 280 (see also next section for discussion on this domain), whereas NapB homologues had ratios between 70 and 85. The detailed domain composition of the various MHC's has a significant impact on the computed values: for example, the penta-heme MHC's that are TorC homologues had ratios at the higher end of the corresponding range of Figure 2, but this was in good part due to the fact that, as detailed in the next section, they are composed by a tetra-heme domain fused to a mono-heme domain, the latter having a higher ratio than the rest of the protein. Another contribution toward the increase of the ratio results from linker regions in multi-domain MHC's. In summary, the typical heme:protein length ratio for MHC's is around 60-70 regardless of the number of motifs, with the possible exception of di-heme MHC's. Multi-domain MHC's tend to have higher ratios

than more compact MHC's.

Functional insights

Further information can be obtained from the dataset of MHC's that we assembled in this work by clustering the proteins with the program CLANS. Figure 3 graphically depicts the 30 largest clusters identified, whose size varied between 10 and 218 proteins and which cumulatively accounted for three quarters of the dataset. Some information is summarized in Table 5 for the largest clusters, which are discussed in the following. The largest cluster contained Torc and TorY homologues, which have five CXXCH motifs in a mono- plus a tetra-heme domain [20], and NapC/NirT homologues, which have four CXXCH motifs in a tetra-heme domain. These MHC's mediate electron transfer from the quinine pool, directly or via other electron transfer proteins, to molybdo-enzymes reducing respectively DMSO or nitrate. In TorC, it has been proposed that the mono-heme cytochrome c domain injects the electron into the enzyme TorA, after receiving it from the menaquinone pool through the tetra-heme domain [20]. Notably, there is no MHC of known structure within this cluster, making this group a clear target for structure determination. Some could be detected similarity to the NrfH structure (PDB entry 2J7A; see also later). Indeed, the four heme-binding motifs in the tetra-heme domain do align to those of NrfH, but there is a significant difference in sequence spacing between heme 3 and 4. A (distant) evolutionary relationship between the TorC/NapC and NrfH families has been proposed [22]. The second largest cluster containts 111 NrfA (nitrite reductase) homologues (Figure 4). NrfA binds five heme groups, of which one (the first in sequence) is bound via an unusual CXXCK motif (PDB Entry 1FS9 [23]). Interestingly, in the present cluster we also identified proteins with five canonical CXXCH motifs in various organisms, e.g. from the *Campylobacter* genus (Table S3). The alignment between the proteins with four motifs plus the CXXCK motif and the proteins with five motifs was very good; in particular, the first CXXCH of the latter proteins corresponded perfectly to the CXXCK motif of the former. The proteins are therefore very likely to be true NrfA enzymes. It is to be noted that the His residue in the first CXXCH motif may actually not bind heme as seen in the octaheme terathionate reductase (OTR, see later). The cluster contains also three proteins with one CXXCK motif and seven canonical motifs. The structure of a homologous protein from *Thiolkalivibrio nitratireducens*, a bacterium not analyzed here, has appeared this year (PDB entry 2OT4 [24]). The five heme cofactors of NrfA can be superimposed well to five of the eight heme cofactors of the *T. nitratireducens* enzyme, whereas the additional three of the latter are contained in an extra structural domain. CXXCK motifs were identified only in the members of this family out of the 1659 sequences constituting our final ensemble of MHC's. Cluster 4 was composed by 96 proteins containing the Pfam DUF1111 domain (DUF stands for Domain of Unknown Function [13]). All these proteins contained two CXXCH motifs, and are

sometimes annotated as being thiol oxidoreductases. We detected some sequence similarity in the C-terminal region to a di-heme cytochrome c peroxidase of known structure (PDB Entry 1RZ5 [25]), suggesting that also these proteins may contain two mono-heme domains [7] rather than a true MHC domain. These proteins are consistently associated in the respective genomes organization to a putative lipoprotein and other uncharacterized proteins. Also due to its relatively large size, the present family clearly is a target for both functional and structural investigation. Members of the MtrF and MtrA families (together with other cytochromes not annotated) constitute clusters 5 and 6 respectively. The majority of these proteins are from organisms of the *Shewanella* genus, which are renowned for their rich content of cytochromes, where they are presumably involved in the dissimilatory reduction of metal (hydr)oxides [26]. The metal-reductase containing locus in Shevanella species is known to encode a variable number of deca- and undecaheme MHC's [21]. They are mostly decaheme cytochromes, but, in particular for cluster 6, the number of motifs can range from 7 up to 26. Cluster 9 contained 40 pentaheme NrfB homologues (Figure 4). NrfB is an electron donor to NrfA [27;28] in proteobacteria, whereas the same role is played by NrfH in other organisms. Other more distant members of the family are contained in cluster 20. NrfH proteins defined cluster 11, with 33 members, and typically had four motifs. At the highly selective threshold of 10^{-20} the sequence of the *D. vulgaris* NrfH protein in the aforementioned 2J7A structure did not fall in cluster 11 (Table 5), but joined the other NrfH proteins at a still tight threshold of 10^{-15}. This is due to the *D. vulgaris* NrfH sequence being 20-30 amino acids shorter than the others. Finally, it is worth commenting cluster 12: it contained 38 proteins having eight motifs in all cases but one, which could be assigned as hydroxylamine reductases [29] or tetrathionate reductase (OTR) (Figure 4). As already mentioned, the structure 1SP3 of OTR [30] shows that one heme group in these enzymes has an anomalous ligation, with the His residue of the CXXCH motif not bound and being replaced by a Lys side chain, in a way that is reminiscent of NrfA ligation [5;23]. MHC variants in this cluster featured from five to ten CXXCH motifs.

By analyzing more closely the 15 largest clusters we noticed that when raising the clustering threshold to 10^{-15}, cluster 6 and 9 merged, even though they typically contain 10 and 5 motifs, respectively. In addition, also clusters 20 and 24 ended up in the same group, again with respectively 5 and 8-10 motifs. From a detailed analysis of alignments, it appeared that the N-terminal 200 amino acids of the larger MHC's aligned well to the entire sequence of the smaller ones, suggesting that the former, which corresponds to the *Shewanella* MtrA family, may a have a modular organization in which the first five hemes are contained in an element quite similar to the entire NrfB protein. The other larger clusters did not merge with one another until a threshold of 10^{-10} was used. However, at this level several functionally unrelated families were grouped together, suggesting that this threshold is inappropriate

to generate meaningful clusters.

Finally, it is interesting to compare the per-organism distribution of MHC's with respect to the distribution of the enzymatic machineries for heme biosynthesis and uptake. For the correct production of MHC's it is expected that an organism is able to either synthesize or acquire from a host the heme cofactor. By comparing the present dataset to that described by us in [31]. We observed only three possible exceptions to this rule. One protein in the organism *Bifidobacterium adolescentis* did not contain any known Pfam or Superfamily domain and thus was included in the final dataset as an unassigned protein only on the basis of the presence of two CXXCH motifs. This protein does not cluster with any other, and it is presumably a false positive. The second case is that of five proteins in *Haemophilus ducreyi*, which include a complete system for nitrite reduction. In this case, the discrepancy should be ascribed to the fact that the heme uptake system of *Haemophilus* is poorly characterized and does not closely resemble that of other proteobacteria. Indeed, the need of *Haemophilus ducreyi* for heme as an iron source is documented [32], and other hemoproteins from this organism have been characterized [33]. Finally, we identified in *Streptococcus thermophilus* a fusion between a MerR-type regulator and a putative MHC domain. *Streptococcus pyogenes*, which is able to take up heme [31], and some *Staphylococci* contained highly similar proteins, but not *Streptococcus pneumoniae*, which also cannot synthesize nor take up heme with the common machineries although its virulence depends on hemin [34]. At present, we can only speculate that *Streptococcus thermophilus* contains an uncharacterized heme uptake system.

Conclusion

We identified 1659 MHC's or unassigned proteins containing multiple CXXCH motifs in 258 organisms (out of 594 analyzed). The presence of MHC's is a good indicator of an organism's ability to take up or synthesize heme; in two cases, that of *Haemophilus ducreyi* and of *Streptococcus thermophilus*, the presence of MHC's in the genome may suggest that they have an uncommon or highly divergent heme uptake pathway. The most common number of heme-binding motifs in a sequence was four (25%) and two (23%), followed by five (13%) and ten (9.8%).

However, the variability also within a group of MHC's with the same number of motifs was relatively high. Only within individual functional families the ratio exhibited little variability. This observation could thus be useful to corroborate functional assignments of novel MHC's.

The detailed comparison of the MHC sequences retrieved provided various hints to direct future experimental work in field. For example, we identified homologues of the NRfA nitrite reductase where the CXXCK motif of the catalytic heme is replaced by a fifth 5 CXXCH motif. In addition, we identified sequence similarity between deca-heme MHC's of the MtrA family and penta-heme MHC's

of the NrfB family. As a general consideration, it appears that the amount of structural information currently available for MHC's is limited with respect to the diversity of this broad class of metalloproteins, and even some of the largest MHC clusters lack a structurally characterized member. This limits the possibility e.g. to perform systematic structural modeling of MHC's because of the poor selection of templates, which would affect key factors such as the reciprocal position and orientation of the heme cofactors and their ligands. Experimental efforts in the structural investigation of MHC's are thus warranted.

References list :

1. Scott RA, Mauk AG (1996) Cytochrome c. A multidisciplinary approach. University Science Books, Sausalito, California

2. Barker PD, Ferguson SJ (1999) Structure Fold Des 7:R281-R290

3. Stevens JM, Daltrop O, Allen JW, Ferguson SJ (2004) Acc Chem Res 37:999-1007

4. Allen JW, Ginger ML, Ferguson SJ (2004) Biochem J 383:537-542

5. Einsle O, Messerschmidt A, Stach P, Bourenkov GP, Bartunik HD, Huber R., Kroneck PMH (1999) Nature 400:476-480

6. Czjzek M, ElAntak L, Zamboni V, Morelli X, Dolla A, Guerlesquin F, Bruschi M (2002) Structure 10:1677-1686

7. Bertini I, Cavallaro G, Rosato A (2006) Chem Rev 106:90-115

8. Mowat CG, Chapman SK (2005) Dalton Trans 7:3381-3389

9. Banci L, Bertini I, Rosato A, Varani G (1999) J Biol Inorg Chem 4:824-837

10. Daltrop O, Ferguson SJ (2003) J Biol Chem 278:4404-4409

11. Pruitt KD, Tatusova T, Maglott DR (2007) Nucleic Acids Res 35:D61-D65

12. Eddy SR (1998) Bioinformatics 14:755-763

13. Finn RD, Tate J, Mistry J, Coggill PC, Sammut SJ, Hotz HR, Ceric G, Forslund K, Eddy SR, Sonnhammer EL, Bateman A (2008) Nucleic Acids Res 36:D281-D288

14. Gough J, Chothia C (2002) Nucleic Acids Res 30:268-272

15. Bamford VA, Bruno S, Rasmussen T, Appia-Ayme C, Cheesman MR, Berks BC, Hemmings AM (2002) EMBO J 21:5599-5610

16. Frickey T, Lupas A (2004) Bioinformatics 20:3702-3704

17. Heitmann D, Einsle O (2005) Biochemistry 44:12411-12419

18. Altschul SF, Madden TL, Schaeffer A, Zhang J, Zhang Z, Miller W, Lipman DJ (1997) Nucl Acids Res 25:3389-3402

19. Hussain H, Grove J, Griffiths L, Busby S, Cole J (1994) Mol Microbiol 12:153-163

20. Gon S, Giudici-Orticoni MT, Mejean V, Iobbi-Nivol C (2001) J Biol Chem 276:11545-11551

21. Fredrickson JK, Romine MF, Beliaev AS, Auchtung JM, Driscoll ME, Gardner TS, Nealson KH, Osterman AL, Pinchuk G, Reed JL, Rodionov DA, Rodrigues JL, Saffarini DA, Serres MH, Spormann AM, Zhulin IB, Tiedje JM (2008) Nat Rev Microbiol 6:592-603

22. Bergmann DJ, Hooper AB, Klotz MG (2005) Appl Environ Microbiol 71:5371-5382

23. Einsle O, Stach P, Messerschmidt A, Simon J, Kroger A, Huber R, Kroneck PM (2000) J Biol Chem 275:39608-39616

24. Polyakov KM, Boyko KM, Tikhonova TV, Slutsky A, Antipov AN, Zvyagilskaya RA, Popov AN, Bourenkov GP, Lamzin VS, Popov VO (2009) J Mol Biol 389:846-862

25. Dias JM, Alves T, Bonifacio C, Pereira AS, Trincao J, Bourgeois D, Moura I, Romao MJ (2004) Structure 12:961-973

26. Shi L, Squier TC, Zachara JM, Fredrickson JK (2007) Mol Microbiol 65:12-20

27. Clarke TA, Dennison V, Seward HE, Burlat B, Cole JA, Hemmings AM, Richardson DJ (2004) J Biol Chem 279:41333-41339

28. Clarke TA, Cole JA, Richardson DJ, Hemmings AM (2007) Biochem J 406:19-30

29. Atkinson SJ, Mowat CG, Reid GA, Chapman SK (2007) FEBS Lett 581:3805-3808

30. Mowat CG, Rothery E, Miles CS, McIver L, Doherty MK, Drewette K, Taylor P, Walkinshaw MD, Chapman SK, Reid GA (2004) Nat Struct Mol Biol 11:1023-1024

31. Cavallaro G, Decaria L, Rosato A (2008) J Proteome Res 11:4946-4954

32. Lee BC (1991) J Med Microbiol 34:317-322

33. Pacello F, Langford PR, Kroll JS, Indiani C, Smulevich G, Desideri A, Rotilio G, Battistoni A (2001) J Biol Chem 276:30326-30334

34. Tai SS, Lee CJ, Winter RE (1993) Infect Immun 61:5401-5405

35. Brige A, Leys D, Meyer TE, Cusanovich MA, Van Beeumen JJ (2002) Biochemistry 41:4827-4836

36. Igarashi N, Moriyama H, Fujiwara T, Fukumori Y, Tanaka N (1997) Nat Struct Biol 4:276-284

37. Gibson HR, Mowat CG, Miles CS, Li BR, Leys D, Reid GA, Chapman SK (2006) Biochemistry 45:6363-6371

Table 1. HMM's describing MHC's identified in the Pfam database and corresponding PDB entries.

Pfam HMMs	**PDB entries**
Cytochrom_CIII	1gmb, 2e84, 2cdv, 3cao, 1a2i, 2cdv, 1z1n, 1upd, 1w7o, 2yxc, 1qn0, 1i77, 2cth, 2bq4, 1up9, 1gws, 2z47, 1mdv, 2ewi, 2cy3, 1j0p, 3car, 1gm4, 1czj, 1gx7, 2a3m, 1wad, 2ewu, 2cym, 2yyx, 1duw, 1ofy, 1it1, 1j0o, 2cth, 2yyw, 1ofw, 2ewk, 19hc, 2a3p, 1qn1, 1aqe, 1h29, 2bpn, 2ffn, 2cy3, 1gyo, 2cvc, 3cyr, 1wr5, 3cyr
Cytochrom_C552	1gu6, 2rf7, 2e81, 2j7a, 3f29, 1fs7, 3f29, 3bnf, 1fs9, 2ot4, 2e80, 2vr0, 1oah, 3bnh, 3bnj, 1fs8, 2rdz, 3bng, 1qdb
DHC	2fw5, 2fwt
CytoC_RC	2prc, 1vrn, 6prc, 1prc, 1dxr, 2jbl, 4prc, 1r2c, 1txw, 3d38, 2i5n, 5prc, 7prc, 1eys, 3prc
NapB	1jni, 1ogy
Cytochrom_NNT	2j7a, 2vr0
Paired_CXXCH_1	-
Gsu_C4xC_C2H	-
DUF1111	-

Table 2. HMM's describing MHC's identified in the Superfamily database and corresponding PDB entries.

SUPERFAMILY HMMs	**PDB entries**
0035167	1aqe, 1czj
0037755	3bnf, 3bng, 3bnh, 3bnj, 2ot4, 2e80, 2e81, 1fs7, 1fs8, 1fs9, 1gu6, 2rdz, 2rf7
0038903	2ffn, 1a2i, 1wr5, 1cdv, 1cth, 1up9, 1upd, 3cyr, 2cth, 2cym, 2cyr, 2ewi, 2ewk, 2ewu, 1gm4, 1gmb, 1gx7, 1i77, 1it1, 1j0o, 1j0p, 2bpn, 2cdv, 2yxc, 2yyw, 2yyx, 2z47
0037345	1lm2, 1new, 1cfo, 1ehj, 2new, 1f22, 1kwj, 1l3o, 1hh5
0042897	1qo8
0045238	1w7o, 1cy3, 2cy3
0036657	2prc, 1dxr, 1vrn, 1prc, 1r2c, 3d38, 3prc, 2jbl, 5prc, 6prc, 7prc, 2i5n, 2jbl, 1vrn
0038266	2ozy, 2p0b
0036613	19hc, 1duw, 1ofw, 1ofy
0036731	1lj1, 1m64, 1y0p, 1e39, 1q9i, 1qjd, 1p2e, 1p2h, 1jrx, 1jry, 1jrz, 1kss, 1ksu, 2b7r, 2b7s
0042718	1qdb
0044945	1mdv, 1wad, 2a3m, 2a3p, 1qn0, 1qn1, 2bq4
0038377	1z1n, 2cvc, 2e84, 1gws, 1h29
0037766	1bvb, 1ft5, 1ft6
0041763	1oah, 2j7a, 2vr0
0043396	3bxu, 1os6, 1rwj
0036263	1d4c, 1d4d, 1d4e
0038331	1gyo
0041853	1ogy
0038376	1ddc, 1h21
0045497	3cao, 3car
0040737	1m1p, 1m1p, 1m1q, 1m1r
0039660	1jni
0037571	1fgj
0043690	1sp3
0037264	1eys, 2j7a, 2vr0

Table 3. Summary of input data and results.

Overview of Initial Data	
Number of organisms analyzed	594
Number of sequences analyzed	1,900,966
Number of sequences containing two CXXCH motifs or more	3,783
Analysis of sequences containing two CXXCH motifs or more	
Number of MHC's assigned by Pfam at 10^{-5} threshold	585
Number of sequences rejected by Pfam at 10^{-5} threshold	1,286
Number of MHC's additionally assigned by Pfam at 10^{-3} threshold	22
Number of sequences additionally rejected by Pfam at 10^{-3} threshold	44
Number of MHC's additionally assigned by Superfamily at 10^{-5} threshold	947
Number of MHC's additionally assigned by Superfamily at 10^{-3} threshold	38
Number of sequences additionally rejected by Superfamily at 10^{-3} threshold	794
Number of unassigned sequences	67
Number of homologues to 2CZS	7
Summary of results	
Total number of MHC's	1,599
Total number of rejected sequences	2,124
Total number of unassigned sequences	60
Number of organisms with at least one MHC or unassigned sequence	258

Table 4. Statistics describing the per-organism distribution of MHC's in the final dataset. Only organisms containing at least one MHC or unassigned protein are taken into account. For the minimum and maximum values of MHC's the number of organisms containing that value is given in parentheses.

Mean	6 ± 10
First quartile	2
Median	3
Third quartile	6
95th percentile	22
Minimum value	1 (51)
Maximum value	75 (1)

Table 5. Properties of the fifteen largest clusters detected by CLANS. Only motif numbers occurring in at least 10% of the members are reported. The number of members with a given number of motifs in each cluster is given in parentheses in the third column.

Cluster #	**Members**	**# of Motifs (members with given # of motifs)**	**PDB representative**
1	218	4 (129), 5 (85)	None
2	111	4 (91), 5 (16)	1FS7, 3BNG, 1QDB, 1GU6 1OAH, 2J7A, 2OT4, 2RF7, 2VR0, 3F29, 1FS9
3	111	2 (111)	1OGY, 1JNI
4	96	4 (96)	None
5	81	10 (72), 11 (9)	None
6	78	10 (56)	None
7	49	7 (16), 8 (27)	None
8	41	From 9 to 45	None
9	40	5 (40)	2P0B
10	38	8 (32)	None
11	33	4 (32)	None
12	31	8 (30)	1FGJ
13	30	2 (30)	2FWT, 2FW5
14	28	5 (24), 6 (4)	None
15	20	7 (4), 9 (9), 6 (3)	None

Figure 1. Distribution of MHC's as a function of the number of heme-binding motifs. The graph shows the number of MHC's in our final dataset that have a given number of CXXCH motifs. The inset reports a vertical expansion to appreciate the values for proteins with more than 10 motifs.

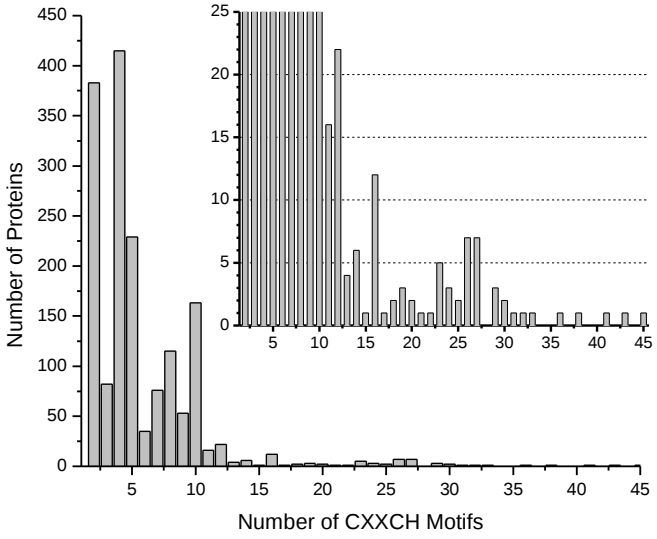


Figure 2. Average ratio between the length of MHC proteins and the number of CXXCH motifs as a function of the number of CXXCH motifs.

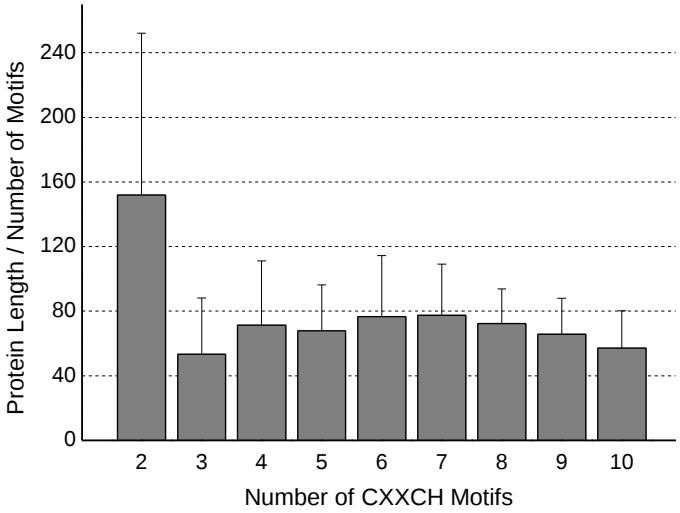


Figure 3. Clustering of the MHC sequences of the present dataset. The graph presents a twodimensional visualization of the results of the CLANS grouping of sequences. Axes units are arbitrary. The sequences are represented by vertices in the graph, and BLAST matches below the threshold E-value of 10^{-20} are shown as edges connecting vertices. Note that the overlap between different clusters is not meaningful.

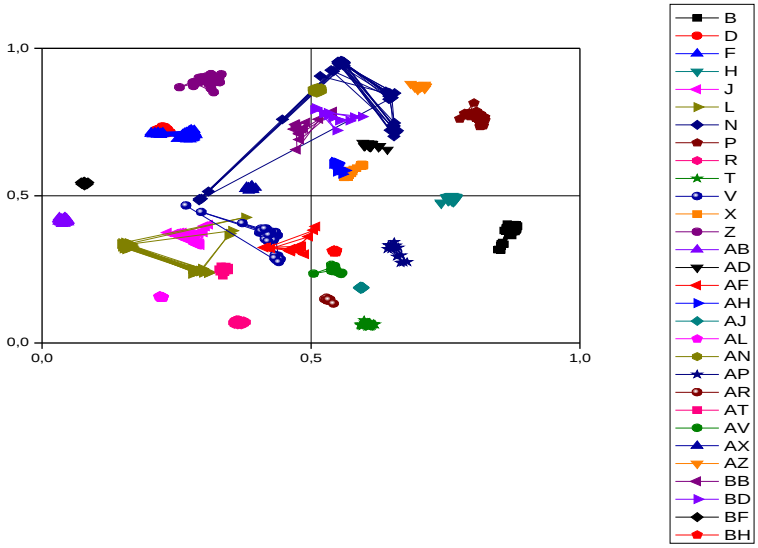
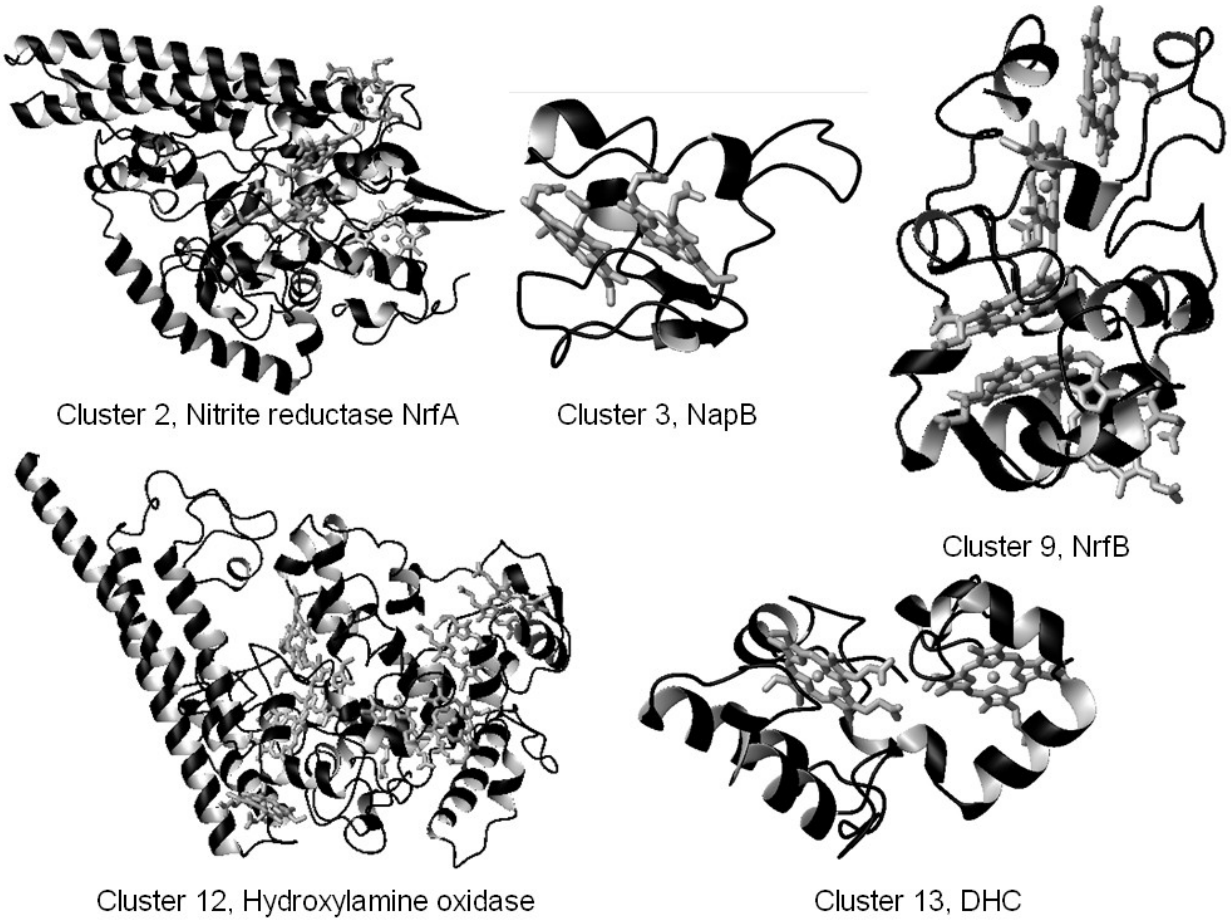


Figure 4. Three-dimensional structures of MHC's. The available structures of MHC's representing the largest clusters identified in this work are shown (cfr. Table 5) are shown as ribbons; the heme cofactors are in grey. The following structures were used: cluster 2, 1FS7 [23]; cluster 3, 1JNI [35]; cluster 9, 2P0B [28]; cluster 12, 1FGJ [36]; cluster 13, 2FWT [37].



CHAPTER 4

Benchmarking protocols for structure determination of proteins from chemical shift data

Introduction :

Chemical shifts are key to protein NMR spectroscopy not only because they allow separate observation of each ^{1}H, ^{13}C, and ^{15}N nucleus in the macromolecule, but also as they carry important information on the local conformation. For example, chemical shift data can be used to obtain secondary structure information [1], or indications on hydrogen bonding [2,3]. Protein structural information derived from chemical shifts, such as the backbone ψ torsion angles predicted by the program TALOS [4] is widely used in NMR structure determination, typically to complement conventional NOE distance restraints.

Recently, several computational approaches have been developed to use the NMR chemical shifts alone as input for protein structure generation [5,6,7,8]. These approaches, represented by CHESHIRE [9], CS-Rosetta [10] and CS23D [11], match the experimental chemical shifts of the backbone and $^{13}C^{\beta}$ atoms, which are commonly available at the early stage of the conventional NMR structure determination procedure, to a structural database to identify protein fragments with similar chemical shifts. Because the structural database of proteins for which actual NMR assignments are available remains relatively small, empirical relationships [12,13,14,15] are commonly used to “assign” chemical shift values to nuclei in proteins of known structure. Selected protein fragments are then used as input for a fragment assembly procedure, which also aims to optimize empirical energy terms related to hydrogen bonding, hydrophobic packing, etc., to generate an all-atom protein structure. These approaches have been evaluated for several proteins, with sizes of up to 15 kD and a wide variety of folds. When the method converges, protein models that compare well with experimental structures are often obtained. For CS-Rosetta in particular, data for structural genomics target proteins, obtained before the conventional NMR structure determination process was available [16], showed that CS-Rosetta could be a viable alternative for medium-size proteins [17].

To date, the chemical shift based structure determination methods have been evaluated for proteins with complete or nearly complete NMR chemical shift assignments. In practice, however, resonance assignments are often incomplete, and also may contain a small fraction of erroneous assignments. The conventional NMR determination strategy is sufficiently tolerant to be successful even in the presence of only 80–90% of the backbone sequence-specific assignments. Some recent

work is available addressing this point [18].

In order to establish a consistent evaluation of chemical shift-based structure determination protocols, also in the frame of the e-NMR project, we set out to build a benchmark of recently determined protein structures, lacking chemical shift assignments or homologs with available chemical shift assignment, to test the available protocols and investigate possible ameliorations.

Material and methods

Protein selection

We decided to start with a benchmark of ten different proteins. The selection was based on the following criteria:

only crystallographic structures with no or very distant homologs having an NMR assignment available were taken into account

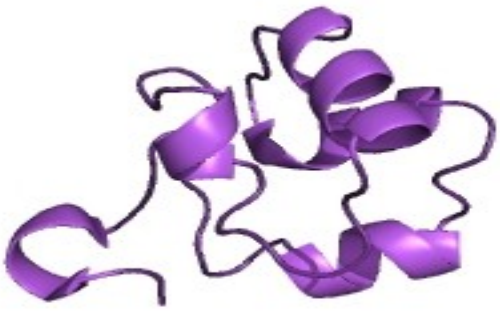
structures solved in 2007 or later

the protein should be in monomeric state

no ligands, metals or other heterogroups should be bound to the protein

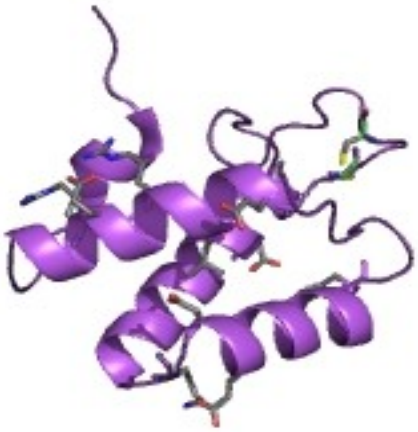
protein size smaller than 150 amino acids

A query was run on the PDB database by using the advance search option to include the above conditions. The absence of homologues already studied by NMR was checked manually. At the end, the following ten structures were selected: 2DUY , 2HL7 , 2J8B , 2PLZ , 2RHF , 2EHS , 2I5M , 2P5K , 2QNW , 3CA7, ranging in size from 35 to 70 amino acids These structures are depicted in Figure 1.

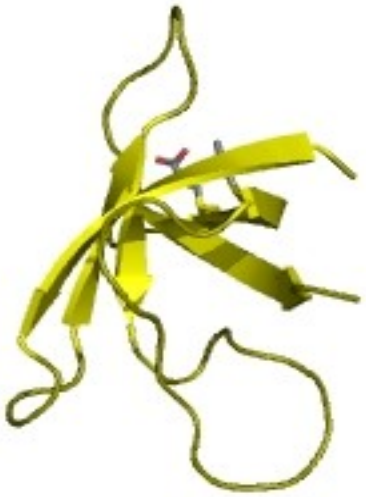


2duy

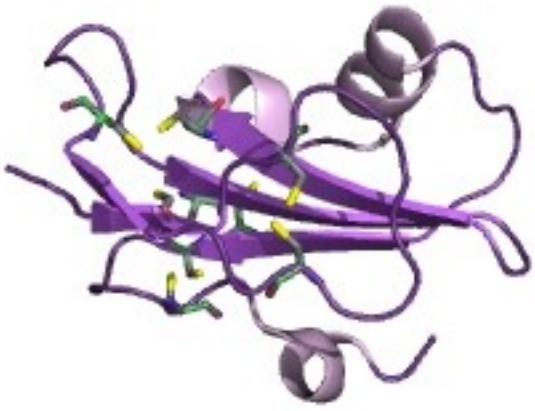
2ehs



2hl7



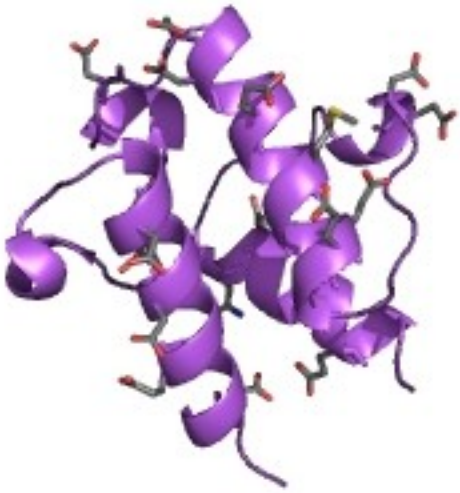
2i5m



2j8b

2p5k

2plz



2qnw

2rh5

3ca7

Figure 1 Structures of the ten selected proteins.

Chemical shift calculations

Chemical shifts were predicted for these proteins using both the shiftS and shiftX programs.

Brief description of SHIFTX: A computer program (SHIFTX) rapidly and accurately calculates the ^{1}H, ^{13}C and ^{15}N chemical shifts of both backbone and side chain atoms in proteins. The program is freely available as a web server at http://redpoll.pharmacy.ualberta.ca. SHIFTX uses a hybrid predictive approach that employs pre-calculated, empirically derived chemical shift hypersurfaces in combination with classical or semi-classical equations (for ring current, electric field, hydrogen bond and solvent effects) to calculate ^{1}H, ^{13}C and ^{15}N chemical shifts from atomic coordinates. The chemical shift hypersurfaces capture the effects of dihedral angle, side chain orientation, secondary structure and nearest neighbor, which cannot easily be translated to analytical formula or predicted via classical means.

Brief description of SHIFTS: SHIFTS takes a protein structure in Brookhaven (PDB) format, and computes proton chemical shifts from empirical formulas. It can also compute ^{15}N, $^{13}C\alpha$, $^{13}C\beta$ and $^{13}C'$ shifts in proteins, using a database based on DFT calculations on peptides. It is freely available as a web server at http://casegroup.rutgers.edu/qshifts/qshifts.htm

CS-Rosetta calculations and results

After calculating shift values we used the e-nmr web portal to use CS ROSETTA program for the calculation of the 3D structure of each protein. The calculation was repeated twice, both excluding and not explicitly excluding the corresponding experimental structure . We wanted to verify the effect of excluding the target structures from the database from which CS Rosetta selects fragments in order to ensure that the calculations did not consist of merely reassembling the structures from their own fragments with the basic Rosetta procedure. The quality of the results was assessed for each of the ten selected proteins by calculating the backbone RMSD between three structures: the experimental protein structure, the CS ROSETTA structure which did not explicitly excluded the experimental structure during calculation and the protein structure calculated by CS ROSETTA explicitly excluding experimental one.

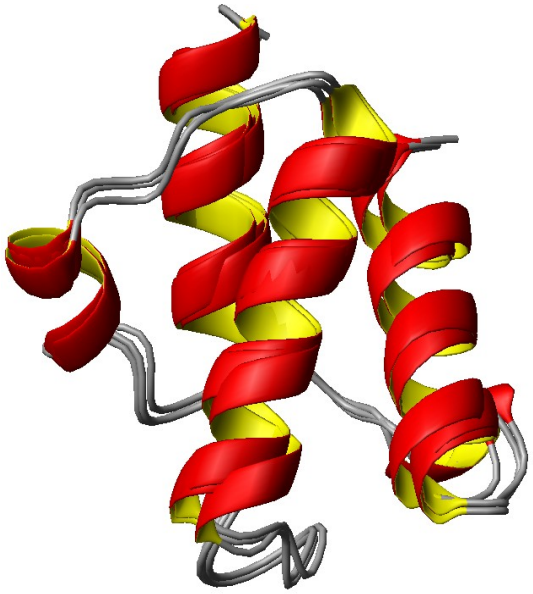


Figure 2 Experimental structure of 2EHS and structures calculated by CS ROSETTA using chemical shifts calculated by shiftX both excluding and including experimental structure . No. of amino acids = 75; RMSD between all 3 structures = 0.546 Å.

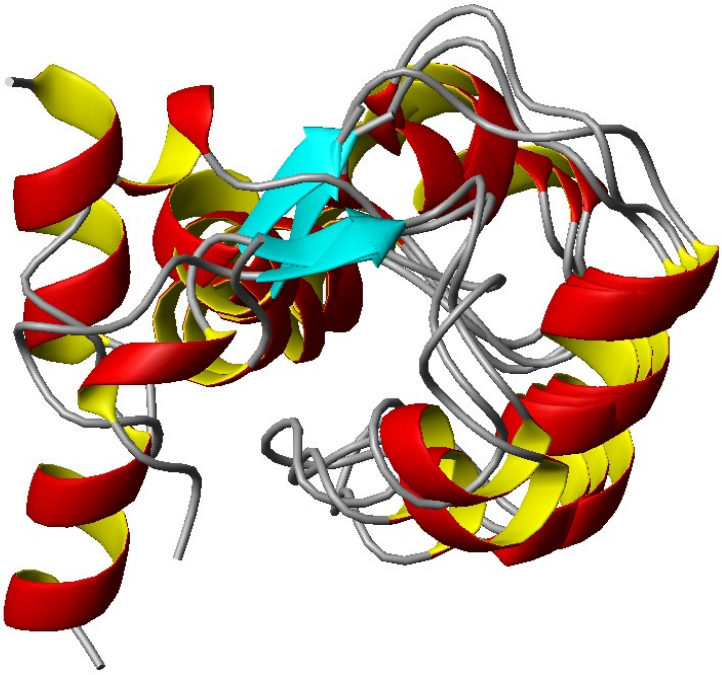


Figure 3 Experimental structure of 2DUY and structures calculated by CS ROSETTA using chemical shifts calculated by shiftX both excluding and including experimental structure. No. of amino acids = 65;RMSD between all 3 structures = 1.731 Å.

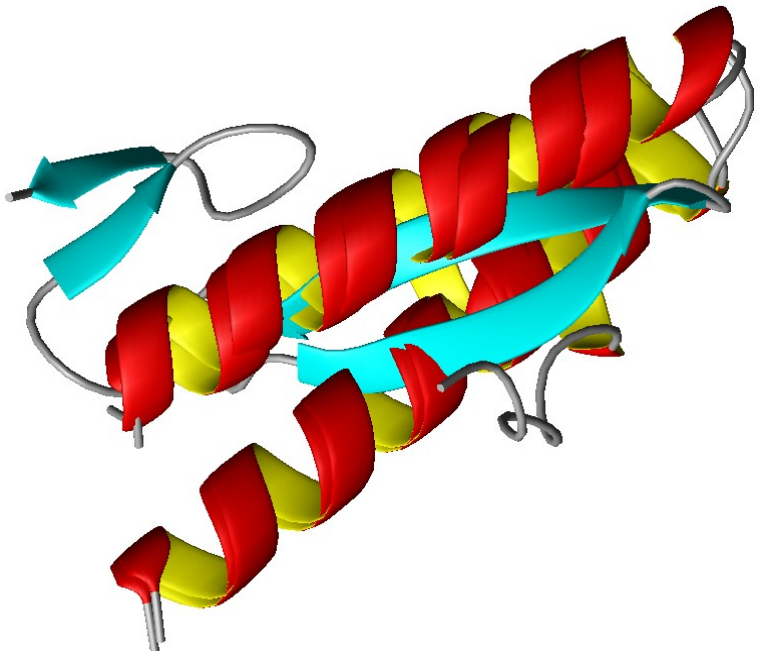


Figure 4 Experimental structure of 3CA7 and structures calculated by CS ROSETTA using chemical shifts calculated by shiftX both excluding and including experimental structure. No. of amino acids = 45;RMSD between all 3 structures = 5.536 Å

We found that CS ROSETTA can provide good results for proteins of 70 - 80 amino acids. This fact was demonstrated by RMSD values of proteins shown in figure 2 and 3. Interestingly, for structure 3CA7, which is of a protein of only 45 amino acids, we obtained a very high RMSD value of 5.63 Å (Figure 4). Further work is needed to assess the reasons for this.

References :

1. Wishart DS, Sykes BD and Richards FM (1991) Relationship between nuclear magnetic resonance chemical shift and protein secondary structure. J. Mol.Biol. 222: 311-333

2. Wagner G, Pardi A and Wuthrich K (1983) Hydrogen-Bond Length And H-1-Nmr Chemical-Shifts In Proteins. J. Am. Chem. Soc. 105: 5948-5949

3. Shen Y and Bax A (2007) Protein backbone chemical shifts predicted from searching a database for torsion angle and sequence homology. J. Biomol. NMR 38: 289-302

4. Cornilescu G, Delaglio F and Bax A (1999) Protein backbone angle restraints from searching a database for chemical shift and sequence homology. J. Biomol. NMR 13: 289-302

5. Cavalli A, Salvatella X, Dobson CM and Vendruscolo M (2007) Protein structure determination from NMR chemical shifts. Proc. Natl. Acad. Sci. U. S. A. 104: 9615-9620

6. Gong HP, Shen Y and Rose GD (2007) Building native protein conformation from NMR backbone chemical shifts using Monte Carlo fragment assembly. Protein Science 16: 1515-1521

7. Shen Y, Lange O, Delaglio F, Rossi P, Aramini JM, Liu GH, Eletsky A, Wu YB, Singarapu KK, Lemak A, Ignatchenko A, Arrowsmith CH, Szyperski T, Montelione GT, Baker D and Bax A (2008) Consistent blind protein structure generation from NMR chemical shift data. Proc. Natl. Acad. Sci. U. S. A. 105: 4685-4690

8 .Wishart DS, Arndt D, Berjanskii M, Tang P, Zhou J and Lin G (2008) CS23D: a web server for rapid protein structure generation using NMR chemical shifts and sequence data. Nucleic Acids Res. 36: 496-502

9. Cavalli A, Salvatella X, Dobson CM and Vendruscolo M (2007) Protein structure determination from NMR chemical shifts. Proc. Natl. Acad. Sci. U. S. A. 104: 9615-9620

10. Shen Y, Lange O, Delaglio F, Rossi P, Aramini JM, Liu GH, Eletsky A, Wu YB, Singarapu KK, Lemak A, Ignatchenko A, Arrowsmith CH, Szyperski T, Montelione GT, Baker D and Bax A (2008) Consistent blind protein structure generation from NMR chemical shift data. Proc. Natl. Acad. Sci. U. S. A. 105: 4685-4690

11. Wishart DS, Arndt D, Berjanskii M, Tang P, Zhou J and Lin G (2008) CS23D: a web server for rapid protein structure generation using NMR chemical shifts and sequence data. Nucleic Acids Res. 36: 496-502

12. Cornilescu G, Delaglio F and Bax A (1999) Protein backbone angle restraints from searching a database for chemical shift and sequence homology. J. Biomol. NMR 13: 289-302

13. Neal S, Nip AM, Zhang HY and Wishart DS (2003) Rapid and accurate calculation of

protein H-1, C-13 and N-15 chemical shifts. J. Biomol. NMR 26: 215-240

14. Kontaxis G, Delaglio F and Bax A (2005) Molecular fragment replacement approach to protein structure determination by chemical shift and dipolar homology database mining. Meth. Enzymol. 394: 42-78

15. Shen Y and Bax A (2007) Protein backbone chemical shifts predicted from searching a database for torsion angle and sequence homology. J. Biomol. NMR 38: 289-302

16. Shen Y, Lange O, Delaglio F, Rossi P, Aramini JM, Liu GH, Eletsky A, Wu YB, Singarapu KK, Lemak A, Ignatchenko A, Arrowsmith CH, Szyperski T, Montelione GT, Baker D and Bax A (2008) Consistent blind protein structure generation from NMR chemical shift data. Proc. Natl. Acad. Sci. U. S. A. 105: 4685-4690

17. Gryk MR and Hoch JC (2008) Local knowledge helps determine protein structures. Proc. Natl. Acad. Sci. U. S. A. 105: 4533-4534

18. J Biomol NMR. 2009 Feb;43(2):63-78. Epub 2008 Nov 26.De novo protein structure generation from incomplete chemical shift assignments. Shen Y, Vernon R, Baker D, Bax A

Conclusion and future prospectives :

In the first part of our study the occurrence and properties of $P_{1B-1,2}$-type ATPases, and partly, of their partner metallochaperones were investigated . $P_{1B-1,2}$-type ATPases may contain multiple metal-binding domains (MBDs) in the N-terminal cytoplasmic part of their sequence. The number of these domains tends to be higher in more complex organisms. For the human MNK and WND proteins, which contain six MBDs, the two domains closest in sequence to the transmembrane part of the protein have been sown to have a different role than the other four . We demonstrated through sequence analysis across a large dataset of organisms and molecular dynamics that in the majority of organisms, these two domains tend to be structurally independent and can reorient one with respect to the other. Therefore, it appears very likely that they can exert their function independently. The MBDs have a role in regulating the ATPase activity through interactions with the other protein domains. The presently observed relative flexibility can thus be instrumental in optimizing the regulation of the activity in multi-MBD ATPases.

Metallochaperones instead are single-MBD proteins that typically deliver copper(I) ions to partner $P_{1B-1,2}$-type ATPases or mercury(II) ions to te MerA reductase in the mercury detoxification system. However, also when involved in the latter process, metallochaperones are detected nearly exclusively in organisms encoding $P_{1B-1,2}$-type ATPases as well, possibly indicating that they evolved originally to interact with the ATPases and then adapted to scavenge also mercury(II). We also described several hints suggesting that metallochaperones can interact also with P-type ATPases lacking MBDs. Conversely, and much more commonly, there are several $P_{1B-1,2}$-type ATPases that likely function in the absence of a partner metallochaperone, analogously to ATPases transporting divalent cations. For these systems, in addition to modulating the overall enzymatic activity, the reciprocal mobility of MBDs could be important to tune the selectivity and/or the affinity for the substrate, which could be the metal ion complexed to either an organic molecule or an unidentified cytoplasmic metallochaperone. There are several perspectives for this work, including: i) the simulation of interaction between the MDBs of ATPases and the metallochaperones, including quantitative estimates of metal-protein interaction energies and their change along the reaction coordinate; ii) the modeling of the metal-binding sites within the transmembrane regions; iii) the simulation of interaction with different metal ions, to understand the possible determinants of metal ion selectivity.

In the second part of our PhD work, we identified 1659 multi-heme cytochromes *c* (MHC's) or unassigned proteins containing multiple CXXCH motifs in 258 organisms (out of 594 analyzed). The presence of MHC's correlated well with, and thus can be taken as a good indicator of, an organism's

ability to take up or synthesize heme; we consequently identified two cases, that of *Haemophilus ducreyi* and of *Streptococcus thermophilus*, where the presence of MHC's in the genome may suggest that they have an uncommon or highly divergent heme uptake pathway. The most common number of heme-binding motifs in a sequence was four (25%) and two (23%), followed by five (13%) and ten (9.8%). The average protein:heme ratio was relatively similar for all MHC's, except di-heme proteins, regardless of the number of motifs at around 60 ± 30. However, the variability also within a group of MHC's with the same number of motifs was relatively high. Only within individual functional families the ratio exhibited little variability. This observation could thus be useful to corroborate functional assignments of novel MHC's. The detailed comparison of the MHC sequences retrieved provided various hints that can be useful to direct future experimental work in the field. For example, we identified homologues of the NRfA nitrite reductase where the CXXCK motif of the catalytic heme is replaced by a fifth 5 CXXCH motif. In addition, we identified sequence similarity between deca-heme MHC's of the MtrA family and penta-heme MHC's of the NrfB family. As a general consideration, it appears that the amount of structural information currently available for MHC's is limited with respect to the diversity of this broad class of metalloproteins, and even some of the largest MHC families lack a structurally characterized member. This limits the possibility e.g. to perform systematic structural modeling of MHC's because of the poor selection of templates, which would affect key factors such as the reciprocal position and orientation of the heme cofactors and their ligands. Experimental efforts in the structural investigation of MHC's are thus warranted. A possible perspective could thus be that of providing priorities for structural determination, e.g., by evaluating the leverage of each new MHC structure (that is, the number of new high-quality structural models can be derived from each new structure). A more in-depth analysis of the domain composition and possible intramolecular interactions of very large MHC's could also constitute an interesting perspective in terms of defining the modularity of these proteins.

In the third part we aimed at establishing a benchmark for the evaluation of the performance of chemical shift-only methods for the calculation of protein structures. We initially focused on CS-Rosetta, and a set of ten proteins of known X-ray structure, but lacking any NMR characterization, to be used for calculations using simulated chemical shift data. The perspective in this work is to extend calculations to other programs and use an extended set of structures in the benchmark.

One Page Description of thesis :

Title of thesis: Bioinformatics of metal binding proteins and genome wide analysis .

Specific area of Discipline : Structural Biology .

5 key words describing my work : bioinformatics; metalloproteins; cytochromes; copper; proteomics

Brief description of my work :

During my PhD training I applied theoretical, bioinformatic methods to address problems in structural biology starting from amino acid sequence going to enzyme function through structure determination. The first two years of my training were focused on protein sequence analysis and their comparison, homology modelling, molecular dynamics. In this context, I developed a computational approach to identify unprecedented members of known protein families on the basis of HMM model databases. The performance of this method was tested by using the PDB database as a “gold standard”. In the third year I applied Nuclear Magnetic Resonance (NMR) chemical shift calculation methods to validate methods for the determination of the tertiary structure of proteins. An objective of these calculations is to develop approaches that reduce the amount of time needed for solution structure determination of proteins based on NMR data. It is likely that this approach in the future will be applied to study protein-protein as well as protein-ligand interactions.

The sequence-to-function analyses of protein families were applied on two different biological systems: i) an ubiquitous intracellular copper transport parthway; ii) multi-heme c-type cytochromes. In i), we focused on two protein partners: a soluble small (ca. 70 amino acids) copper(I) binding protein (called a metallochaperone) and P_{1B}-type ATPases, which can transport copper(I) ions across membranes at the expenses of ATP hydrolysis. The latter may contain multiple copper(I)-binding domains, separated by a flexible protein linker. Molecular dynamics simulations showed large differences in the behavior of systems having different linker lengths. Indeed, the inter-domain interaction energy was inversely proportional to the length of the linker. S-order parameter values were higher for smaller linker lengths. This is clearly demonstrating that the reciprocal motions of domains was also dependent on the length of the linker. These results were comparable for both apo and holo forms, indicating that the copper(I) ion does not playing contribute significantly to modulating the inter domain interactions, which are only dependent on the length of the linker region. This has implications for the mechanism of copper(I) transfer to the ATPase transmembrane binding site. For ii), we investigated the genomes of the 594 prokaryotic organisms sequenced till date, thereby analyzing a total of 19,00,966 protein sequences. Of these, 3783 contained two or more CXXCH motifs, which are responsible for heme attachment to the protein, and therefore were potential multi-heme cytochromes.

After a systematic search of HMM models of two databases (Pfam and SUPERFAMILY), we retained 1659 proteins as true MHC's, and remained with 67 unassigned proteins.

For our NMR application, we have focused on protein structure calculations based on only the amino acid sequence and NMR chemical shifts. It involves homology modeling based on the sequence-structure alignment, SPARTA , Cheshire ,CS - ROSETTA, cs23d and PSVS web servers complemented by locally written programs.