UNIVERSITÀ
DEGLI STUDI
FIRENZE

DOTTORATO DI RICERCA
INTERNATIONAL DOCTORATE IN STRUCTURAL BIOLOGY

CICLO XXXII

COORDINATOR Prof. Claudio Luchinat

APPLICATION OF MOLECULAR DYNAMICS TO THE UNDERSTANDING OF METAL-BINDING MACROMOLECULES AND THEIR ADDUCTS

Settore Scientifico Disciplinare CHIM/03

PhD student
Dott. Davide Sala

Tutor
Prof. Antonio Rosato

Coordinator
Prof. Claudio Luchinat

November 2016 – 2019

This thesis has been approved by the University of Florence, the University of Frankfurt and the Utrecht University

ACKNOWLEDGEMENTS

First and foremost, I would like to thank my PhD tutor and mentor, Prof. Antonio Rosato for his constant support during my work period at CERM. He has been always present every time I needed a word of encouragement, thereby demonstrating his inestimable human worth other than of being an excellent guidance. The same can be said of the other member of our research group and my office mate, Dr. Andrea Giachetti, whose patience and help have never been missed.

I would like to express my gratitude also to the directors of CERM Prof. Lucia Banci and Prof. Claudio Luchinat for giving me the opportunity to conduct my doctoral research work in this excellent research infrastructure.

Furthermore, my gratitude is due to my colleagues at CERM. In particular, Dr. Silvia Ciambellotti, Dr. Linda Cerofolini and to their respective supervisors Prof. Paola Turano and Prof. Marco Fragai for their precious collaboration in two relevant works of my PhD.

My time in CERM would have been less pleasurable without the company of the metabolomics group for having shared many afternoon breaks, as well as the fantaCERM group for three years of battles and of the other guys with whom I spent funny moments inside and outside the workplace.

Lastly, I am deeply indebted to my family and to Giulia, that actually are the same thing, for their support and unconditional love without which I would surely be lost.

TABLE OF CONTENTS

LIST OF FIGURES

Chapter 1

PHD PROJECT

1.1 Introduction

1.1.1 Role of Metals in Biology

More than 90 of the 118 elements in the periodic table are metals. Each metal has unique chemical and physical properties that can play crucial roles in biological reactions. In this regard, metals play an important role since the era of the so-called "RNA world". Despite the fact that the concentrations of most of these metals in the prebiotic ocean are not known, it is believed that evolution took advantage of the most plentiful ones to develop numerous and fundamental biological processes[1]. As essential metals for all living organisms that are abundant in both the environment and *in vivo*, we find sodium (Na^+) and potassium (K^+) of the first group and magnesium (Mg^{2+}) and calcium (Ca^{2+}) of the second group[2]. Sodium and potassium ions play major roles in biological systems, from maintaining the membrane potential, a mechanism underlying muscle contraction and nerve impulse transmission, to osmotic pressure regulation in cell[3]. Magnesium is mainly involved in the maintenance of nucleotides stability and ATP activation[4]. Calcium is the most abundant metallic element of the human body and takes part in many crucial physiological activities as signal transduction, muscle contraction and relaxation, neurotransmission, bone formation, mitosis, secretion or can simply act as enzyme stabilizer and others[5]. The other essential metal ions are also known as "trace" elements because they act in low concentration and are cytotoxic in high amount. Molybdenum (Mo), tungsten (W), manganese (Mn), iron (Fe), cobalt (Co), nickel (Ni), copper (Cu), and zinc (Zn) are all "trace" elements and belong to the d-block of the periodic table. In general, essential trace elements take part in different biological processes depending on their properties. For instance, the propensity of iron to make complexes and exchange electrons is extremely important in hemoglobin formation underlying oxygen transport and oxido-reduction reactions, respectively[6]. Again, zinc is a fundamental cofactor for many enzymes involved in metabolism and cell growth[7]. The maintenance of an optimal cellular or subcellular concentration requires the presence of mechanisms to regulate the uptake, storage and excretion of metal ions[8]. The balance of these processes is called metal homeostasis. Several and severe pathologies are directly related to disfunctions in metal homeostasis regulation[9]. Furthermore, there are also a number of non-essential metals that are present in many organisms and affect physiological processes. They are usually toxic even in very small concentration, hence specific homeostatic systems regulate their availability. Finally, some metal-dependent physio-chemical properties of some non-biological metals can be specifically exploited to better understand physiological processes (for instance in imaging) or to therapeutic effects[10].

1.1.2 Metalloproteins

Proteins that require one or more metal ions to perform their biochemical function are called metalloproteins. To date, metal ions are critical to the activity of up to an estimated one-third of all proteins in most of the organisms. Metals can play different functional roles in the interaction with proteins. As mentioned above, transport of electrons is a crucial activity that often relies on transition metal cofactors. As catalytic cofactors, it was estimated that 41% of enzymes contain metal ions at the catalytic centre and thus are named metalloenzymes[11]. As structural cofactors, metals ions are mainly required for correct protein folding or to stabilize the folded functional state, in both monomeric conformation and mediating protein-protein interactions in complex formation. Furthermore, a big portion of metalloproteins are directly involved in the maintenance of an optimal intracellular metal concentration i.e. in the metal homeostasis regulation. For instance, metallochaperones are employed for intracellular metal trafficking and for controlling the delivery of the appropriate metal ion to the target metalloprotein in a specific cellular compartment and through specific protein-protein interactions[12]. Metalloproteins usually use a specific metal ion to perform their function that, in most of the cases, cannot be substituted by other metals, thus requiring a metal-specific fine-tuned homeostasis regulation system[13]. In this context, a crucial role is played by metal-responsive transcriptional regulators, also known as metalloregulatory proteins or metal sensors[14]. The specific binding of the appropriate metal on the regulator triggers a rearrangement of protein structure and/or dynamics that modulates its affinity/specificity to DNA. The binding on the DNA in turn triggers up- or down-regulation of proteins involved in metal homeostasis or that require metal to their biochemical activity. Thus, this class of metalloproteins can modulate gene expression in response to the metal intracellular concentration. A quick mention deserves also mellathioneins, a family of metalloproteins enriched of sulfhydryl groups mainly involved in the protection against metal toxicity and oxidative stress and in the maintenance of essential metal ion homeostasis[15].

1.1.3 Iron and Zinc

Iron and zinc are the first and second most abundant trace element in the human body, respectively. They participate in many important physiological processes. Thus, disorders in their metabolism can lead to severe diseases. In particular, iron propensity to catalyze the propagation of ROS (reactive oxygen species) and the generation of highly reactive radicals (such as the hydroxyl radical) through Fenton chemistry requires a tight control of its availability by diverse mechanisms at different levels[16]. Iron exists *in vivo* in two different oxidation states, the ferrous +2 and the insoluble ferric +3. Iron-containing systems can bind the metal as single ions or complexed with minerals and other cofactors, such as in ferritin, heme or iron-sulfur clusters.

In contrast to iron, zinc is ubiquitous within cells and can be found in cellular compartments approximately with the following distribution: 50% in the cytoplasm, 30-40% in the nucleus and the remaining 10% in plasma and organelle membranes[17]. Zinc is required by approximately 10% of human proteins[18]. As catalytic or structural cofactor it contributes to many physiological functions, from cell growth, development and differentiation, to processes such as gene expression, DNA synthesis, hormone storage and release, neurotransmission, memory and apoptosis[19]. Specific protein families are responsible of zinc transport inside and outside cells and cellular compartments. The removal of zinc from cell or from organelles is due to the Cation Diffusion Facilitator (CDF) proteins ZnT/Slc30. On the contrary, ZIP/Slc39 proteins uptake zinc in the cytoplasm from the extracellular space. There are at least 10 ZnT and 14 Zip transporters in human cells. Both ZnT and ZIP family members feature unique tissue-specific expression and respond differentially to zinc overload or deficiency.

1.2 Aim of the Project

My doctoral project was focused on the investigation of structural properties of metalloproteins *in silico.* In this regard, the large diffusion of powerful and cost-effective GPGPU units has allowed the computational exploration of a number of biological processes using techniques based on molecular dynamics (MD). In particular, the research projects addressed during my PhD can be grouped in two main chapters: i) structure determination of proteins and ii) role of metal ions in metalloprotein dynamics.

In the context of proteins structure prediction, three studies have been carried out. In the first one, the PseudoContact Shifts (PCS) module of the AMBER MD package was extended to handle multiple PCS datasets simultaneously and to refine NMR structures with the restrained molecular dynamics (rMD) method. In the second study, NMR-derived ambiguous contacts were simulated and provided to predictors (in addition to real data) of the CASP13 NMR-assisted competition. Eventually, evaluation of the models generated by NMR-assisted competitor groups was performed relative to the models generated using baseline methods and relative to the models generated using non-assisted regular methods. In the third study, we took advantage of coevolution analysis to develop a protocol for the prediction of homo-oligomeric complexes from ambiguous NMR-data.

In the chapter "role of metal ions in metalloprotein dynamics" are grouped the four projects based on MD methods. In the first one, the metal-dependent folding process of rubredoxin was investigated using accelerated MD. In the second project, the iron release mechanism from human ferritin was explored as a function of pH. In the third study, zinc-dependent conformational transitions of the YiiP transporter were sampled. Finally, in an ongoing project the conformational motions of two variants of the human zinc transporter ZnT8 associated with type-2 diabetes are under investigation.

1.3 Methodological Aspects

1.3.1 Molecular Dynamics

An important step in the knowledge of biological systems is the characterization of their three-dimensional (3D) or quaternary structure and dynamics. Experimental methods for structural determination of molecules such as X-ray crystallography, NMR and Cryo-EM (and others) very often can resolve proteins structures in the corresponding global energetic minimum but less often the transient conformations corresponding to local energetic minima. However, transient conformations can provide important contribution in the description of molecular mechanisms, therefore they are fundamental for getting a complete overview of biological processes[20]. In this regard, the computational approach most suitable for the investigation of dynamics properties is the molecular dynamics (MD) method. Molecular dynamics can simulate the evolution of a system over time based on the forces (expressed as force-fields) affecting system particles. Basically, MD simulations solve Newton's equations of motion (equation 1) for a system of N interacting atoms:

$$m_i \frac{\partial^2 r_i}{\partial t^2} = F_i \;\; ; \;\; i = 1, \cdots, N \qquad (1)$$

where m_i, r_i and F_i are the mass, the position and the force acting on the i-th atom, respectively. Because of the high number of atoms typically involved in these systems, MD simulations are performed with the use of classical force-fields $[V(r_1, \cdots, r_N)]$, which are functions of the atomic positions only:

$$F_i = -\frac{\partial V(r_1, \cdots, r_N)}{\partial r_i} \;\; ; \;\; i = 1, \cdots, N \qquad (2)$$

Equation 2 is integrated in short time steps, providing an atomistic view of the evolution of the system over time (the so-called trajectory). In fact, MD simulations are worthwhile to obtain an atomic-level description of protein systems and their mobility under different conditions such as during conformational transitions, protein folding or transport of molecules through the cellular membrane[21]. In the context of metalloproteins, MD can be effectively used in the investigation of metal perturbation on systems. By analyzing the trajectory after its initial "equilibration" phase, during which system properties reach target mean values e.g. temperature, pressure or water molecules and ions distribution around the macromolecule, the macroscopic properties of interest can be extracted from the simulation. Furthermore, MD simulations can be harnessed to improve the structural features of protein conformations through restrained molecular dynamics (rMD).

1.3.2 Classical Parametrization Strategies for Metal Ions

A common issue in performing MD simulations of metalloproteins is the parametrization of metal ions and/or metal centers. For instance, metal ions or clusters covalently bond to proteins need a reasonable distribution of their electronic densities on single particle charges. Furthermore, common force-fields used in MD simulations do not include parameters for d-block metals. Thus, using different approaches a number of models have been developed during years with the goal to obtain the best possible (closest to natural behavior) reproduction of metal properties and their effects on the interacting atoms[22]. In the following sections the most popular parametrization strategies for metal ions are discussed.

The Nonbonded Model

The most widespread potential for two atoms not covalently interacting is undoubtedly the sum of the Van Der Waals and electrostatic terms[23] as described in Equation 3. The electrostatic function for two atoms *i* and *j* is given by their partial charges Q_i and Q_j, that for metal ions is usually assigned as integer charge corresponding to their oxidation state, separated by a distance r_{ij} in which a dielectric constant ε_0 defines the free space permittivity. The function describing Van Der Waals energy contributions is known as Lennard-Jones function (LJ), where ε_{ij} is the well depth and measures how strongly the two particles attract each other, σ_{ij} is the distance at which the intermolecular potential between the two particles is zero and r_{ij} is the distance separating the two atoms. Because in a system with N atom types there will be as many as $N(N+1)/2$ unique atom pair types, which require a considerable parametrization effort, different rules have been introduced to derive *ij* pair coefficients from each individual atom type. In this way, only two parameters per atom type (e.g., ε and σ) are needed ($2N$ in total). The resulting potential is a good approximation especially for low electronegative monovalent ions, where the polarization and charge transfer effects are small[24].

$$U(r_{ij}) = E_{ele} + E_{VDW} = \frac{Q_iQ_j}{4\pi\varepsilon_0 r_{ij}} + 4\varepsilon_{ij}\left[\left(\frac{\sigma_{ij}}{r_{ij}}\right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}}\right)^{6}\right] \qquad (3)$$

The parametrization of metal ions can follow different strategies with the common aim to reproduce one or more structural, kinetic and dynamic properties in addition to thermodynamic quantities. In particular, the three key parameters that should be accurately reproduced are: i) the hydration free energy (HFE) ii) the coordination number (CN) and iii) the ion-oxygen distance (IOD) of the first water shell. Among the parametrization strategies that can be adopted with MM computational methods, the force field or single point energy calculations, Monte Carlo or MD simulations are commonly used. Parameters derived by fitting against specific properties should be tested before being applied, since the derived model can be affected by transferability problem if it is used on bigger or more complex systems where multiple atom types are present simultaneously. Instead, metal ions parameters derived from QM fitting typically consider solely the first water molecule, therefore neglecting the many-body effects arising from fitting on multiple molecules. As a result, ion-water interactions are often overestimated especially for the HFE and CN properties. This can be partially compensated by using small basis sets, except for highly charged metal ions for which the many-body effect is too high. In the latter case, various approaches can be deployed to circumvent the many-body issue: (A) introducing many-body terms; (B) working with an effective potential; (C) working with a combined model. Parameterizations derived from experimental data are often effectively based on free energy. Thus, free-energy changes can be calculated using different computational methods such as free-energy perturbation[25] (FEP), thermodynamic integration[26] (TI), MM Poisson–Boltzmann surface area[27] (MM/PBSA), MM generalized Born surface area (MM/GBSA) and linear response approaches[28].

In general, parametrization based on the LJ model are affected by an error the increases with the square of the ion charge. Thus, a good agreement with experimental values is particularly difficult to achieve for all the parameters simultaneously (HFE, CN and IOD) in the case of highly charged ions. A straightforward way to reduce errors is to refit specific parameters to satisfy individual experimental features. However, this approach usually results in some parameters reproduced with high accuracy but big errors in the others. Furthermore, classical approaches neglect charge-induced dipole and dipole-induced dipole interactions, a drawback particularly relevant for multivalent ions. To correct this, a new 12-6-4 LJ model was proposed in 2014 and applied to parametrize various metal ions in conjunction with three common water models[29–31]. The new 12-6-4 model has an additional term representing the ion-induced dipole interaction. Similarly, classical force fields adopting single point charges that are kept fixed during simulations lack of polarization effects. In this case, the charge scaling (CS) approach can be applied to introduce long-range screening effects into account as in the MD electronic continuum (MDEC) approach[32]. A model of water based on the CS approach features fluctuating charge and, because it does not introduce additional terms, has the advantage to be suitable for use in conjunction with classical force fields while preserving computing performances[33]. Despite CS models are unable to reproduce correctly the microscopic environment, they can be used to predict some statistical properties such as different polarization of water molecules binding metal ion at the macromolecular surface and in the bulk phase.

The Bonded Model

Together with the nonbonded terms, the functional used in MM-based calculations is formed by the bonded interactions illustrated in the Equation 4. In the first three terms we find bond, angle and dihedral interactions described as follow: for each group of bonds, the k_r (force constant) and r_{eq} (equilibrium length), for each group of angles, k_θ (force constant) and θ_{eq} (equilibrium value), for each group of torsion angles, V_n (energy barrier), n (periodicity), and γ (phase). The latter two terms define electrostatic, with atomic charges, and VDW interactions that largely contribute to the quality of the force field. The most widely used classical force fields for biological systems are based on the same functional shown in Equation 3, that can slightly differ in the LJ term. The parameters for dihedral and improper torsion angles have distinct minima depending on the repulsions between bond electrons[34]. However, in the case of metal ions all angle values are considered accessible at physiological temperature, therefore the corresponding energy barriers are normally set equal to zero[35].

$$U = \sum_{bonds} k_i (r - r_{eq})^2 + \sum_{angles} k_\theta (\theta - \theta_{eq})^2 + \sum_{torsions} \sum_{n=1,2,3,4} V_n \left[1 + \cos(n^\phi - \gamma) \right] + E_{ele} + E_{VDW} \quad (4)$$

Even though modern force fields can have high power and cross terms[36,37], most of the classic force fields are still based on a harmonic representation of the bonded and angle terms, making the parameters transferability feasible and easy. Force field parameters are usually derived based on agreement with QM or experimental data and then transferred on larger target systems. This basic transferability assumption is a good approximation for equilibrium bond and angle distances, whose values are often similar in different environments. However, in QM calculations the bonded and nonbonded interactions are coupled together, making it difficult in some case to disentangle the two contributions.

Among the QM methods to derive force constants, the potential energy surface (PES) scanning approach is one of the simplest. Despite simplicity, every bond and angle that is part of the metal environment requires a PES scan, making the computational cost expensive. This approach was applied to develop a force field for various heme species[38]. Instead, the Z-matrix method relies on the computation of the Hessian matrix that usually is executed calculating the derivative of the energy analytically and in combination with the DFT level of theory[39]. This method has some well-known drawbacks, we list here the most relevant ones. First, it is dependent on the internal coordinates chosen. Second, it considers only harmonic terms, thus requiring corrections of final force constants. Finally, the harmonic potential used prevent the application of calculated dihedral force constants within modern force fields. Similarly, the Seminario method calculates force constants using a sub-matrix of the Cartesian Hessian matrix[40]. However, these constants incorporate the influence of the environment and may not be suitable for isolated bonded terms. To date, several metal ion parameters are present in literature based on the use of this method within the MCPB toolkit[41]. Furthermore, it was recently used for the development of a python based metal center parameter builder[42] and a visual force field derivation toolkit[43] (VFFDT). A mention deserves also the automated parametrization method (APM), an approach to derive parameters directly from the parameter space. APM has the advantage of not presenting the double counting issue of bonded and nonbonded interactions that affect other methods as the Z-matrix approach. Nevertheless, the number of possible combinations raises exponentially with the parameter space dimension. This can be partially managed dividing the parameter set in subsets or performing the parameter optimization as a geometry optimization. A number of free programs are available to perform APM, such as Parmfit[96] and ForceBalance[97].

Charge parameters for classical force fields are in most of the cases inferred by QM calculations. Among them, the most used methods to derive partial charges of metal sites is the electrostatic potential fitting (ESP)[46] and in particular the restrained charge fitting (RESP)[47]. The former method is applied in combination with algorithms able to pick ESP points and their spatial location for fitting. For instance, the CHarges from Electrostatic Potentials (CHELP)[48] approach uses points that are symmetrically distributed on four spherical shells from the VDW surface separated by 1 Å, with 14 points on each shell. Similarly, the CHarges from ELectrostatic Potentials using a Grid-based (CHELPG)[49] strategy picks the points between the VDW surface of the molecule and a surface 2.8 Å away based on a cubic grid where the points are separated by 0.3 Å. The RESP method using the DFT functional with the 6-31G* showed excellent performance in calculating the atomic charges of metal sites, despite sometimes an effective core potential can be preferable (for instance with negative charged clusters[50]). Both the mentioned methods are unable to take in account charge fluctuations due to conformational dynamics around the metal site. However, the metal-protein bonds usually confer rigidity to the protein region, making the calculation on a single conformation a reasonable approximation.

In addition to QM calculations, atomic charges can be inferred by empirical methods, by fluctuating charges methods (FQ) and by experimental derivation methods. However, all these approaches have limited applicability for different reasons. Empirical charge methods are time consuming and perform poorly in addressing transition and excited states. The FQ model is difficult to apply on macromolecules. Finally, the presence of a structure with very high resolution and low B-factors allows to derive charges from XRD experiments.

The Polarizable Model

Highly charged metal sites are affected by strong local polarization effects that are impossible to be accurately reproduced with fixed partial charges. As consequence, the polarizable model was developed to account for charge delocalization as a function of the coordination environment. The three most used polarization models are, the FQ model[51], the Drude oscillator[52] and the induced dipole model[53]. Nowadays, MD simulations with polarization models have a reasonable computational cost if performed in combination with extended Lagrangian algorithms[54]. However, the applicability of these models is still limited.

The Cationic Dummy Atom and the Combined Model

The cationic dummy atom model (CDAM) representation splits the total charge and mass of the metal atom in a number of dummy atoms around the metal core that depends on the coordination number of the metal ion (Figure 1). This model was originally developed for Mg^{2+} and then extended for other cations where it showed excellent performances in reproducing the native coordination of ions in solution[55]. Despite it might reduce the energy of the interaction between metal and amino acids, CDAM was applied also in nonbonded models to improve experimental HFE and IOD parameters together[56].

Alternatively, the combined model incorporates the first water shell around the metal core in a combined unit exploitable in MM calculations[57]. This approach has the advantage of immediately including many-body effects in the first water shell, while taking into account electronic state crossing. However, it works well only for highly charged metal ions (+2 or higher) and in short simulations because of long mean residence time of water molecules in the first shell.

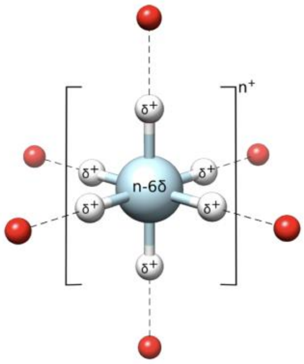


Figure 1. Example of an octahedral dummy model coordinating water molecules[58].

Chapter 2

PROTEINS STRUCTURE PREDICTION

2.1 Introduction

2.1.1 NMR-based Structure Determination of Metalloproteins

Function of metalloproteins strongly depends by the physical-chemical properties of the bound metal ion/cluster, in turn affecting the coordination chemistry of the metal center. In this regard, the investigation of the metal coordination by NMR can be performed only on the metals with a magnetic susceptibility. Unfortunately, a significant number of biologically active metals lack of nuclear magnetic susceptibility (e.g. ^{67}Zn, ^{43}Ca) or have unpaired electrons (e.g. Mn^{2+} and Cu^{2+}), making the investigation of the metal coordination unsuitable. In some cases, this issue can be addressed by replacing physiological metals with NMR-active metals, for instance using ^{113}Cd in zinc binding proteins[59]. In the case of diamagnetic metals, the information collected mainly derive from the metal-dependent NMR chemical shift perturbation of the environment around the metal, both through chemical bonds and through space. Stoichiometry of metal binding and its affinity can be estimated simply by mapping the chemical shift perturbations on the protein structure. However, in most of the cases this method alone is not sufficient to derive the 3D structure of a metalloprotein. Thus, the NMR structure determination of metal-binding proteins usually relies on the combination of NMR with other techniques such as SAXS or X-ray crystallography. Interestingly, despite the complexity of the approach, quantum mechanical/molecular mechanical (QM/MM) molecular dynamics simulations restrained by NMR-derived data has proved effective in the structural determination and refinement of zinc metal binding sites in absence of other experimental information[60].

In the case of metalloproteins binding paramagnetic ions, additional information can be collected from the interaction between the magnetic nuclei of the protein and the unpaired electron spin density of the metal. This kind of interactions typically cannot be used to derive the coordinates of the nuclei of the protein residues that coordinate the paramagnetic metal because of signal broadening or because of negligible contributions. However, paramagnetic interactions between residues distant in space can be exploited as long-range restraints that, in combination with classical short-range restraints as NOEs, can be applied in structure refinement. Besides residual dipolar couplings (RDCs), the two most used paramagnetic restraints are called paramagnetic relaxation enhancements (PREs), derived from the enhancements of nuclear relaxation rates, and pseudocontact shifts[61] (PCSs).

The collection of a PCS dataset relies on the determination of the magnetic susceptibility anisotropy tensor ($\Delta\chi$), that arise from the paramagnetic metal ion, and its fitting on a protein structural model[62]. Then, the structural model is iteratively refined against PCS restraints (together with other restraints) and its $\Delta\chi$ updated[63]. In particular, this approach has been extensively applied in the study of calcium-binding proteins by replacing Ca^{2+} ions with lanthanide ions or by attaching lanthanide binding tags to proteins[64,65] (Figure 2). Each lanthanide ion has a different number of unpaired electrons and therefore can induce a paramagnetic effect in a shell located at variable distance from the metal ion. This lanthanides feature is typically exploited to collect multiple PCS datasets that potentially bring structural information on metalloproteins in a distance range from the metal site up to ca. 40 Å[66].

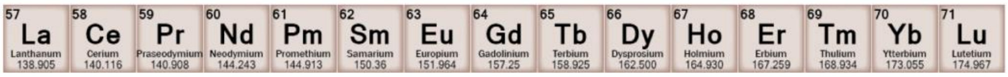

57 La Lanthanum 138.905
58 Ce Cerium 140.116
59 Pr Praseodymium 140.908
60 Nd Neodymium 144.243
61 Pm Promethium 144.913
62 Sm Samarium 150.36
63 Eu Europium 151.964
64 Gd Gadolinium 157.25
65 Tb Terbium 158.925
66 Dy Dysprosium 162.500
67 Ho Holmium 164.930
68 Er Erbium 167.259
69 Tm Thulium 168.934
70 Yb Ytterbium 173.055
71 Lu Lutetium 174.967

Figure 2. Lanthanide series in the periodic table.

2.1.2 Refinement of Proteins Structure with restrained MD

The calculation of protein structures based on NMR data usually is executed with specific computer packages as Cyana or Xplor-NIH[67,68]. These programs can manage a number of NMR data that can be used as restraints in simulated annealing calculations, at the end of which a 3D model of the molecule of interest is generated. Usually, NOE-derived distance restraints are given as pseudopotential in the calculations to the determination of the molecular structures (Figure 3). Moreover, NMR data can provide further relevant structural information, such as dihedral angles (derived from chemical shifts) or other restraints derived from paramagnetic data (RDCs, PCSs and others). Once that a data-driven simulated annealing calculation is completed, the generated NMR structure needs a final energetic refinement to improve its structural quality, in particular for stereochemical parameters[69]. In this regard, other possible approaches are the so-called structure rebuilding or statistical potential[70,71].

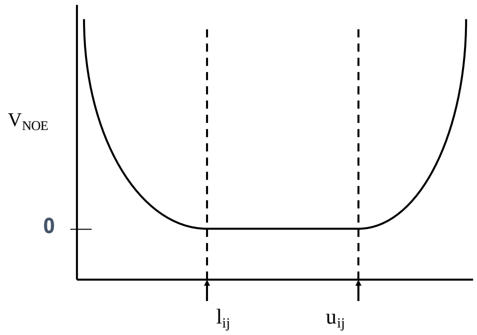


Figure 3. Parabolic flat-bottom potential often used for NOE-driven structural determination. Upper (u_{ij}) and lower (l_{ij}) limit for the NOE distance r_{ij}.

Energetic refinement relies on the force fields and water treatment (explicit or implicit), implemented in NMR-based structure calculation programs, that often have lower quality than those implemented in classical MD packages as AMBER or GROMACS[72,73]. Thus, restrained molecular dynamics (rMD) can improve significantly the accuracy of the generated structure both in term of agreement with experimental data and stereochemical quality. The restraints given in input to MD packages need to be converted to the supported format in order to retain the structure consistent with the experimental data. This step can be tedious due to conflicting rules and other format differences. To facilitate the format conversion and speed up the calculation, a web interface exploiting computational grid infrastructure called AMPS-NMR (AMBER-based Portal Server for NMR structures) is available free of charge for academics at http://py-enmr.cerm.unifi.it[74]. Furthermore, the portal implements predefined protocols greatly simplifying rMD set up for who is not familiar with MD packages.

2.1.3 Solid-State NMR in the Investigation of Protein Assemblies

Protein assemblies are defined as a combination of different proteins or multiple copies of the same monomeric unit. In the first case are called hetero-complexes and in the second case are called homo-complexes. The structural characterization of protein complexes is extremely important since many proteins can perform their biological function only as biological assembly. The correct complex formation depends on specific protein-protein interactions (PPIs) that play a crucial role in the partner recognition. The diffusion of Cryo-EM methods has exponentially increased the investigation of big complexes and molecular machines. However, other techniques based on X-ray crystallography and NMR spectroscopy can provide information on structure and dynamics by giving a contribution to the characterization of individual subunits and by pinpointing PPIs.

Among the experimental methods, solid-state NMR (SSNMR) has an increasing potential in the study of complexes because is not limited by protein size, solubility, crystallization difficulties, presence of inorganic/organic matrices or lack of long-range order that often make the application of other structural biology methods extremely difficult or unsuitable. In the NMR context, SSNMR can yield good quality spectra by the identification and assignment of through-space nucleus-nucleus interactions also in experiments on high molecular weights, such as protein assemblies. For this purpose, DARR is a diffused pulse sequence based on ^{13}C-^{13}C magnetization transfer through proton-driven spin diffusion[75]. By tuning experimental DARR parameters, users can select the range of distances at which inter-nuclear interactions are sampled. Although solid-state resonance lines of protein complexes are narrow, spectral congestion coming from resonance overlap makes the assignment of DARR peaks a challenging task (Figure 4). As a result, SSNMR DARR experiments collect a number of ambiguous contacts in which the tertiary and quaternary contacts cannot be distinguished, thus precluding the correct identification of protein-protein contacts and the determination of the complex structure. In hetero-complexes this problem can be mitigated by using different schemes for the enrichment of stable NMR-active isotopes (^{13}C, ^{15}N) in different complex units. Unfortunately, this procedure is extremely more tedious for homo-complexes, in which the identification of inter-monomeric contacts remains largely a manual task that requires a lot of time and user efforts.

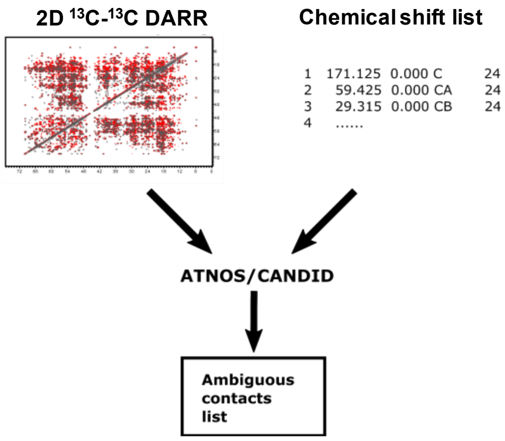


Figure 4. Example of a scheme for the generation of ambiguous contacts from 2D ^{13}C-^{13}C DARR experiment.

2.1.4 Coevolution in the Investigation of Protein Assemblies

PPIs of protein assemblies can be also predicted by investigating their evolutionary conservation across many different proteins of the same family. The so-called coevolution analysis assumes that the conservation of fundamental residue interactions is favored during protein evolution[76]. This implies that residues in contacts in the tertiary structure or among the complex subunits show correlations in their aminoacidic occurrences. The information derived can be exploited by structural biologists, alone or in combination with experimental data, to predict tertiary or quaternary structures or even putative multiple protein conformational states. In the case of homo-complexes, the identification of evolutionary couplings (ECs) consistent with PPIs is based on the construction of a single large multiple-sequence alignment (MSA) in which are present the homologue protein family sequences. On the contrary, in hetero-complexes the ECs prediction requires a *joint* MSA of two protein families, in which each line corresponds to an interacting protein pair. This crucial step is very often much difficult to implement due to the presence of paralogs that must be identified and removed prior to the covariance analysis. Although the construction of the MSA is simpler for homo-complexes, the identification of ECs belonging to inter-molecular contacts is much more complicated because such information is hidden among thousands of ECs that can belong to either tertiary or quaternary contacts[77]. The removal of ECs corresponding to tertiary contacts requires the knowledge of the tridimensional structure of the monomeric subunit of the complex. Nevertheless, at the time of writing (2019) there is a relevant number (about 2000) of protein families annotated as forming homo-oligomeric assemblies *in vivo* with a deposited monomeric structure in the Protein Data Bank (PDB). These families potentially constitute an interesting target for homo-oligomeric structural predictions.

2.2 Research Projects

As described above, PCS are information-rich paramagnetic restraints that can be measured when the protein binds a paramagnetic metal ion. For lanthanide-substituted proteins it is possible to measure multiple sets of PCS data. In principle, these data sets can be used simultaneously and in combination with diamagnetic restraints to generate structural models, thanks to the incorporation of specific routines in different structure generation programs. However, there were no available protocols for rMD using multiple PCS datasets as part of the restraints. Thus, the PSEUDOAMBER routine was extended to allow the energetic refinement of structural models in the presence of all available PCS restraints. This new extension implements a refinement protocol that is a finely tuned version of the AMPS-NMR standard protocol. To test the new implementation, NOEs and multiple independent PCS data sets were used to refine bovine calbindin D_{9k}. In each PCS dataset, a different member of the entire series of lanthanide trivalent cations (excluding the radioactive Pm^{3+} and the isotropic Gd^{3+}) was used. How much the protocol affects the structural quality was evaluated performing rMD on the NMR bundle models generated from the beginning (i.e. in CYANA) only with traditional restraints or with traditional restraints in combination with multiple PCS datasets. With respect to improving the local energetics and geometric features of the protein structure, the protocol was equally effective in the refinement of structures initially generated with or without inclusion of the PCS data. The content of secondary structure elements was not significantly affected by the refinement, yet there was a small but consistent improvement in the distribution of residues in the most favoured regions of the Ramachandran plot. The Procheck G-factor all (Z-score) featured a much greater improvement than its counterpart focusing only on the backbone conformation. This significant difference was due to the impact of the refinement on the rotameric states of all side chains. Another improvement was the extensive removal of bumps between atoms, as indicated by the large improvement of the MolProbity clashscore. Finally, correlation between experimental and back-calculated PCS data before and after rMD refinement indicated a relevant improvement of the agreement in the refined models. Overall, this new implementation delivered the expected improvement of protein geometry, resulting in final structures that were of suitable quality for deposition in PDB.

In the CASP-NMR project, CASP13 predictors were provided with "ambiguous contact lists" derived from the analysis of simulated or real NOESY peak lists using the program ASDP[78]. In the case of generating simulated NOESY data, the procedure was started by using structure coordinates as input to the SHIFTX2 program to generate chemical shift values for ^{1}H, ^{13}C, and ^{15}N atoms (Figure 5). The simulated NOESY peak lists and the final table of the simulated chemical shifts were used to generate a list of ambiguous distance restraints using the EC0-NMR method from the ASDP program[79]. In total, data for 13 different proteins or protein domains were provided to competitors. In so doing, for each dataset a number of resonance assignments and NOESY peaks were deleted to simulate line broadening due to conformational dynamics and incomplete assignments. In fact, protein dynamics and signal overlap are the main causes of errors in the conversion of NMR observables into structural restraints. The first case can lead to missing observable peaks involving a given nucleus. The second case can lead to less reliable assignments of resonance frequencies e.g. due to accidental degeneracies. In the latter case, peaks are observable but cannot be converted to the appropriate structural restraints. To mimic realistic data, both types of problems were introduced in the simulations by randomly selecting loop residues. For each extracted residue, it was randomly decided whether to remove the assignments of all its nuclei and all corresponding peaks, or to discharge only the frequency assignments while retaining the peaks, hence leading to the inclusion of ambiguous restraints that are impossible to satisfy. In addition, random noise peaks were added by randomly combining shifts of the direct and indirect dimensions. In general, considering both the removal of assignments from the chemical shift list and the addition of random noise to the spectra, in each simulated dataset the fraction of ambiguous restraints that were impossible to satisfy was kept around 10% of the whole list. Unfortunately, the program Talos+ used to predict the chemical shifts not always was consistent with the X-ray structure. Thus, backbone dihedral angles of all the structured residues present in the final table of chemical assignments were randomized in a range +/- 30 of the value observed in the X-ray structure.

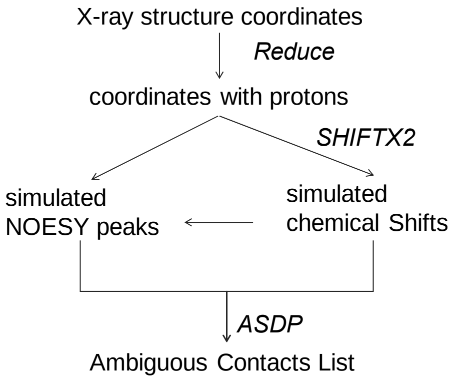


Figure 5. Scheme adopted for the simulation of ambiguous contacts in CASP13.

As already anticipated, many proteins carry out their functional role acting as part of protein assemblies. The assembly of the correct biological complex strongly depends upon specific protein-protein interactions (PPIs) that often are evolutionary conserved. If this is the case, coevolution analysis can detect quaternary contacts besides tertiary contacts. However, in homo-oligomers the separation of evolutionary couplings (ECs) belonging to protein-protein interface from the others is not straightforward. Similarly, solid-state NMR (SSNMR) is often used in the investigation of protein assemblies, as the quality of SSNMR spectra does not decrease with increasing molecular weight, but spectral congestion makes the investigation of homo-complexes largely a manual task.

In this project, a protocol was developed to automatically predict the structure of homo-complexes from SSNMR-derived ambiguous contact lists using coevolution analysis. The same protocol can be also applied using only solution-state NMR data. Our protocol calculates a list of putative interface residues to be used as input for docking calculations. It needs four inputs: one or more files with the list of ECs, the structure of the monomer, the experimental NMR-derived list of ambiguous contacts and the Naccess file (rsa format) with the per-residue relative solvent accessible area. The protocol was validated on a tetrameric and a dimeric protein. The ECs for both the proteins were collected using 3 servers available online: Gremlin, RaptorX and ResTriplet. The monomer-monomer docking calculations were carried out with HADDOCK. The structural prediction of the tetrameric E. coli L-asparaginase II (PDB: 6EOK) is not trivial since this protein features a D2 symmetry, hence two distinct dimeric conformations must be recognized to reconstruct the functional complex. Nevertheless, the calculated ECs were matched with a solid state 2D^{13}C-^{13}C

DARR dataset yielding a relevant number of predicted interface residues for both the protein interfaces. In fact, the docking calculation mainly sampled a structural configuration with an RMSD of about 1 Å from the crystallographic dimer with the largest interface in the tetrameric protein. After the removal of the restraints already satisfied by the most favored configuration, the second docking run resulted in a second dimer that, despite being less favored, had an RMSD of 1.3 Å from the crystallographic dimer with the smaller interface. The final tetrameric structure can be easily reconstructed by symmetry. In addition, the agreement between side chains orientation and docking accuracy was assessed building template-based monomeric conformations with side chains randomly oriented. Finally, the robustness of the protocol in the identification of residues belonging to small interface regions was tested in the prediction of dimeric human apo Sod1 (PDB: 3ECU). The match between the calculated ECs and solution-state 3D ^{1}H-^{15}N NOESY-HSQC dataset resulted in residues that, given to docking calculation, yielded clusters with comparable HADDOCK score values. Nevertheless, the distribution of the desolvation energies discriminated the best cluster with high accuracy, showing this energetic contribution as an excellent scoring function.

2.3 Results

2.3.1 A protocol for the refinement of NMR structures using simultaneously pseudocontact shift restraints from multiple lanthanide ions

*Davide Sala[1], Andrea Giachetti[1], Claudio Luchinat[1, 2, *]and Antonio Rosato[1, 2, *]*

[1]Magnetic Resonance Center (CERM) – University of Florence, Via L. Sacconi 6, 50019 Sesto Fiorentino, Italy

[2]Department of Chemistry – University of Florence, Via della Lastruccia 3, 50019 Sesto Fiorentino, Italy

J Biomol NMR. 2016 Nov;66(3):175-185. Epub 2016 Oct 22.

J Biomol NMR (2016) 66:175–185
DOI 10.1007/s10858-016-0065-6

ARTICLE

A protocol for the refinement of NMR structures using simultaneously pseudocontact shift restraints from multiple lanthanide ions

Davide Sala[1] · Andrea Giachetti[1] · Claudio Luchinat[1,2] · Antonio Rosato[1,2]

Received: 5 July 2016 / Accepted: 29 September 2016 / Published online: 22 October 2016

Abstract The binding of paramagnetic metal ions to proteins produces a number of different effects on the NMR spectra of the system. In particular, when the magnetic susceptibility of the metal ion is anisotropic, pseudocontact shifts (PCSs) arise and can be easily measured. They constitute very useful restraints for the solution structure determination of metal-binding proteins. In this context, there has been great interest in the use of lanthanide(III) ions to induce PCSs in diamagnetic proteins, e.g. through the replacement native calcium(II) ions. By preparing multiple samples in each of which a different ion of the lanthanide series is introduced, it is possible to obtain multiple independent PCS datasets that can be used synergistically to generate protein structure ensembles (typically called bundles). For typical NMR-based determination of protein structure, it is necessary to perform an energetic refinement of such initial bundles to obtain final structures whose geometric quality is suitable for deposition in the PDB. This can be conveniently done by using restrained molecular dynamics simulations (rMD) in explicit solvent. However, there are no available protocols for rMD using multiple PCS datasets as part of the restraints. In this work, we extended the PCS module of the AMBER MD package to handle multiple datasets and tuned a previously developed protocol for NMR structure refinement to achieve consistent convergence with PCS restraints. Test calculations with real experimental data show that this new implementation delivers the expected improvement of protein geometry, resulting in final structures that are of suitable quality for deposition. Furthermore, we observe that also initial structures generated only with traditional restraints can be successfully refined using traditional and PCS restraints simultaneously.

Keywords Pseudocontact shift · Lanthanide ion · Metalloprotein · Protein structure · Structure refinement

Introduction

Since the 1980s NMR spectroscopy has been used as a technique complementary to X-ray to determine the solution structure of proteins (Wüthrich 1986; Cavanagh et al. 2007). The usefulness of addressing protein structure determination using both NMR and X-ray approaches has been extensively documented, especially within Structural Genomics initiatives (Synder et al. 2005; Yee et al. 2005; Serrano et al. 2016). The standard protocols for NMR-based protein structures determination involve three main steps: (i) determining the chemical shift assignments of the target protein; (ii) measuring as many as possible NOEs (nuclear Overhauser enhancements) between pairs of ^{1}H nuclei to compute internuclear distance restraints (upper distance limits, upls); and (iii) using the NOE-derived upls to generate the 3D structure of the protein. Additional restraints that are normally used include dihedral angle restraints [which often are derived from J-couplings (Karplus 1959) or from

Electronic supplementary material The online version of this article (doi:10.1007/s10858-016-0065-6) contains supplementary material, which is available to authorized users.

✉ Claudio Luchinat
luchinat@cerm.unifi.it

✉ Antonio Rosato
rosato@cerm.unifi.it

[1] Magnetic Resonance Center (CERM), University of Florence, Via Luigi Sacconi 6, 50019 Sesto Fiorentino, Italy

[2] Department of Chemistry, University of Florence, Via della Lastruccia 3, 50019 Sesto Fiorentino, Italy

chemical shifts (Shen et al. 2009)] and residual dipolar couplings (Tjandra and Bax 1997). Alternative protocols based on the use of chemical shift data or chemical shift and residual dipolar couplings without any NOE information have been described (Cavalli et al. 2007; Shen et al. 2008; Vernon et al. 2013; Cavalli and Vendruscolo 2015; van der Schot and Bonvin 2015). As a further alternative, limited NOE information (sparse NOEs) can be used together with various other types of data, including evolutionary-based restraints, to define the protein fold with reasonable accuracy (Lange et al. 2012; Tang et al. 2015).

Within standard protocols, the energetic refinement of the initial 3D structures has a deep impact on the quality of the final NMR structure that will be eventually deposited in the PDB. Aside from the accuracy and completeness of experimental data, the quality of NMR structures thus depends on the programs utilized in the generation of initial structures with simplified force fields and in their subsequent energetic refinement (Linge and Nilges 1999; Linge et al. 2003; Chen et al. 2004). In particular, as demonstrated by many studies, the quality of NMR structures can be improved by refinement either based on molecular dynamics simulations with state-of-the-art force field and explicit or, less often, implicit solvent (Xia et al. 2002; Linge et al. 2003; Chen et al. 2004; Feig et al. 2004; Jao et al. 2008). Other approaches based on structure rebuilding (Mao et al. 2014) or on the use of statistical potentials (Ryu et al. 2016) have been proposed. Such refinement approaches can improve significantly the geometric parameters that are often used as indicators of structural quality and can also have a positive impact on the accuracy of the structure, as measured from the agreement with the NMR data (Nabuurs et al. 2004; Nederveen et al. 2005). On the other hand, these procedures may sometimes mask errors in the NMR structures, at least as far as the normality of geometric parameters is concerned (Nabuurs et al. 2006; Saccenti and Rosato 2008). Therefore, NMR structure validation approaches should be based on a combination of quantitative parameters describing geometric quality and agreement with NMR data (Huang et al. 2012; Rosato et al. 2013; Ragan et al. 2015; Vuister et al. 2014; Montelione et al. 2013; Doreleijers et al. 2012).

Pseudocontact shifts (PCSs) arise in paramagnetic molecules when the magnetic susceptibility tensor of the paramagnetic center is anisotropic, as it is often the case when the paramagnetic center is a metal ion. Because of the anisotropic magnetic susceptibility, macromolecular systems featuring measurable PCSs spontaneously orient in solution in the presence of a magnetic field, thereby allowing residual dipolar couplings (RDCs) to be measured as well (Banci et al. 1998b). PCSs contain structural information that proved very helpful for solving protein structures both in solution (Gochin and Roder 1995; Banci et al. 1996, 1998a; Allegrozzi et al. 2000; Bertini et al. 2001b; Gaponenko et al. 2004; Banci et al. 2004; Schmitz et al. 2012; Yagi et al. 2013; Camilloni and Vendruscolo 2015), in the solid state (Balayssac et al. 2008; Bertini et al. 2010; Li et al. 2013; Jaroniec 2015) and in living cells (Pan et al. 2016; Muntener et al. 2016). PCSs can also be used for structural refinement in combination with X-ray data (Rinaldelli et al. 2014; Carlon et al. 2016). In this context, the use of lanthanide ions is attractive due to their large magnetic anisotropy that induces large PCSs and RDCs (Bertini et al. 2008). Lanthanide ions can be introduced in proteins in different ways. Among these, there is extensive literature describing metal-substitution in calcium-binding proteins (Allegrozzi et al. 2000; Bertini et al. 2003) and the use of so-called lanthanide binding tags that are chemically attached to proteins (Barthelmes et al. 2011; Hass and Ubbink 2014; Rodriguez-Castañeda et al. 2006; Su and Otting 2010). Owing to the different number of unpaired electrons in the lanthanide series, respectively, the induced PCSs as well as paramagnetic relaxation enhancements, which lead to signal broadening even beyond detection, are significantly different for different ions. Thus, the combination of PCS data from different paramagnetic lanthanides provides structural information on different shells at variable distances from the binding site of the metal ion (Allegrozzi et al. 2000). This has been exploited not only for the structural determination of individual globular proteins (Schmitz et al. 2012; Yagi et al. 2013) but also for two-domain proteins and for protein–protein docking (Hass and Ubbink 2014; Hulsker et al. 2008; Brewer et al. 2015; Chen et al. 2014). The software tools for structure calculation and for structure refinement can handle RDCs induced by self-orientation and the more popular RDCs induced by the presence of orienting media in solution in exactly the same way. Instead, it is necessary to implement specific routines for the use of PCSs as structural restraints. It is currently possible to generate structural models of proteins by combining traditional NMR restraints with PCS restraints from multiple lanthanides, for example using torsion angle dynamics (Banci et al. 1998a). However, is no protocol for the energetic refinement of such models. This makes it difficult to obtain structures based on multiple PCS datasets that are of sufficient quality for deposition in the PDB. Here, we report on an extension of PSEUDOAMBER (Banci et al. 1997) that addresses this limitation, by providing a protocol that closely resembles the protocols routinely used for structure determination without PCS restraints and thus can be readily adopted by NMR structural biologists. We validated the above protocol by using the PCS data measured for bovine calbindin D9k where the full series of the lanthanide ions (except the radioactive promethium and the isotropic gadolinium) was bound into its C-terminal calcium-binding site (Bertini et al. 2001a).

Materials and methods

Implementation of the restraints

We extended our previous implementation of PCS in the *sander* module of the AMBER suite (Banci et al. 1997), in order to permit the use of multiple experimental datasets, each from a different lanthanide ion, all referring to the position of a single metal ion. For the present implementation, we used version 12 of the suite. For each lanthanide ion, the following holds (McConnell and Robertson 1958)

$$\delta^{pcs} = \frac{1}{12\pi r^3}\left[\Delta\chi_{ax}\left(3\cos^2\vartheta - 1\right) + \frac{3}{2}\Delta\chi_{rh}\sin^2\vartheta\cos 2\phi\right] \tag{1}$$

where $\Delta\chi_{ax}$ and $\Delta\chi_{rh}$ are the axial and the rhombic anisotropies of the magnetic susceptibility tensor, and r, θ, and φ are the polar coordinates of the nucleus under observation with respect to the orthogonal reference system formed by the principal axes of the magnetic susceptibility tensor. To use the PCS induced by a single paramagnetic center in a protein as restraints in molecular dynamics calculations, in previous work we defined a pseudo-potential term based on the deviation between experimental (*exp*) and back-calculated (*calc*) PCS values as

$$U^{pcs} = k^{pcs}\sum_{i=1}^{N} w_i\Delta_i^2$$
$$\text{where } \Delta_i = \left(\left|\delta_i^{exp} - \delta_i^{calc}\right| - err_i\right)\text{if}\,\Delta_i > 0, 0 \text{ otherwise} \tag{2}$$

where k^{PCS} is the weight for PCS restraints with respect to the other terms of the complete potential, w_i is the relative weight of the each of the N PCS restraints and Δ_i is the deviation between the experimental and back-calculated values of the i-th restraint, beyond the experimental uncertainty (err_i). In practice, the summation includes only violations that are greater than the uncertainty on each PCS measurement. In our experience, it is not necessary to define w_i weights different from 1.0, except when using null weights for PCS data the assignment of which one wants to confirm based on the calculations results, e.g. when the assignment of an amide proton in the paramagnetic system is ambiguous.

To extend the above approach to the case of L different lanthanide ions, we modified Eq. 2 as follows

$$U^{pcs} = \sum_{l=1}^{L} w_l U_l^{pcs} \tag{3}$$

where U_l^{PCS} is the contribution of the data from the l-th lanthanide ion, as defined in Eq. 2. The experimental PCS restraints associated to the l-th lanthanide ion are back-calculated using Eq. (1) with the specific $\Delta\chi$ values of that ion. Such values, as well as the orientation of the magnetic susceptibility axes with respect to the laboratory frame, are kept constant during the simulation. Therefore, each of the L datasets contributes independently to the overall pseudo-potential, and its contribution can be globally scaled with respect to the other datasets by adjusting the w_l weight. Note that setting a w_l value different from 1 in Eq. (3) is equivalent to multiplying all the w_i weights of the individual PCS restraints of the l-th dataset by the same scaling factor. We assume that the origin of all L tensors coincides. This is the relevant scenario when multiple lanthanides are substituted in the same metal-binding site.

Generation of initial structures

We used the program CYANA (Herrmann et al. 2002) to generate the initial structural models (without or with PCS restraints) to be used for the demonstration of our refinement protocol, using the methods and datasets described in (Bertini et al. 2001a, b). Briefly, we used 1539 NOE-based upper distance limits, 6 distance limits derived from paramagnetic relaxation data and dihedral angles for all residues in secondary structure elements as the ensemble of "traditional restraints". 1097 PCS restraints were derived from measurements on eleven different calbindin D_{9k} (Ca_2Cb) samples, where one of the two calcium ions was selectively replaced by a lanthanide ion (CaLnCb). The PCS values were determined by subtracting the shifts of the CaLaCb or CaLuCb sample from the particular lanthanide derivative (Bertini et al. 2001a). Structure bundles containing 100 conformers each were generated using the standard annealing protocol of CYANA in 10,000 steps. All restraints were used from the beginning of annealing procedure. We generated two different bundles: one using only traditional restraints, which we dubbed CYnoPCS, and one with traditional restraints and PCS restraints, which we dubbed CYwithPCS.

Refinement protocol

The refinement procedure is applied to each conformer of the input bundle independently by computing a single rMD trajectory. The latter consists of six stages: two minimizations (of water only and then of the entire system), three simulated annealing stages (heating, constant-T rMD and cooling), and a final energy minimization (Fig. 1 and Supplementary material). All the minimizations are carried out with a combination of steepest descent followed by conjugate gradient minimization. Before starting the refinement, every conformer of the input bundle of structures (e.g. generated with CYANA) is embedded in a rectangular or octahedral box of TIP3P water molecules with a user-selected distance (10 Å in the present test case)

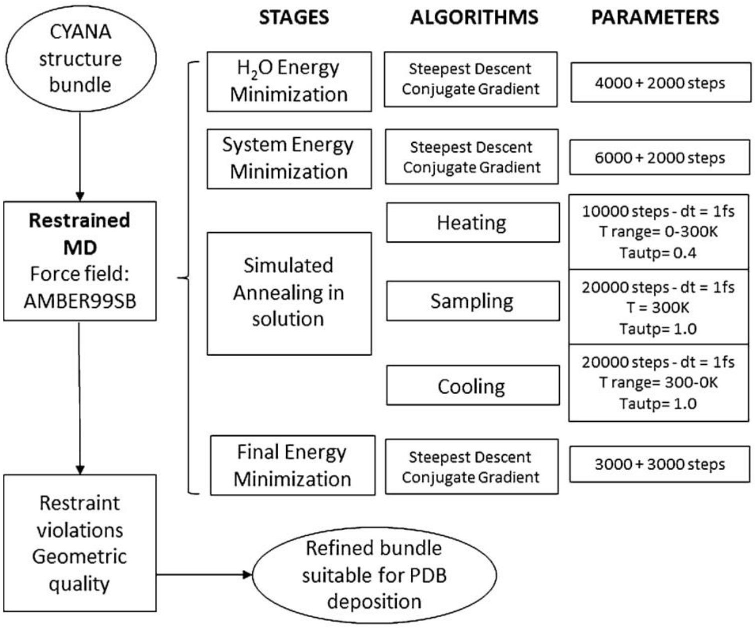

CYANA structure bundle
STAGES
ALGORITHMS
PARAMETERS
H$_2$O Energy Minimization
Steepest Descent Conjugate Gradient
4000 + 2000 steps
System Energy Minimization
Steepest Descent Conjugate Gradient
6000 + 2000 steps
Restrained MD Force field: AMBER99SB
Simulated Annealing in solution
Heating
10000 steps - dt = 1fs T range= 0-300K Tautp= 0.4
Sampling
20000 steps - dt = 1fs T = 300K Tautp= 1.0
Cooling
20000 steps - dt = 1fs T range= 300-0K Tautp= 1.0
Final Energy Minimization
Steepest Descent Conjugate Gradient
3000 + 3000 steps
Restraint violations Geometric quality
Refined bundle suitable for PDB deposition

Fig. 1 Flowchart summarizing the rMD refinement protocol presented in this article

between the protein surface and the box walls. In the present work, the 20 conformers with the best agreement with all experimental data after rMD were retained as the refined NMR structure.

In the first stage, the protein is restrained with a harmonic potential, so that the minimization mostly affects the water molecules in the hydration shell. In the subsequent stages of the protocol, the protein moves freely in presence of the active pseudopotential of all the NMR restraints. For the rMD stages, the integration step can be set to 1–2 fs (we used 1 fs for the present text case). The rotational motion of the protein is turned off, as the reference frames of the magnetic susceptibility tensors are integral with the laboratory frame, and PBC are not used. The length of all bonds involving hydrogen atoms is constrained with the SHAKE and SETTLE (for the water molecules) algorithms. The weak-coupling algorithm is applied to control the system temperature. A tight temperature regulation is maintained during the heating phase by setting the time constant for heat bath coupling to 0.4 ps until the temperature reached its target of 300 K. The temperature coupling is then relaxed to 1 ps during the constant-T rMD trajectory. In the subsequent cooling stage, the temperature of the system returns to 0 K to complete the simulated annealing. During the warming and cooling phases, the temperature is scaled linearly with time, as commonly performed in NMR-based structure determination algorithms (Fossi et al. 2005; Mareuil et al. 2015). Finally, all the conformers are minimized to reach the closest local minimum of energy.

The user can modify most of the parameters in the protocol, such as the temperature for the rMD stages or the number of steps (i.e. simulation time in ps) of rMD. Traditional restraints (NOE-derived upper distance limits, upls, and dihedral angle restraints) are applied using a flat-bottom parabolic potential. A crucial parameter is the weight of the PCS restraints (Eq. 2) versus the weight of all other NMR restraints. For the present text, we chose weights of 32 kcal mol^{-1} Å^{-2} and 30 kcal mol^{-1} ppm^{-2} for traditional and PCS restraints, respectively. The user can adjust one or both of these parameters to scale the relative contribution of the different restraints, as well as the impact of the AMBER force field on the final structures. The used parameter files for AMBER, the input CYwithPCS structure and the corresponding output structure are provided as Supplementary material.

The *sviol* script of the AMBER package provides users with a report on the violations of traditional restraints. The default cutoffs are >0.1 Å for upls and 10° for dihedral angles. We used the FANTEN software (Rinaldelli et al. 2015), which is available on the WeNMR portal (Wassenaar et al. 2012), to fit the PCS data before and after the rMD, in order to assess the stability of the $\Delta\chi$ parameters (Banci et al. 2004). FANTEN provides a very convenient graphical interface to visualize immediately the agreement between back-calculated and experimental PCS data. This is done for each PCS dataset against the structure bundles before and after refinement. As the rMD protocol starts from conformations that are already properly folded and not random coils (Fig. 1) significant changes of the $\Delta\chi$ parameters are not observed. Finally, the overall geometric quality of the bundles was assessed with the Protein Structure Validation Software suite (Version 1.5) (Bhattacharya et al. 2007). The secondary structure content was computed with the DSSP program (Kabsch and Sander 1983).

Results and discussion

In the context of protein structure determination based on NMR data, molecular dynamics (MD) methods are used both to generate structural models in agreement with the data and, at a later stage, to optimize the energetics and several geometric aspects of the final structures, prior to deposition in the PDB. In this work, we wanted to enable the use of MD for structure refinement also when using multiple independent PCS datasets, available thanks to the substitution of different lanthanide ions in the same metal-binding site. Our standardized refinement approach exploits short restrained MD (rMD) trajectories. rMD provides some kinetic energy to each conformer of the NMR bundle input to the refinement procedure, permitting the sampling of the conformational space around the initial conformation. In rMD the pseudo-potential of the NMR-based restraints remains always active in order to prevent the structure from sampling regions of conformational space incompatible with the data. After the rMD run at room temperature, the input conformers are cooled down to 0 K, resulting in the final bundle. This procedure is based on the concept of the RECOORD protocol (Nederveen et al. 2005), and has been extensively used by several research teams via the AMPS-NMR web portal for NMR structure refinement (Bertini et al. 2011). In a nutshell, this refinement scheme allows the protein structure to settle in local minima that are energetically sound, thanks to the use of the AMBER force field, while retaining full agreement with the experimental data. This is particularly effective for parts of the protein structure that are not well defined by the data themselves, such as the backbone conformation of long loops or the rotameric states of side chains not in the core of the structure (Bertini et al. 2011; Nederveen et al. 2005; Rosato et al. 2012).

Our new implementation of PCS restraints in the AMBER package permits the refinement of structures generated without or with the inclusion of PCS data from the very beginning of the calculation procedure (i.e. in CYANA). Hereafter, we will refer to these two initial bundles as CYnoPCS and CYwithPCS, respectively. Both scenarios converge successfully, and produce similar effects at the level of the overall protein fold (Fig. 2). The fold is actually defined mostly by the NOE and dihedral restraints data, so the structural impact of the addition of PCS restraints for the present system is relatively modest. This aspect of the use of PCS restraints has been addressed in detail in previous work (Allegrozzi et al. 2000). The structural refinement in explicit solvent does not, and it is not expected to, change significantly the overall protein fold (Fig. 2). Nevertheless, it appears that the definition of some elements of secondary structure improves after AMBER refinement.

The main motivation to perform the energetic refinement of macromolecular structures is to remediate deviations from optimal local configuration, as gauged by structure validation programs, while maintaining their agreement with the experimental data. The currently available software tools are lacking a protocol to perform such refinements using also multiple PCS data sets as part of the input data sets. In this work, we implemented such a protocol in the AMBER MD package. With respect to improving the local energetics and geometric features of the protein structure, the protocol is equally effective in the refinement of structures initially generated with or without inclusion of the PCS data (Table 1). The content of secondary structure elements is not significantly affected by the refinement, yet there is a small but consistent improvement in the distribution of residues in the most favored regions of the Ramachandran plot. This is apparent also from the values of the Procheck G-factor phi/psi (Z-score), especially in the case of the refinement of the CYwithPCS structure, for which this parameter increased by 0.63 (from −1.42 to −0.79). The distribution of residues in the Ramachandran plot is a sensitive indicator of the presence of errors and of local structure quality (Saccenti and Rosato 2008). The Procheck G-factor all (Z-score) features a much greater improvement than its counterpart focusing only on the backbone conformation, mentioned above. The variation for the CY with PCS structure is as large as 2.78. This significant difference is due to the impact of the refinement on the rotameric states of all side chains. Such an impact, which the Ramachandran plot analysis does not address, is actually much more significant than for the backbone, as measured by the larger increase observed for the G-factor computed on all angles with respect to the G-factor computed on the backbone dihedrals only.

The second prominent improvement afforded by the refinement procedure is the extensive removal of too-close contacts between atoms (bumps), as indicated by the large improvement of the MolProbity clashscore, which is also accompanied by a dramatic reduction in the spread of this parameter over the structure bundles (Table 1). The Verify3D parameter, which describes the likelihood of the overall fold, is less sensitive to the refinement, because the

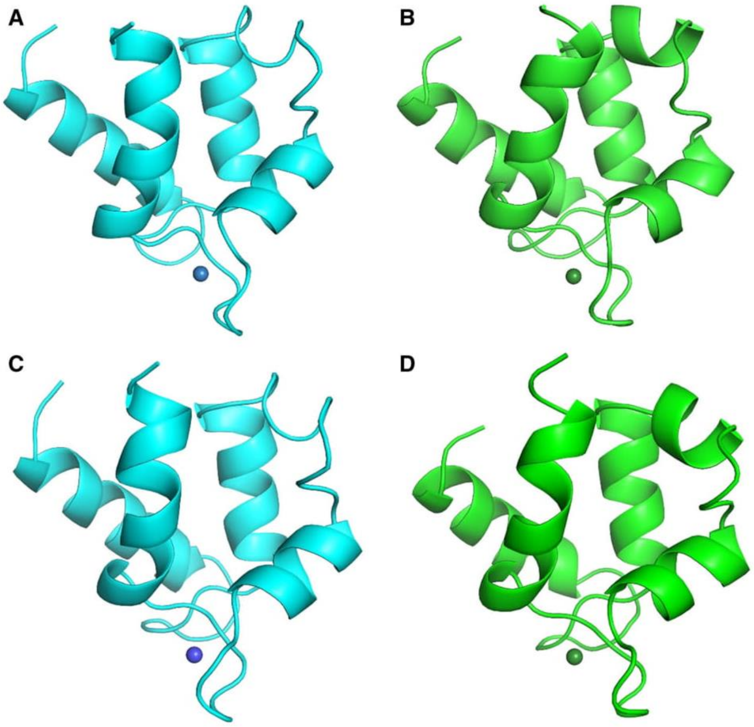


Fig. 2 Refinement of calbindin structures generated with CYANA (**a**, **c**) using AMBER with the standard protocol of the AMPS-NMR portal, incorporating all available PCS restraints (**b**, **d**). **a**, **b** Refinement of the CYANA structure generated without PCS restraints (CYnoPCS). Panels **c**, **d** Refinement of the CYANA structure generated with all PCS restraints (CYwithPCS). In all panels, the structure of the bundle closest to the mean structure is shown. The lanthanide ion is shown as a *sphere*

Table 1 Quality parameters for calbindin structures before and after rMD refinement with all available PCS restraints

	CYANA without PCS (CYnoPCS)		CYANA with PCS (CYwithPCS)	
	Before rMD	After rMD	Before rMD	After rMD
DSSP secondary structure				
Secondary structure content	0.86 ± 0.13	0.84 ± 0.24	0.83 ± 0.14	0.85 ± 0.18
Helical content	0.59 ± 0.06	0.53 ± 0.14	0.45 ± 0.06	0.54 ± 0.09
Sheet content	0.04 ± 0.01	0.03 ± 0.02	0.02 ± 0.01	0.05 ± 0.01
PROCHECK Ramachandran plot summary				
Most favored regions (%)	84.6	90.9	85.5	89.4
Allowed regions (%)	15.3	9.0	14.0	10.4
Generously allowed regions (%)	0.2	0.1	0.5	0.3
Disallowed regions (%)	0.0	0.0	0.0	0.0
Structure quality factors[a]				
Procheck G-factor φ-ψ (Z-score)	−0.75	−0.59	−1.42	−0.79
Procheck G-factor all (Z-score)	−3.78	**−1.54**	−4.55	**−1.77**
Verify3D (Z-score)	−0.96 ± 0.02	−0.80 ± 0.03	−1.28 ± 0.02	−0.96 ± 0.03
MolProbity clashscore (Z-score)	−1.01 ± 3.79	**1.25 ± 1.34**	−1.69 ± 3.37	**1.28 ± 1.05**

The input structures for the rMD procedure were CYANA structures generated with NOE-based upper distance limits and dihedral angle restraints, and without (CYnoPCS) or with (CYwithPCS) PCS restraints

[a] A positive Z-score indicates that the parameter is better than the average value observed in the PSVS reference dataset (composed by high-resolution X-ray structures). Changes larger than a Z-score unit after refinement are highlighted in bold

NOEs and dihedral angle restraints already define the topology at the level of CYANA. Clearly, even though the parameters of Table 1 separately describe specific aspects of the stereochemical and geometric quality of the structure, the refinement procedure addresses all of them simultaneously. In other words, different unsatisfactory aspects of local structure may be fixed together (Fig. 3).

Indeed, a crucial aspect to assess the quality of a structure is to quantify its agreement with experimental data (Huang et al. 2012; Rosato et al. 2013; Ragan et al. 2015; Vuister et al. 2014; Montelione et al. 2013; Doreleijers et al. 2012). CASD-NMR in particular has shown that high quality as measured by the parameters of Table 1 does not imply that a structure is correct, i.e. it may not fulfil all the experimental data (Rosato et al. 2012). In practice, these two aspects of structure quality are somewhat independent and should be simultaneously verified. The application of a structure refinement procedure thus should not affect adversely the agreement with the data, with respect to the situation of the structure before refinement. This is indeed true for the present protocol (Table 2). Regardless of the inclusion of PCS restraints in the initial CYANA structure, the refinement procedure outputs structures with a consistent level of agreement with the data. The agreement before and after refinement is essentially unchanged.

A specific aspect of this work is the use of multiple PCS data as restraints. Thus, we focus on the impact of the energetic refinement on these datasets, by measuring the agreement between the experimental data and the data back-calculated from the structure bundles via Eq. 1. The inclusion of PCS restraints in the refinement leads to a great improvement of the agreement of the CYnoPCS structure (Table 3). This is summarized by the count of the total number of PCS deviations (Δ_i of Eq. 2) larger than 0.5 ppm, which diminishes from 375 to 85. The latter value is however higher than the result obtained for the CYwithPCS structure, both before (71 deviations) and after refinement (63 deviations). The improvement from 71 to 63 deviations, albeit very small, is noteworthy because the force field of rMD prevents the protein from sampling energetically unfavorable conformations that could be previously allowed by CYANA. Nevertheless, our protocol is capable of identifying a slightly better minimum for PCS data. By inspecting the results for each lanthanide, we observe that most of the deviations arise from the Yb dataset. If this dataset is excluded the improvements described above become more significant (not shown). For a single dataset, excluding Yb, the greatest improvement is observed for Nd (Table 3; Fig. 4) for the refinement of the CYnoPCS structure. It is possible to observe that the PCSs back-calculated from the CYnoPCS bundle deviate from the diagonal in Fig. 4, whereas they move much closer to it after refinement, as indicated by the Pearson coefficient value of 0.98. Instead, changes are marginal for each individual lanthanide in the case of the CYwithPCS, as the latter was

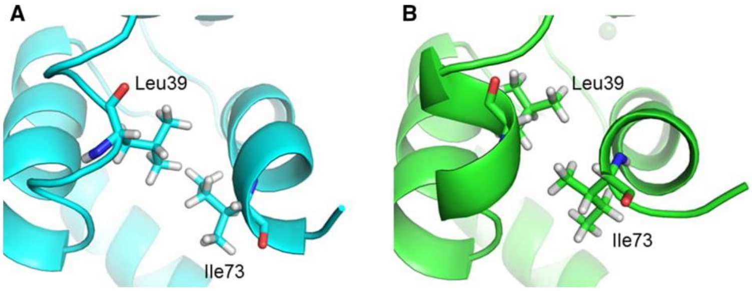


Fig. 3 Effects of refinement on local structure. A close contact between the side chains of Leu39 and Ile73 is present in the unrefined CYANA structure (**a**). Refinement with the present protocol removes the clash and optimizes the rotameric states of the involved residues (**b**)

Table 2 Violation of traditional (upper distance limits and dihedral angle) restraints for calbindin structures before and after rMD refinement with all available PCS restraints

	CYANA without PCS (CYnoPCS)		CYANA with PCS (CYwithPCS)	
	Before rMD	After rMD	Before rMD	After rMD
Violation analysis				
Distance restraint violations >0.1 Å	3 ± 0	3 ± 0	0 ± 0	0 ± 0
Mean distance violations >0.1 Å	0.13 ± 0.01	0.14 ± 0.04	0.00 ± 0.00	0.00 ± 0.00
Dihedral restraint violations >10°	0 ± 0	1 ± 0	0 ± 0	1 ± 0
Mean dihedral violations >10°	0.00 ± 0.00	11.9 ± 0.00	0.00 ± 0.00	10.1 ± 0.00

See the caption to Table 1 for details on the column headers

Table 3 Correlation between experimental and back-calculated PCS data for calbindin structures before and after rMD refinement with all available PCS restraints

	CYANA without PCS (CYnoPCS)		CYANA with PCS (CYwithPCS)	
	Before rMD	After rMD	Before rMD	After rMD
Cerium				
Correlation (%)	95.11	97.19	97.47	96.85
Mean deviation (ppm)	0.07 ± 0.10	0.06 ± 0.07	0.06 ± 0.07	0.06 ± 0.08
# Violations >0.5 ppm	28	9	3	3
Preseodymium				
Correlation (%)	98.53	99.73	99.55	99.71
Mean deviation (ppm)	0.10 ± 0.15	0.05 ± 0.06	0.07 ± 0.07	0.06 ± 0.06
# Violations >0.5 ppm	74	0	3	1
Neodymium				
Correlation (%)	94.96	98.04	97.61	98.05
Mean deviation (ppm)	0.09 ± 0.12	0.06 ± 0.07	0.07 ± 0.08	0.06 ± 0.07
# Violations >0.5 ppm	49	8	20	3
Samarium				
Correlation (%)	86.77	89.02	89.04	88.90
Mean deviation (ppm)	0.07 ± 0.08	0.06 ± 0.07	0.06 ± 0.07	0.06 ± 0.08
# Violations >0.5 ppm	3	0	0	0
Europium				
Correlation (%)	98.20	99.47	99.55	99.49
Mean deviation (ppm)	0.17 ± 0.13	0.09 ± 0.07	0.09 ± 0.06	0.09 ± 0.07
# Violations >0.5 ppm	16	0	0	0
Terbium				
Correlation (%)	98.53	99.59	99.54	99.61
Mean deviation (ppm)	0.19 ± 0.17	0.10 ± 0.09	0.11 ± 0.10	0.10 ± 0.09
# Violations >0.5 ppm	70	0	2	0
Dysprosium				
Correlation (%)	98.16	99.55	99.39	99.57
Mean deviation (ppm)	0.15 ± 0.14	0.08 ± 0.06	0.09 ± 0.08	0.07 ± 0.06
# Violations >0.5 ppm	40	0	0	0
Holmium				
Correlation (%)	98.10	98.80	98.68	98.84
Mean deviation (ppm)	0.07 ± 0.07	0.05 ± 0.05	0.06 ± 0.06	0.05 ± 0.05
# Violations >0.5 ppm	0	0	0	0
Erbium				
Correlation (%)	98.79	99.37	99.10	98.38
Mean deviation (ppm)	0.14 ± 0.11	0.10 ± 0.08	0.12 ± 0.10	0.10 ± 0.08
# Violations >0.5 ppm	12	1	1	0
Thulium				
Correlation (%)	95.77	97.19	97.62	97.45
Mean deviation (ppm)	0.13 ± 0.13	0.11 ± 0.11	0.10 ± 0.10	0.10 ± 0.10
# Violations >0.5 ppm	39	24	2	15
Ytterbium				
Correlation (%)	55.13	57.74	58.74	60.60
Mean deviation (ppm)	0.06 ± 0.10	0.06 ± 0.09	0.06 ± 0.09	0.06 ± 0.09
# Violations >0.5 ppm	44	43	40	41

Table 3 continued

	CYANA without PCS (CYnoPCS)		CYANA with PCS (CYwithPCS)	
	Before rMD	After rMD	Before rMD	After rMD
Total				
Mean deviation (ppm)	0.11 ± 0.12	0.08 ± 0.08	0.08 ± 0.08	0.07 ± 0.08
# Violations >0.5 ppm	375	85	71	63

The correlation is defined by the Pearson coefficient. The mean deviation and the number of violations >0.5 ppm are also given. See the caption to Table 1 for details on the column headers

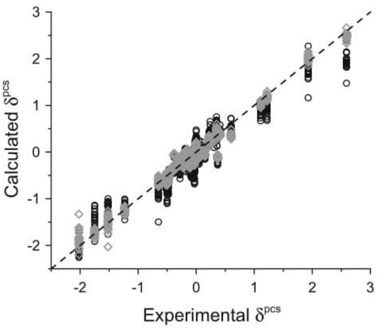


Fig. 4 Correlation between experimental (*x* axis) and back-calculated (*y* axis) PCS data of Nd^{3+}—substituted calbindin for the CYANA structures generated without PCS restraints (CYnoPCS) before (*black circles*) and after (*grey diamonds*) refinement. Each bundle was fit separately. The *dashed line* is $y = x$ and is shown only to guide the eye

generated including also PCS restraints and thus is already in very good agreement with the PCS data before rMD.

The CYnoPCS and CYwithPCS CYANA structures did not reach the same level of agreement with the PCS data after refinement (Table 3). This is presumably due to the relatively low temperatures and high weight on the "traditional" NMR restraints used in the rMD protocol. Indeed, it is the purpose of our refinement strategy not to disrupt at any stage of the calculation the fold initially output by CYANA, which is entirely driven by the experimental data, but to optimize its features that are less well defined by the data. On the other hand, the present setup results in a convergence of the refinement protocol close to 100 % (i.e. all input structures are correctly refined). This is important to allow less expert users to adopt successfully relatively complicated computational procedures.

Concluding remarks

We have implemented a rMD protocol that allows NMR structural biologists to simultaneously use PCS restraints derived from the substitution of multiple lanthanide ions in the same site in the energetic refinement of structures. This tool complements the available portfolio of software tools for the use of paramagnetic restraints in protein structure determination. The protocol exploits the AMBER package for molecular dynamics simulations, and can be incorporated in the AMPS-NMR portal (Bertini et al. 2011) for NMR structure refinement provided by the WeNMR electronic infrastructure (Wassenaar et al. 2012). The rMD refinement affords the expected improvement of initial structures in terms of their geometric quality, especially for rotamer distributions and interatomic bumps. Importantly, our implementation allows a structure initially calculated without PCS restraints to be refined using such data as additional restraints, obtaining both good agreement with all experimental data and good geometry. The refined bundles are of the typical quality for NMR structures deposited in the PDB. This is achieved thanks to an approach that seamlessly integrates into routine procedures for NMR-based protein structure determination (Bertini et al. 2011).

Acknowledgments This work has been supported by the European Commission through contracts EGI-Engage No. 654142 and WeNMR No. 261572. The FP7 WeNMR and H2020 West-Life (no. 675858) European e-Infrastructure projects are acknowledged for the use of their web portals, which make use of the EGI infrastructure and DIRAC4EGI service with the dedicated support of CESNET-MetaCloud, INFN-PADOVA, NCG-INGRID-PT, RAL-LCG2, TW-NCHC, SURFsara and NIKHEF, and the additional support of the national GRID Initiatives of Belgium, France, Italy, Germany, the Netherlands, Poland, Portugal, Spain, UK, South Africa, Malaysia, Taiwan and the US Open Science Grid.

References

Allegrozzi M, Bertini I, Janik MBL, Lee Y-M, Liu G, Luchinat C (2000) Lanthanide induced pseudocontact shifts for solution structure refinements of macromolecules in shells up to 40 A from the metal ion. J Am Chem Soc 122:4154–4161

Balayssac S, Bertini I, Bhaumik A, Lelli M, Luchinat C (2008) Paramagnetic shifts in solid-state NMR of proteins to elicit strucutral information. Proc Natl Acad Sci USA 105: 17284–17289

Banci L, Bertini I, Bren KL, Cremonini MA, Gray HB, Luchinat C, Turano P (1996) The use of pseudocontact shifts to refine solution structures of paramagnetic metalloproteins: Met80Ala cyanocytochrome *c* as an example. J Biol Inorg Chem 1:117–126

Banci L, Bertini I, Gori Savellini G, Romagnoli A, Turano P, Cremonini MA, Luchinat C, Gray HB (1997) Pseudocontact shifts as constraints for energy minimization and molecular dynamic calculations on solution structures of paramagnetic metalloproteins. Proteins Struct Funct Genet 29:68–76

Banci L, Bertini I, Cremonini MA, Gori Savellini G, Luchinat C, Wüthrich K, Güntert P (1998a) PSEUDODYANA for NMR structure calculation of paramagnetic metalloproteins using torsion angle molecular dynamics. J Biomol NMR 12:553–557

Banci L, Bertini I, Huber JG, Luchinat C, Rosato A (1998b) Partial orientation of oxidized and reduced cytochrome b_5 at high magnetic fields: magnetic susceptibility anisotropy contributions and consequences for protein solution structure determination. J Am Chem Soc 120:12903–12909

Banci L, Bertini I, Cavallaro G, Giachetti A, Luchinat C, Parigi G (2004) Paramagnetism-based restraints for Xplor-NIH. J Biomol NMR 28:249–261

Barthelmes K, Reynolds AM, Peisach E, Jonker HRA, DeNunzio NJ, Allen KN, Imperiali B, Schwalbe H (2011) Engineering encodable lanthanide-binding tags into loop regions of proteins. J Am Chem Soc 133:808–819

Bertini I, Janik MBL, Lee Y-M, Luchinat C, Rosato A (2001a) Magnetic susceptibility tensor anisotropies for a lanthanide ion series in a fixed protein matrix. J Am Chem Soc 123:4181–4188

Bertini I, Janik MBL, Liu G, Luchinat C, Rosato A (2001b) Solution structure calculations through self-orientation in a magnetic field of cerium (III) substituted calcium-binding protein. J Magn Reson 148:23–30

Bertini I, Gelis I, Katsaros N, Luchinat C, Provenzani A (2003) Tuning the affinity for lanthanides of calcium binding proteins. Biochemistry 42:8011–8021

Bertini I, Luchinat C, Parigi G, Pierattelli R (2008) Perspectives in NMR of paramagnetic proteins. Dalton Trans 2008:3782–3790

Bertini I, Bhaumik A, De Paepe G, Griffin RG, Lelli M, Lewandowski JR, Luchinat C (2010) High-resolution solid-state NMR structure of a 17.6 kDa protein. J Am Chem Soc 132:1032–1040

Bertini I, Case DA, Ferella L, Giachetti A, Rosato A (2011) A grid-enable web portal for NMR structure refinement with AMBER. Bioinformatics 27:2384–2390

Bhattacharya A, Tejero R, Montelione GT (2007) Evaluating protein structures determined by structural genomics consortia. Proteins: Struct, Funct, Bioinf 66:778–795

Brewer KD, Bacaj T, Cavalli A, Camilloni C, Swarbrick JD, Liu J, Zhou A, Zhou P, Barlow N, Xu J, Seven AB, Prinslow EA, Voleti R, Haussinger D, Bonvin AM, Tomchick DR, Vendruscolo M, Graham B, Sudhof TC, Rizo J (2015) Dynamic binding mode of a Synaptotagmin-1-SNARE complex in solution. Nat Struct Mol Biol 22:555–564

Camilloni C, Vendruscolo M (2015) Using pseudocontact shifts and residual dipolar couplings as exact NMR restraints for the determination of protein structural ensembles. Biochemistry 54:7470–7476

Carlon A, Ravera E, Andralojc W, Parigi G, Murshudov GN, Luchinat C (2016) How to tackle protein structural data from solution and solid state: an integrated approach. Progress NMR Spectrosc 92–93:54–70

Cavalli A, Vendruscolo M (2015) Analysis of the performance of the CHESHIRE and YAPP methods at CASD-NMR round 3. J Biomol NMR 62:503–509

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

Cavanagh J, Fairbrother WJ, Palmer AG III, Rance M, Skelton NJ (2007) Protein NMR spectroscopy. Principles and practice. Academic Press, SanDiego

Chen J, Im W, Brooks CL III (2004) Refinement of NMR structures using implicit solvent and advanced sampling techniques. J Am Chem Soc 126:16038–16047

Chen WN, Loscha KV, Nitsche C, Graham B, Otting G (2014) The dengue virus NS2B-NS3 protease retains the closed conformation in the complex with BPTI. FEBS Lett 588:2206–2211

Doreleijers JF, Sousa da Silva AW, Krieger E, Nabuurs SB, Spronk CA, Stevens TJ, Vranken WF, Vriend G, Vuister GW (2012) CING: an integrated residue-based structure validation program suite. J Biomol NMR 54:267–283

Feig M, Im W, Brooks CL III (2004) Implicit solvation based on generalized Born theory in different dielectric environments. J Chem Phys 120:903–911

Fossi M, Oschkinat H, Nilges M, Ball LJ (2005) Quantitative study of the effects of chemical shift tolerances and rates of SA cooling on structure calculation from automatically assigned NOE data. J Magn Reson 175:92–102

Gaponenko V, Sarma SP, Altieri AS, Horita DA, Li J, Byrd RA (2004) Improving the accuracy of NMR structures of large proteins using pseudocontact shifts as long/range restraints. J Biomol NMR 28:205–212

Gochin M, Roder H (1995) Use of pseudocontact shifts as a structural constraint for macromolecules in solution. Bull Magn Reson 17:1–4

Hass MA, Ubbink M (2014) Structure determination of protein-protein complexes with long-range anisotropic paramagnetic NMR restraints. Curr Opin Struct Biol 24:45–53

Herrmann T, Güntert P, Wüthrich K (2002) Protein NMR structure determination with automated NOE assignment using the new software CANDID and the torsion angle dynamics algorithm DYANA. J Mol Biol 319:209–227

Huang YJ, Rosato A, Singh G, Montelione GT (2012) RPF—a quality assessment tool for protein NMR structures. Nucleic Acids Res 40:W542–W546

Hulsker R, Baranova MV, Bullerjahn GS, Ubbink M (2008) Dynamics in the transient complex of plastocyanin-cytochrome f from Prochlorothrix hollandica. J Am Chem Soc 130:1985–1991

Jao CC, Hedge BG, Chen J, Haworth IS, Langen R (2008) Structure of membrane-bound alpha-synuclein from site-directed spin labeling and computational refinement. Proc Natl Acad Sci USA 105:19666–19671

Jaroniec CP (2015) Structural studies of proteins by paramagnetic solid-state NMR spectroscopy. J Magn Reson 253:50–59

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

Karplus M (1959) Contact electron-spin coupling of nuclear magnetic moments. J Chem Phys 30:11–15

Lange OF, Rossi P, Sgourakis NG, Song Y, Lee HW, Aramini JM, Ertekin A, Xiao R, Acton TB, Montelione GT, Baker D (2012) Determination of solution structures of proteins up to 40 kDa using CS-Rosetta with sparse NMR data from deuterated samples. Proc Natl Acad Sci USA 109:10873–10878

Li J, Pilla KB, Li Q, Zhang Z, Su X, Huber T, Yang J (2013) Magic angle spinning NMR structure determination of proteins from pseudocontact shifts. J Am Chem Soc 135:8294–8303

Linge JP, Nilges M (1999) Influence of non-bonded parameters on the quality of NMR structures: a new force field for NMR structure calculation. J Biomol NMR 13:51–59

Linge JP, Williams MA, Spronk CAEM, Bonvin AMJJ, Nilges M (2003) Refinement of protein structures in explicit solvent. Proteins: Struct, Funct, Bioinf 50:496–506

Mao B, Tejero R, Baker D, Montelione GT (2014) Protein NMR structures refined with Rosetta have higher accuracy relative to corresponding X-ray crystal structures. J Am Chem Soc 136:1893–1906

Mareuil F, Malliavin TE, Nilges M, Bardiaux B (2015) Improved reliability, accuracy and quality in automated NMR structure calculation with ARIA. J Biomol NMR 62:425–438

McConnell HM, Robertson RE (1958) Isotropic nuclear resonance shifts. J Chem Phys 29:1361–1365

Montelione GT, Nilges M, Bax A, Guntert P, Herrmann T, Richardson JS, Schwieters CD, Vranken WF, Vuister GW, Wishart DS, Berman HM, Kleywegt GJ, Markley JL (2013) Recommendations of the wwPDB NMR validation task force. Structure 21:1563–1570

Muntener T, Haussinger D, Selenko P, Theillet FX (2016) In-cell protein structures from 2D NMR Experiments. J Phys Chem Lett 7:2821–2825

Nabuurs SB, Nederveen AJ, Vranken W, Doreleijers JF, Bonvin AM, Vuister GW, Vriend G, Spronk CA (2004) DRESS: a database of REfined solution NMR structures. Proteins: Struct, Funct, Bioinf 55:483–486

Nabuurs SB, Spronk CA, Vuister GW, Vriend G (2006) Traditional biomolecular structure determination by NMR spectroscopy allows for major errors. PLoS Comput Biol 2:e9

Nederveen AJ, Doreleijers JF, Vranken W, Miller Z, Spronk CA, Nabuurs SB, Güntert P, Livny M, Markley JL, Nilges M, Ulrich EL, Kaptein R, Bonvin AM (2005) RECOORD: a recalculated coordinate database of 500 + proteins from the PDB using restraints from the BioMagResBank. Proteins: Struct, Funct, Bioinf 59:662–672

Pan BB, Yang F, Ye Y, Wu Q, Li C, Huber T, Su XC (2016) 3D structure determination of a protein in living cells using paramagnetic NMR spectroscopy. Chem Commun (Camb) 52:10237–10240

Ragan TJ, Fogh RH, Tejero R, Vranken W, Montelione GT, Rosato A, Vuister GW (2015) Analysis of the structural quality of the CASD-NMR 2013 entries. J Biomol NMR 62:527–540

Rinaldelli M, Ravera E, Calderone V, Parigi G, Murshudov GN, Luchinat C (2014) Simultaneous use of solution NMR and X-ray data REFMAC5 for joint refinement/detection of structural differences. Acta Cryst D D70:958–967

Rinaldelli M, Carlon A, Ravera E, Parigi G, Luchinat C (2015) FANTEN: a new web-based interface for the analysis of magnetic anisotropy-induced NMR data. J Biomol NMR 61:21–34

Rodriguez-Castañeda F, Haberz P, Leonov A, Griesinger C (2006) Paramagnetic tagging of diamagnetic proteins for solution NMR. Magn Reson Chem 44:S10–S16

Rosato A, Aramini JM, Arrowsmith C, Bagaria A, Baker D, Cavalli A, Doreleijers JF, Eletsky A, Giachetti A, Guerry P, Gutmanas A, Guntert P, He Y, Herrmann T, Huang YJ, Jaravine V, Jonker HR, Kennedy MA, Lange OF, Liu G, Malliavin TE, Mani R, Mao B, Montelione GT, Nilges M, Rossi P, van der Schot G, Schwalbe H, Szyperski TA, Vendruscolo M, Vernon R, Vranken WF, de Vries S, Vuister GW, Wu B, Yang Y, Bonvin AMJJ (2012) Blind testing of routine, fully automated determination of protein structures from NMR data. Structure 20:227–236

Rosato A, Tejero R, Montelione GT (2013) Quality assessment of protein NMR structures. Curr Opin Struct Biol 23:715–724

Ryu H, Lim G, Sung BH, Lee J (2016) NMRe: a web server for NMR protein structure refinement with high-quality structure validation scores. Bioinformatics 32:611–613

Saccenti E, Rosato A (2008) The war of tools: how can NMR spectroscopists detect errors in their structures? J Biomol NMR 40:251–261

Schmitz C, Vernon R, Otting G, Baker D, Huber T (2012) Protein structure determination from pseudocontact shifts using ROSETTA. J Mol Biol 416:668–677

Serrano P, Dutta SK, Proudfoot A, Mohanty B, Susac L, Martin B, Geralt M, Jaroszewski L, Godzik A, Elsliger M, Wilson IA, Wuthrich K (2016) NMR in structural genomics to increase structural coverage of the protein universe. FEBS J. doi:10.1111/febs.13751

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

Shen Y, Delaglio F, Cornilescu G, Bax A (2009) TALOS + : a hybrid method for predicting protein backbone torsion angles from NMR chemical shifts. J Biomol NMR 44:213–223

Su XC, Otting G (2010) Paramagnetic labelling of proteins and oligonucleotides for NMR. J Biomol NMR 46:101–112

Synder DA, Chen Y, Denissova NG, Acton T, Aramini JM, Ciano M, Karlin R, Liu J, Manor P, Rajan PA, Rossi P, Swapna GV, Xiao R, Rost B, Hunt J, Montelione GT (2005) Comparisons of NMR spectral quality and success in crystallization demonstrate that NMR and X-ray crystallography are complementary methods for small protein structure determination. J Am Chem Soc 127:16505–16511

Tang Y, Huang YJ, Hopf TA, Sander C, Marks DS, Montelione GT (2015) Protein structure determination by combining sparse NMR data with evolutionary couplings. Nat Methods 12:751–754

Tjandra N, Bax A (1997) Direct measurement of distances and angles in biomolecules by NMR in a diluite liquid crystalline medium. Science 278:1111–1114

van der Schot G, Bonvin AM (2015) Performance of the WeNMR CS-Rosetta3 web server in CASD-NMR. J Biomol NMR 62:497–502

Vernon R, Shen Y, Baker D, Lange OF (2013) Improved chemical shift based fragment selection for CS-Rosetta using Rosetta3 fragment picker. J Biomol NMR 57:117–127

Vuister GW, Fogh RH, Hendrickx PM, Doreleijers JF, Gutmanas A (2014) An overview of tools for the validation of protein NMR structures. J Biomol NMR 58:259–285

Wassenaar TA, van Dijk M, Loureiro-Ferreira N, van der Schot G, de Vries SJ, Schmitz C, van der Zwan J, Boelens R, Giachetti A, Ferella L, Rosato A, Bertini I, Herrmann T, Jonker HRA, Bagaria A, Jaravine V, Guntert P, Schwalbe H, Vranken WF, Doreleijers JF, Vriend G, Vuister GW, Franke D, Kikhney A, Svergun DI, Fogh RH, Ionides J, Laue ED, Spronk C, Jurksa S, Verlato M, Badoer S, Dal Pra S, Mazzucato M, Frizziero E, Bonvin AMJJ (2012) WeNMR: structural biology on the grid. J Grid Comput 10:743–767

Wüthrich K (1986) NMR of proteins and nucleic acids. Wiley, New York

Xia B, Tsui V, Case DA, Dyson HJ, Wright PE (2002) Comparison of protein solution structures refined by molecular dynamics simulation in vacuum, with a generalized Born model, and with explicit water. J Biomol NMR 22:317–331

Yagi H, Pilla KB, Maleckis A, Graham B, Huber T, Otting G (2013) Three-dimensional protein fold determination from backbone amide pseudocontact shifts generated by lanthanide tags at multiple sites. Structure 21:883–890

Yee AA, Savchenko A, Ignatchenko A, Lukin J, Xu X, Skarina T, Evdokimova E, Liu CS, Semesi A, Guido V, Edwards AM, Arrowsmith CH (2005) NMR and X-ray crystallography, complementary tools in structural proteomics of small proteins. J Am Chem Soc 127:16512–16517

2.3.2 Protein Structure Prediction Assisted with Sparse NMR Data in CASP13

Davide Sala[1,2,#], *Yuanpeng Janet Huang*[3,#], *Casey A. Cole*[4], *David Snyder*[5], *Gaohua Liu*[3,6], *Yojiro Ishida*[3,7], *G.V.T. Swapna*[3], *P. Brock*[8], *C. Sander*[8], *Krzysztof Fidelis*[9], *Andriy Kryshtafovych*[9], *Masayori Inouye*[7], *Roberto Tejero*[10], *Homay Valafar*[4], *Antonio Rosato*[1,2*], *Gaetano T. Montelione*[3,7,*]

[#] These authors are designated as co-first authors.

[1] Magnetic Resonance Center, University of Florence, 50019 Sesto Fiorentino, Italy

[2] Department of Chemistry, University of Florence, 50019 Sesto Fiorentino, Italy

[3] Center for Advanced Biotechnology and Medicine, and Department of Molecular Biology and Biochemistry, Rutgers, The State University of New Jersey, Piscataway, NJ 08854, USA

[4] Department of Computer Science & Engineering, University of South Carolina, Columbia, South Carolina 29208

[5] Department of Chemistry, College of Science and Health, William Paterson University, 300 Pompton Rd, Wayne, NJ, USA, 07470

[6] Nexomics Biosciences, 1200 Florence Columbus Rd., Bordentown, NJ 08505, USA

[7] Department of Biochemisty and Molecular Biology, The Roberto Wood Johnsone Medical School, Rutgers, The State University of New Jersey, Piscataway, NJ 08854, USA.

[8] Harvard Medical School, Boston, MA, 02115 USA

[9] Genome Center, University of California, Davis, CA 95616, USA

[10] Departamento de Quimica Fisica, Universidad de Valencia, Avenida Dr. Moliner 50, 46100 Burjassot, Valencia, SPAIN

Accepted - *Proteins* (2019). doi:10.1002/prot.25837

Received: 9 September 2019 | Revised: 26 September 2019 | Accepted: 27 September 2019

DOI: 10.1002/prot.25837

PROTEINS WILEY

RESEARCH ARTICLE

Protein structure prediction assisted with sparse NMR data in CASP13

Davide Sala[1,2] | Yuanpeng Janet Huang[3,4] | Casey A. Cole[5] | David Snyder[6] | Gaohua Liu[3,7] | Yojiro Ishida[3,8] | Gurla V. T. Swapna[3] | Kelly P. Brock[9] | Chris Sander[10,11] | Krzysztof Fidelis[12] | Andriy Kryshtafovych[12] | Masayori Inouye[8] | Roberto Tejero[13] | Homayoun Valafar[5] | Antonio Rosato[1,2] | Gaetano T. Montelione[3,4,8]

[1]Magnetic Resonance Center, University of Florence, Sesto Fiorentino, Italy

[2]Department of Chemistry, University of Florence, Sesto Fiorentino, Italy

[3]Center for Advanced Biotechnology and Medicine, and Department of Molecular Biology and Biochemistry, Rutgers, The State University of New Jersey, Piscataway, New Jersey

[4]Department of Chemistry and Center for Biotechnology and Interdisciplinary Studies, Rensselaer Polytechnic Institute, Troy, New York

[5]Department of Computer Science & Engineering, University of South Carolina, Columbia, South Carolina

[6]Department of Chemistry, College of Science and Health, William Paterson University, Wayne, New Jersey

[7]Nexomics Biosciences, Bordentown, New Jersey

[8]Department of Biochemistry and Molecular Biology, The Robert Wood Johnson Medical School, Rutgers, The State University of New Jersey, Piscataway, New Jersey

[9]Department of Systems Biology, Harvard Medical School, Boston, Massachusetts

[10]Department of Cell Biology, Harvard Medical School, Boston, Massachusetts

[11]cBio Center, Dana-Farber Cancer Institute, Boston, Massachusetts

[12]Genome Center, University of California, Davis, California

[13]Departamento de Quimica Fisica, Universidad de Valencia, Valencia, Spain

Correspondence
Gaetano T. Montelione, CIBS, Rensselaer Polytechnic Institute, 110 Eighth St, Troy, NY 12180.
Email: monteg3@rpi.edu.

Antonio Rosato, University of Florence, 50019 Sesto Fiorentino, Italy.
Email: rosato@cerm.unifi.it

Funding information
National Institute of General Medical Sciences, Grant/Award Numbers: 1R01GM120574, P20 RR-016461, R01GM100482

Peer Review
The peer review history for this article is available at https://publons.com/publon/10.1002/prot.25837.

Abstract

CASP13 has investigated the impact of sparse NMR data on the accuracy of protein structure prediction. NOESY and ^{15}N-^{1}H residual dipolar coupling data, typical of that obtained for ^{15}N,^{13}C-enriched, perdeuterated proteins up to about 40 kDa, were simulated for 11 CASP13 targets ranging in size from 80 to 326 residues. For several targets, two prediction groups generated models that are more accurate than those produced using baseline methods. Real NMR data collected for a de novo designed protein were also provided to predictors, including one data set in which only backbone resonance assignments were available. Some NMR-assisted prediction groups also did very well with these data. CASP13 also assessed whether incorporation of sparse NMR data improves the accuracy of protein structure prediction relative to nonassisted regular methods. In most cases, incorporation of sparse, noisy NMR data results in models with higher accuracy. The best NMR-assisted models were also compared with the best regular predictions of any CASP13 group for the same target. For six of 13 targets, the most accurate model provided by any NMR-assisted

Davide Sala and Yuanpeng Janet Huang are as co-first authors.

prediction group was more accurate than the most accurate model provided by any regular prediction group; however, for the remaining seven targets, one or more regular prediction method provided a more accurate model than even the best NMR-assisted model. These results suggest a novel approach for protein structure determination, in which advanced prediction methods are first used to generate structural models, and sparse NMR data is then used to validate and/or refine these models.

KEYWORDS

CASP, contact prediction, protein modeling, residual dipolar coupling, simulated NMR spectra, sparse NMR data, structure prediction

1 | INTRODUCTION

Since its inception, CASP has been a driving force in the field of contact prediction and contact-directed modeling (see eg References 1-5. Conceptually, even a few accurate native contacts could reliably guide de novo fold predictions, or provide valuable information for selecting among alternate models. During the CASP10 International Meeting, it was suggested that rather than using predicted contacts, which at the time were not very reliable, it might be more productive to explore the impact of a few real experimental contacts, as can be obtained from NMR, cross-linking, fluorescence energy transfer, or other experimental methods. This concept developed in CASP11 into the first NMR-assisted contact prediction experiment.[6-8] In the meantime, the accuracy of contact prediction based on evolutionary sequence covariance analysis and machine learning has increased dramatically,[5,9-14] making the original proposal of replacing predicted contacts with real, sparse experimental contacts moot. None the less, the concept of combining sparse experimental data with sophisticated modeling methods is an important and emerging area of integrative structural biology, and CASP provides an important venue for testing and developing such hybrid methods. In CASP12 and CASP13, this integrative approach was also explored using small angle X-ray scattering[15] and chemical cross-link data.[15] Such integrative data-driven protein prediction is evolving into an important approach for structural biology.

The NMR Community has also explored automated NMR structure analysis in the context of the Critical Assessment of Protein Structure Determination by NMR (CASD-NMR). In this series of studies,[16-19] NOESY peak lists and NMR resonance assignments for 20 small proteins were distributed to several groups developing fully automated nuclear Overhauser effect spectroscopy (NOESY) assignment and structure determination methods. In the first phase, CASD-NMR 2010, NOESY peak lists for 10 proteins were preprocessed to be relatively free of noise peaks, and automated structure determination was carried out in a blind fashion, without knowledge of the manually refined reference structure. It was observed that with such data, several fully automated NOESY analysis program could consistently deliver structures with backbone rmsd's < 2.5 Å from the manually refined reference structure, demonstrating the feasibility of routine, fully automated protein structure determination by NMR. These results were extended in CASD-NMR 2013, a similar blinded study using uncurated, noisy NOESY peak lists. Across the entire set of more than 140 models submitted in this phase, 70% of all entries had a backbone accuracy relative to the reference NMR structure better than 1.5 Å backbone rmsd, with some methods having up to 100% of their submitted models within 1.5 Å rmsd. However, using these uncurated NOESY peak lists, some automated structure determination methods did not converge for some targets. These studies provide benchmark results demonstrating strengths and weaknesses of several programs for fully automated NOESY analysis and structure generation of small (<15 kDa) proteins.

While these CASD-NMR studies were very successful with these relatively small proteins, determining larger-sized protein (20-70 kDa) structures by solution NMR is extremely challenging but highly feasible.[20,21] For such larger proteins, perdeuteration becomes necessary to circumvent the efficient spin relaxation properties resulting from their slow rotational correlation times. Backbone and sidechain amide hydrogens (H^N) can be exchanged back into the protein structure, allowing collection of H^N-H^N NOE data, and some methyl and/or aromatic groups can be protonated by biosynthetic methods.[22,23] However, aside from such selectively protonated side-chain moieties, replacing most protons in the protein structure with deuterons also eliminates most long-range and sidechain NOESY information. The difficulty in determining accurate structures with no, or limited, side-chain information (ie, sparse NMR data) is a major technological challenge to the modeling community that currently limits routine application of solution NMR to larger systems.

In CASP11, we explored this challenge together with the global CASP community,[6-8] by providing interatomic contacts derived from sparse nuclear Overhauser effect (NOE) data simulated from the X-ray crystal structure coordinates of 19 CASP template-free modeling targets, assuming perdeuteration with selective protonation of backbone amide and certain methyl groups. These targets ranged in size from 108 to 462 residues. The results were compared with baseline modeling results using some of the more successful automated structure determination programs assessed in CASD-NMR, including the *ASDP* program.[24,25] While most NMR-assisted CASP11 methods could not provide accurate models using these sparse experimental data, a few groups (eg, Lee, Baker) submitted models for several targets that were more accurate than those generated using "conventional" baseline automated NOESY analysis methods. These results demonstrate the strong synergy between the computational NMR and protein prediction communities, as each has the potential to learn from one another.

In CASP13, we extended our NMR-assisted structure prediction study begun in CASP11. NMR data were simulated from CASP free modeling targets with realistic degrees of incompleteness and noise, typical to that observed in real NMR spectra of perdeuterated, selectively protonated proteins. These free modeling targets ranged from 80 to 326 residues. In addition to contacts based on simulated 3D NOESY data, simulated residual ^{15}N-^{1}H dipolar coupling (RDC) data, and dihedral restraints as derived from backbone chemical shift data, for a subset of residues were also provided. In some cases, contact predictions from evolutionary sequence covariance analysis were also provided to predictors. Two real NMR data sets were also made available to the CASP13 prediction community. These results further drive the field of integrated protein structure modeling by exploring the impact of sparse experimental data in enhancing the power of protein structure prediction methods.

2 | METHODS

2.1 | Experimental NMR structure determination

NMR studies were performed using a uniformly ^{15}N,^{13}C-enriched sample of a de novo designed protein, named foldit3,[26] CASP13 target 1008. The synthetic codon optimized gene (Genscript, Inc), designed to exclude ACA nucleotide sequences,[27,28] was cloned into plasmid pET15TEV_NESG.[29] The resulting protein product includes a short N-terminal 6xHis purification tag, followed by a TEV protease cleavage site, which was removed prior to data collection. Details of the production and characterization of this sample have been described elsewhere,[26] and are also provided in the Supporting Information. Homogeneity (>97%) was validated by SDS polyacrylamide gel electrophoresis. The purified protein was dialyzed against 20 mM potassium phosphate, pH 6.5, and the protein concentration was adjusted to between 0.3 and 0.4 mM for NMR studies.

All NMR spectra were recorded at 25°C using cryogenic NMR probes. NMR data were collected on a Bruker AVANCE III 600 MHz spectrometer, processed using the program *NMRPipe*,[30] and analyzed using the programs *SPARKY*[31] and *XEASY*.[32] Spectra were referenced to external DSS. Sequence-specific resonance assignments were determined using *AutoAssign* software[33,34] together with interactive manual analysis. NMR data collection included simultaneous ^{15}N,^{13}C-edited 3D NOESY and ^{15}N-edited 3D NOESY, both recorded with mixing time τ_m = 120 ms. Backbone dihedral angle constraints were then derived from the assigned chemical shifts using the program *TALOS_N*[16-19] for residues located in well-defined secondary structure elements. The programs *ASDP*[24,25] and *CYANA*[35,36] were used to automatically assign NOEs and to generate 3D structures, respectively. NOESY peak lists used for NMR-assisted predictions in CASP13 were all based on fully automated NOESY peak assignment with *ASDP*.

For structure refinement, RPF analysis,[37,38] comparing observed and predicted NOESY peak lists, was used to guide iterative cycles of noise/artifact peak removal, peak picking, and NOESY peak assignments. The 20 conformers with the lowest target *CYANA* function value were then refined by restrained molecular dynamics in explicit water[39] using the program *CNS*.[40] Structural statistics and global structure quality factors were assessed using the *PSVS* 1.5[41] and *PDBStat*[42] software packages. The global goodness-of-fit of the final structure ensembles with the NOESY peak list data, the NMR DP score, was determined using the *RPF* program.[37,38]

2.2 | Baseline modeling with *ASDP-Cyana*

Baseline modeling was carried out using state-of-the-art "conventional" methods for modeling protein structures from NMR data. NMR structures were modeled using the *ASDP* program for NOESY peak assignment, together with *Cyana* for structure generation from the resulting restraints. This pipeline, described in detail elsewhere,[25] was one of the top performing automated NOESY analysis methods in the CASD-NMR experiments.[16-19] *ASDP* uses expert system methods to assign NOESY cross peaks, and to generate distance restraints. These restraints are then input to a structure generation program. In this case, structures were generated from the restraints using the restrained molecular dynamics in torsion angle space module of the program *Cyana*. The resulting intermediate structure models are then used to iteratively rule in/rule out additional NOESY cross peak assignments.[24] The final structures generated in several cycles of NOESY peak assignment and model generation are then refined with these NMR restraints active using *Rosetta*, with loop remodeling and core repacking, as described elsewhere (Mao et al. 2013).

2.3 | CASP assessment units and assessment metrics

Simulated sparse NMR data were provided to CASP13 predictors for 11 proteins and protein domains (Table 1, CASP Targets), ranging in size from 80 to 326 residues. In addition, real NMR data were provided for one target, protein T1008 (aka, foldit3),[26] of 80 residues. In this case, two different real NMR were made available, differing in the completeness of the assignment of the NMR frequencies. Submitted prediction models were assessed by standard CASP metrics.[43-45] Summed or averaged *Z* scores for each metric were computed by the CASP Prediction Center.[46] NMR DP scores,[37] comparing the short ^{1}H-^{1}H distances in prediction models with the NOESY peak list, and ^{15}N-^{1}H RDC Q scores[47,48] were provided for each model submitted to the CASP Prediction Center for statistical analysis.

Accuracy of submitted models was evaluated at the domain level. *Z* scores were calculated for a total of 14 assessment units (domains) listed in the third column of Table 1. The *Z* score analysis excluded the combined domain constructs N0957s1-D1.D2 and N0989-D1.D2, and also the n1008 target for which full NMR assignments were available, although structure accuracy metrics for these are also available

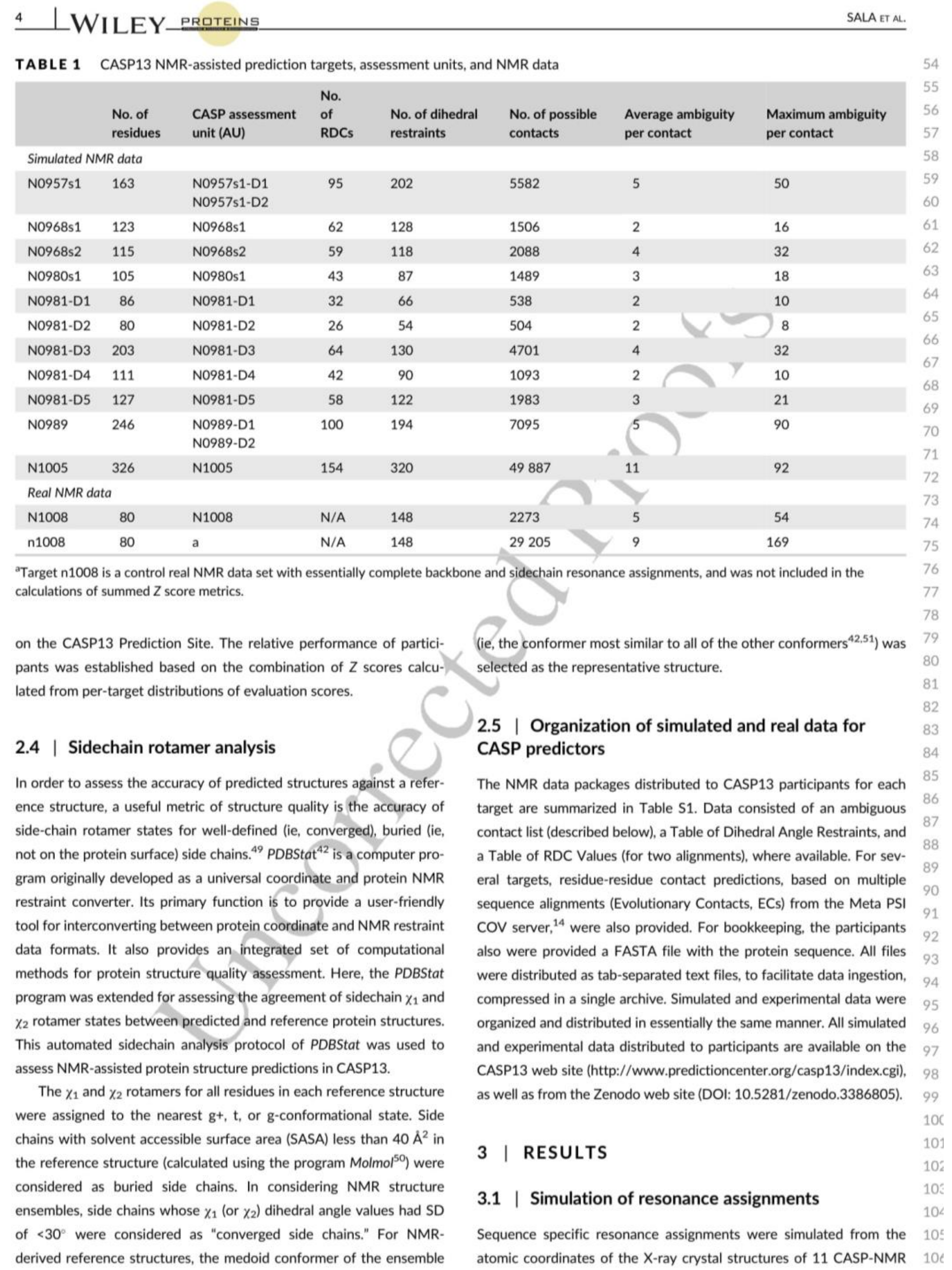


TABLE 1 CASP13 NMR-assisted prediction targets, assessment units, and NMR data

	No. of residues	CASP assessment unit (AU)	No. of RDCs	No. of dihedral restraints	No. of possible contacts	Average ambiguity per contact	Maximum ambiguity per contact
Simulated NMR data							
N0957s1	163	N0957s1-D1 N0957s1-D2	95	202	5582	5	50
N0968s1	123	N0968s1	62	128	1506	2	16
N0968s2	115	N0968s2	59	118	2088	4	32
N0980s1	105	N0980s1	43	87	1489	3	18
N0981-D1	86	N0981-D1	32	66	538	2	10
N0981-D2	80	N0981-D2	26	54	504	2	8
N0981-D3	203	N0981-D3	64	130	4701	4	32
N0981-D4	111	N0981-D4	42	90	1093	2	10
N0981-D5	127	N0981-D5	58	122	1983	3	21
N0989	246	N0989-D1 N0989-D2	100	194	7095	5	90
N1005	326	N1005	154	320	49 887	11	92
Real NMR data							
N1008	80	N1008	N/A	148	2273	5	54
n1008	80	a	N/A	148	29 205	9	169

[a]Target n1008 is a control real NMR data set with essentially complete backbone and sidechain resonance assignments, and was not included in the calculations of summed *Z* score metrics.

on the CASP13 Prediction Site. The relative performance of participants was established based on the combination of *Z* scores calculated from per-target distributions of evaluation scores.

2.4 | Sidechain rotamer analysis

In order to assess the accuracy of predicted structures against a reference structure, a useful metric of structure quality is the accuracy of side-chain rotamer states for well-defined (ie, converged), buried (ie, not on the protein surface) side chains.[49] *PDBStat*[42] is a computer program originally developed as a universal coordinate and protein NMR restraint converter. Its primary function is to provide a user-friendly tool for interconverting between protein coordinate and NMR restraint data formats. It also provides an integrated set of computational methods for protein structure quality assessment. Here, the *PDBStat* program was extended for assessing the agreement of sidechain χ_1 and χ_2 rotamer states between predicted and reference protein structures. This automated sidechain analysis protocol of *PDBStat* was used to assess NMR-assisted protein structure predictions in CASP13.

The χ_1 and χ_2 rotamers for all residues in each reference structure were assigned to the nearest g+, t, or g-conformational state. Side chains with solvent accessible surface area (SASA) less than 40 Å^2 in the reference structure (calculated using the program *Molmol*[50]) were considered as buried side chains. In considering NMR structure ensembles, side chains whose χ_1 (or χ_2) dihedral angle values had SD of <30° were considered as "converged side chains." For NMR-derived reference structures, the medoid conformer of the ensemble (ie, the conformer most similar to all of the other conformers[42,51]) was selected as the representative structure.

2.5 | Organization of simulated and real data for CASP predictors

The NMR data packages distributed to CASP13 participants for each target are summarized in Table S1. Data consisted of an ambiguous contact list (described below), a Table of Dihedral Angle Restraints, and a Table of RDC Values (for two alignments), where available. For several targets, residue-residue contact predictions, based on multiple sequence alignments (Evolutionary Contacts, ECs) from the Meta PSI COV server,[14] were also provided. For bookkeeping, the participants also were provided a FASTA file with the protein sequence. All files were distributed as tab-separated text files, to facilitate data ingestion, compressed in a single archive. Simulated and experimental data were organized and distributed in essentially the same manner. All simulated and experimental data distributed to participants are available on the CASP13 web site (http://www.predictioncenter.org/casp13/index.cgi), as well as from the Zenodo web site (DOI: 10.5281/zenodo.3386805).

3 | RESULTS

3.1 | Simulation of resonance assignments

Sequence specific resonance assignments were simulated from the atomic coordinates of the X-ray crystal structures of 11 CASP-NMR targets (excluding targets N1008 and n1008 for which real NMR data were generated for this study). First, any selenomethione (MSE) residues in the original PDB coordinate file of the reference X-ray structures were changed to methionine (MET). Hydrogen atoms were then added to the coordinates of X-ray crystal structures with the program *Reduce*.[52] The resulting coordinates were then used to simulate ^{1}H, ^{13}C, and ^{15}N chemical shift values using the program *SHIFTX2*.[53] As one goal of the CASP13 experiment is to explore the impact of NMR data obtainable from NMR studies of larger proteins on the accuracy of structure prediction, resonance assignments were all simulated assuming a perdeuterated protein sample with typical selective reprotonation.[22,23] Specifically, only backbone HN, N, C$^\alpha$, and C′ atoms, sidechain C$^\beta$ atoms, and the C and H atoms of Ile(δ1), Leu, Val, and Ala methyl groups were included in the simulated chemical shift table. This proton labeling pattern corresponds to that provided by the application of typical selective labeling strategies used for studies of proteins in the size range 20-70 kDa. It was assumed that individual stereospecific assignments of the isopropyl methyls of Leu and Val were not available, and no corrections were made to account for deuterium isotope effects on bound ^{13}C chemical shift values.

3.2 | Simulation of NOESY peak lists

In order to create incomplete NOESY peak lists like those observed with real NMR data, a number of resonance assignments were deleted prior to simulating the NOESY spectra. In this process, illustrated for a F1 representative target in Figure 1, our expertise in protein NMR studies was used to simulate the effects of line broadening due to conformational dynamics and/or weak spectra in causing "missing resonances." First, we created a list containing selected regions for each CASP target proposed to exhibit missing resonances and/or NOESY cross peaks (eg, yellow residues in Figure 1A). The choice of region to select was made so as to simulate the effects of local dynamics which could plausibly result in exchange broadening. This typically included surface loop residues and/or potentially dynamic secondary structures. Within each of these regions, we randomly selected 25% of the residues and deleted all chemical shift assignments for these residues (eg, red residues in Figure 1B).

These chemical shifts were then used, together with ^{1}H-^{1}H distances from the atomic coordinates, to simulate 3D ^{13}C-edited and 3D ^{15}N-edited NOESY peak lists (frequencies and intensities) which would be obtained for perdeuterated ^{13}C,^{15}N-enriched, backbone ^{1}H^N and ILVA ^{13}C-^{1}H$_3$ methyl labeled proteins. For all potential ^{1}H-^{1}H NOEs, a summation distance[42] was calculated from the atomic coordinates of all degenerate proton resonances,

$$r_{ij} = \left[\sum_{k=1}^{n}\left(\frac{1}{r_{ij,k}}\right)^{6}\right]^{-1/6} \qquad (1)$$

If the resulting r_{ij} was less than a cutoff distance D_{cutoff}, 3D ^{13}C- or ^{15}N-edited NOE cross peaks were simulated with intensity of 10 000/r^6_{ij}. The maximum observable interproton distance was set to 5 Å. NOESY "cross peaks" (frequencies and intensities in a NOESY peak list), representing these short interproton distances, were then created between these resonances. Adjacent NOESY peaks created within tolerances of 0.02 ppm for the direct H dimension, 0.2 ppm for the C/N, and 0.03 ppm for the indirect H dimension were merged, and the corresponding resonance frequencies were averaged, to simulate overlapped regions in the NOESY spectra. Short ^{1}H-^{1}H distances between protons in residues for which resonance assignments were deleted did not generate any NOESY peak, corresponding to False Negative, FNs.

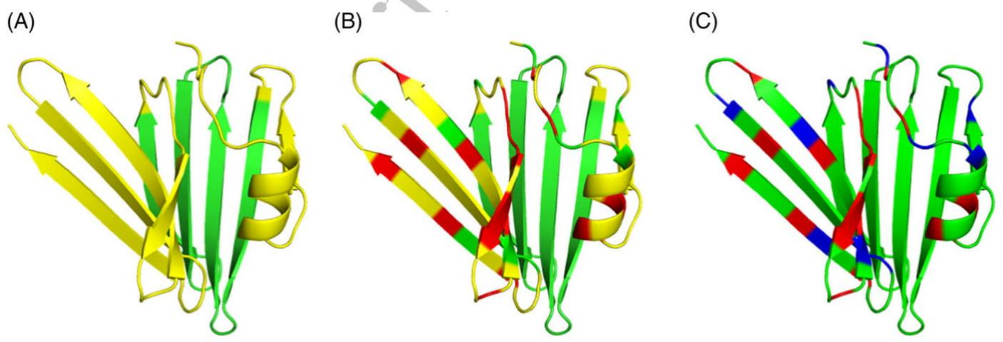


FIGURE 1 Schematic depiction of process for simulating missing NOESY peak and resonance assignment data for CASP-NMR target N0968s1. (A) Ribbon representation of the backbone structure of T0968s1 showing the locations (in yellow) of candidate regions for NOESY peak and/or resonance assignment deletion. (B) Residues (red) for which all resonance assignments are deleted prior to generating the NOESY peak list. This results in the elimination of potential NOESY peaks from the simulated spectra due to, for example, conformational exchange broadening. (C) Residues (blue) for which, after the NOESY cross peaks are generated, all resonance assignments are deleted. These data simulate the case where NOESY peaks are observed, but one or more of the corresponding ^{1}H resonances are not assigned. These NOESY cross peaks are present in the NOESY peak list, but cannot be correctly assigned, and often result in incorrect NOESY cross peak assignments due to accidental degeneracies with observable resonances

After simulating these NOESY peak lists (ie, the resonance frequencies and intensities of observable NOESY cross peaks, excluding cross peaks involving resonances that are missing due to exchange broadening), an additional 15% of the residues in the resonance assignment list were randomly selected and their chemical shift values were removed from the chemical shift list, while preserving the corresponding NOESY peaks (eg, blue residues in Figure 1C). This process simulates the situation where the NOESY cross peak is present, but one or both of the corresponding sequence-specific resonance assignments cannot be determined. These NOESY cross peaks cannot be correctly matched to the original ^{1}H-^{1}H pair, and have the potential to be incorrectly assigned. Finally, weak NOESY noise peaks were added to the NOESY peak lists, at frequency positions consistent with assigned resonances, but not corresponding to short ^{1}H-^{1}H distances in the reference structure. These are false positive NOESY peaks, FPs. The contacts indicated by these FP peaks are generally inconsistent with the native protein structure. This similar process of resonance assignment deletions (resulting in FNs and incorrect NOESY cross peak assignments) and random addition of weak NOESY peaks (resulting in FPs) was applied in simulating ^{13}C- and ^{15}N-edited 3D NOESY spectra peak lists from reference X-ray structures of 11 CASP-NMR targets (excluding targets N1008 and n1008 for which real NMR data were available).

NMR data were simulated assuming a monomer structure for all targets. No efforts were made to simulate interfacial X-filtered NOESY data; homodimers were simulated as the single protomer. Targets N0957s1 and N0989 are two-domain proteins, in which each domain is well-defined relative to the other. Hence, the complete two-domain coordinates were used to simulate NOE and RDC data assuming the same static orientation of domains as observed in the corresponding X-ray crystal structures. Target N0980 is a dimer of heterodimers (2:2 tetramer). Since one chain is a small polypeptide, we simply used the coordinates of the single protomer of the larger subunit as the target for simulating NMR data. Target N0981 is a five-domain structure, in which the domains are likely to be independent of one another. Hence, each domain was treated as an independent target, and NOESY and RDC data were simulated separately for each of them.

3.3 | Statistics on simulated and real NOESY peak lists

Statistics on the NOESY peak list data for each target are summarized in Table S1 and Figure S1. The process of simulating NOESY peak lists outlined here provided data sets with properties similar to those generally obtained for uniformly ^{15}N,^{13}C-enriched, perdeuterated proteins with ILVA methyl ^{13}C-^{1}H$_3$ labeling. Analysis of these NOESY peak lists against the reference atomic coordinates showed that 5%-18% of short distances in the reference structures have no corresponding NOESY peak, and 5%-10% of the NOESY peaks in these lists cannot be assigned to any true short ^{1}H-^{1}H distance. For these simulated NOESY peak lists, the data ranged from 3.5 to 9.3 NOESY peaks per residue. This compares to 8.4 NOESY peaks per residue for the sparse real data set N1008, and 43.4 NOESY peaks per residue for the generally complete experimental dataset of n1008. Generally speaking, the low restraint density (10 NOESY peaks/residue) of the simulated data sets, and of the real data set N1008, make these targets challenging for NMR-based structural determination using traditional methods.

These NOESY data sets (summarized in Table S1) are not only sparse (incomplete), but they also include significant numbers of false peaks which cannot be satisfied by the correct structure. For the simulated datasets, 6.2%-9.1% of the peaks in the NOESY spectra cannot be satisfied by the reference structure. For real data set N1008, the fraction of peaks with possible assignments for which none are consistent with the native structure is even higher, 19.7%. This is because this protein sample was fully protonated, but its NOESY peak list was analyzed using only the backbone resonance assignments, including H^N and H$^\alpha$ resonances; cross peaks involving sidechain atoms with chemical shifts similar to backbone atoms were thus often incorrectly assigned (uniquely or ambiguously) as backbone-backbone NOEs, making this real data set particularly challenging. For the real data set n1008, which included essentially complete backbone and sidechain assignments, only about 0.2% of the NOESY peaks cannot be explained by the final, refined solution NMR structure, a hallmark of high quality NMR data and structures.[54]

3.4 | Generation of ambiguous contact lists

The ideal input for NMR-assisted prediction would be the unassigned NOESY peak lists together with sequence-specific NMR assignments and RDC data, as was done in the CASD-NMR project.[16-19] However, in order to reduce the extent of domain specific NMR spectroscopy knowledge required for participation in CASP13, the organizers instead provided these NOESY data as “ambiguous contact lists” (-Figure S2). For each NOESY peak, the ambiguous contact list provides the set of ^{1}H-^{1}H pairs which, within a defined frequency tolerance of matching the NOESY peak frequencies to the chemical shift frequencies, are possible assignments for each NOESY peak.

Ambiguous contact lists were generated by analyzing simulated (or real) NOESY peak lists together with the corresponding resonance assignment lists, without knowledge of the target 3D structure. The resonance assignment list was first modified to simulate the small inconsistencies generally seen between peak frequencies measured in the NOESY spectrum and the corresponding frequencies in the resonance assignment list. Random noise shifts were added to the chemical shift values in each dimension, with SD 0.01 ppm for direct ^{1}H dimension, 0.20 ppm for indirect C/N, and 0.02 ppm for the indirect ^{1}H dimension. 3D NOESY peak lists were then analyzed together with these resonance assignment lists using the Cycle 0 module of the program *ASDP*. This algorithm assigns NOESY cross peaks to one or more potential ^{1}H-^{1}H interactions based on chemical shift matching. These initial assignments use information, based on backbone chemical shift, on the locations of α-helices and β-strands, inter strand alignments, and other topological rules derived automatically from distances within standard secondary structures, to reduce the ambiguity of

NOESY cross peak assignments, as described elsewhere.[24] The Cycle 0 ASDP analysis was executed with match tolerances of 0.03 ppm for H atoms and 0.30 ppm for C/N atoms, and with parameters D_{cutoff} = 5.0 Å and D_{upper} = 7.5 Å.

In this protocol, outlined in Figure 2, only one cycle of *ASDP* was executed. If peaks could be uniquely assigned by the algorithms of *ASDP* cycle 0, the unique assignments for these particular NOESY peaks, including any unique long-range H^N-H^N NOEs important for β-strand alignments, were included in the ambiguous contact lists. Short-range intraresidue and sequential NOESY peak assignments (|i-j| ≤ 2) were excluded. For each remaining NOESY peak, the output of *ASDP* Cycle 0 provided all possible proton pair assignments within the defined resonance frequency match tolerances. In practice, each NOESY peak is assigned to a set of ambiguous ^1H-^1H pair assignments whose chemical shifts are compatible with the resonance frequencies associated with the 3D NOESY peak. In the absence of experimental errors, at least one of these ^1H-^1H pairs should correspond to a short-distance interproton interaction that is consistent with the native protein structure. However, NOESY peaks that arise from unassigned resonances, as well as random noise peaks, will provide a set of ambiguous contacts, or possibly even a unique assignment, none of which are consistent with the native structure. We calibrated the number of added noise peaks so that number of FP contacts due to these random noise peaks did not exceed 10% (except for real data set N1008) of the final ambiguous restraint list. The resulting ambiguous contact lists (Figure S2) were provided to CASP predictors.

3.5 | Backbone dihedral angle restraints

For relatively static protein conformations, backbone chemical shift data can be used to make reliable predictions of backbone dihedral angle values based on statistical assessment against the database of protein chemical shifts and local structures.[53] In this work, we observed that dihedral angle restraints computed using the program Talos_N,[55] from chemical shifts predicted from the atomic coordinates with the program SHIFTX2, were not always consistent with the X-ray crystal structure used to predict the chemical shifts. This probably reflects shortcomings in the accuracy of these chemical shift predictions. In order to provide the kind of restraint data based on backbone chemical shifts that would be available using real NMR data, the backbone dihedral angle restraints for residues with "observed" and "assigned" backbone chemical shifts (ie, for residues that were not deleted from the chemical shift list) were also provided to CASP predictors. These dihedral restraints were provided as ranges, in which two random numbers between 5 and 30° were added and subtracted from the dihedral angle value observed in the reference X-ray structure.

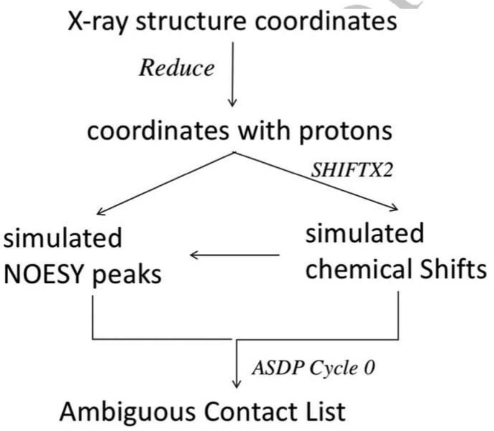


FIGURE 2 Process flow chart for generating ambiguous contact lists for NMR-Assisted CASP13

3.6 | Simulation of ^{15}N-^1H RDC data

RDCs arise from the interaction of two magnetically active nuclei in the presence of the external magnetic field of an NMR instrument. In solution NMR studies, this interaction is normally reduced to zero due to the isotropic tumbling of molecules in their aqueous environment. The introduction of partial order to the molecular alignment reintroduces dipolar interactions by minutely limiting isotropic tumbling. This partial order can be introduced in numerous ways, including inherent magnetic anisotropy susceptibility of molecules, incorporation of artificial tags (such as lanthanides) that exhibit magnetic anisotropy, or using a liquid crystal or otherwise partially ordered aqueous solution.

The RDC interaction phenomenon has been formulated in different ways. To harness the computational synergy of RDC data, in this study we have utilized the matrix formulation of this interaction as shown in Equation (2). The matrix S shown in Equations (2) and (3) represents the Saupe order tensor matrix (the "order tensor") that can be described as a 3×3 symmetric and traceless matrix. $D_{\max}$ in Equation (2) is a nucleus-specific collection of constants, r_{ij} is the separation distance between the two interacting nuclei (in units of Å), and v_{ij} is the corresponding normalized internuclear vector.

$$D_{ij} = \left(\frac{D_{max}}{r_{ij}^3}\right) v_{ij} * S * v_{ij}^T \tag{2}$$

$$S = \begin{bmatrix} S_{xx} & S_{xy} & S_{xz} \\ S_{xy} & S_{yy} & S_{yz} \\ S_{xz} & S_{yz} & S_{zz} \end{bmatrix} \tag{3}$$

$$v_{ij} = \begin{pmatrix} cos(\theta_x) \\ cos(\theta_y) \\ cos(z) \end{pmatrix} \tag{4}$$

The software package *REDCAT*[56,57] used this formalism to simulate ^{15}N-$^1H^N$ RDCs for the target proteins. *REDCAT* uses the protein structure and an order tensor S to calculate RDCs using Equations (2)-

(4). For each of the target protein structures, two different order tensors were calculated using the software package *PALES*.[58] *PALES* utilizes a steric collision model to calculate order tensors in different simulated alignment media. In this work, two different simulated alignment media were utilized: bicelle (wall-like structures) and phage (rod-like structures). The concentration used for both simulations was 0.05 units. The resulting ^{15}N-^{1}H RDC for the "observed" and "assigned" residues of each target were provided to CASP13 predictors.

3.7 | Summary of NMR data

Table 1 also summarizes the numbers of RDCs, dihedral restraints, and ambiguous contacts (NOESY peaks) provided to predictors for each target. The number of all possible atom pair assignments for all NOESY peaks provide by ASDP Cycle 0 ranged from 504 (for target data set N0981-D2) to 49 887 (for target data set N1005). These NOE-based contacts, provided to CASP predictors, had an average of 2-11 possible atom pair assignments per contact, with maximum ambiguity of 92 possible assignments per contact.

3.8 | Assessment of baseline NMR assisted modeling with *ASPD*

Baseline modeling was carried out for each target using the *ASDP* software program for NOESY peak assignment and restraint generation. These "baseline structures" were modeled in "blinded fashion," in which simulated NOESY and RDC data were provided to one of the authors (JH) without her knowledge of the reference structures, and structures were generated from these data using the conventional automated NOESY assignment and modeling algorithms of the *ASDP* software. *ASDP* takes as input the NOESY peak and resonance assignment lists, from which the ambiguous contact lists were derived. For this reason, the *ASDP* baseline calculations used these NOESY peak lists, rather than the ambiguous contact lists, as input. Peaks in these unassigned NOESY peak lists were labeled only by the corresponding resonance frequencies, and did not include any link to the table of chemical shift assignments. NOESY peaks were assigned and disambiguated using *ASDP*, these data were interpreted as calibrated distance restraints, and the resulting assigned distance restraints were used to generate structural models with the software *CYANA*.[36] The resulting structures were further refined using restrained Rosetta refinement,[59] as outlined in the Section 2.

Baseline models were generated using three protocols. Baseline_Group 321 provided five models generated using the simulated (or real) sparse NOESY, dihedral, and RDC data, without EC contact predictions. Baseline_Group 459 provided five models generated using these same data, plus EC contact predictions from the Meta PSI COV server,[14] which were also provided to all predictors. A third set of models was generated for each target using ECs from the EVFold contact prediction pipeline,[60] run locally for this study. Alignments were generated using five jackhmmer[61] iterations against the Uniref100 sequence database (February 2018 release), with multiple normalized bitscore thresholds ranging from 0.1 to 0.9 (with T0981 subsequently run at 0.03). Alignments were chosen based on maximizing both the effective number of sequences and the nongap coverage of each position. Pseudolikelihood maximization was then used to compute evolutionary couplings using the alignments, with the default settings found at the evcouplings.org webserver. These ECs were then combined with NMR data to generate EC-NMR structures, and the top five scoring models were selected. The resulting 15 models for each target (five from NMR alone, five from EC-NMR using Meta PSI COV, and five from EC-NMR using EVFold ECs) were then assessed using the DP score "NMR R-factor" metric, which compares the contact map for the NMR-derived model against the NOESY peak list.[37,38] The five models with highest DP score were then submitted as Baseline Group 313.

These baseline models were then assessed by the CASP Prediction Center. All three baseline groups (ASDP Baseline_Groups 313, 321, and 459) had similar overall accuracy performance based on GDT-TS, GDT_HA, GDT_ALL, GDC_SC, SphereGrinder, and RPF assessment metrics (these metrics are described in References 44 and 45). In general, modeling accuracy was highest for ASDP Baseline_Group 313 (best DP score), followed by ASDP Baseline_Group 459 (with Meta PSI COV ECs), and then ASDP Baseline_Group 321 (without ECs). Interestingly, using the knowledge-based MolProbity assessment score, the highest quality structures were those generated by protocol ASDP Baseline_Group 321 (without ECs), while ASDP Baseline_Group 459 (with PSI COV ECs) had significantly poorer MolProbity scores, suggesting that inclusion of contact predictions in these protocols can distort models from their best atomic packing conformations. For the sake of simplicity, in the remaining analysis we utilize only the ASDP Baseline_Group 321 (without ECs) and ASDP Baseline_Group 459 (with PSI COV ECs) as the baseline comparison results.

3.9 | Initial assessment of NMR assisted predictions

Six CASP13 predictor groups participated in this NMR-guided prediction experiment; Forbidden (122), KIAS-Gdansk (208), Meilerlab (250), UNRES (288), Laufer (431), and wf-Baker-UNRES (492). An initial ranking of the six NMR-guided prediction groups, along with two baseline groups, was done using summed GDT-TS *Z* scores for the first-ranked model submitted for each predictor group (Figure 3A), as described elsewhere.[43,62,63] For the calculation of the summed GDT-TS *Z* scores we used the common convention of setting the *Z* score = −2 for any model with *Z* score ≤ −2.[43,62,63] This is done so as not to heavily penalize the worst models, and to encourage the exploration of new (perhaps less successful) methods in CASP.

Figure 3A demonstrates that two prediction groups (Laufer and Meilerlab) generally provided more accurate models than the baseline groups. The same conclusion was drawn by considering the "best of 5" models from each group, and also for the individual assessment measures GDT_HA, GDT_All, GDC_SC, Sphere Grinder, and RPF. However, in ranking using only MolProbity scores, groups Laufer and Meilerlab are reversed in their relative order; apparently group Meilerlab, using the Rosetta force field and fragment libraries, does a better job of

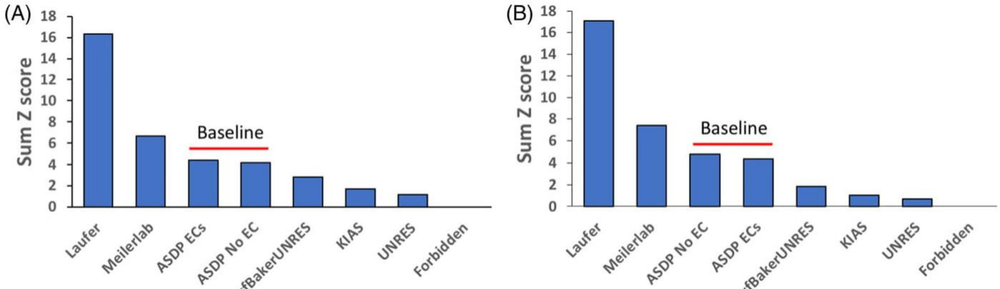


FIGURE 3 Summed *Z* score ranking of CASP13 NMR-assisted predictions. (A) Summed GDT-TS *Z* scores. (B) Weighted summed *Z* scores for multiple metrics, as described in text

generating better packed and more physically plausible conformations, with better MolProbity scores.

3.10 | Principal component analysis on Assessment Metrics

A good CASP prediction should be both similar to its corresponding experimentally derived target structure and physically reasonable. Therefore, the ranking of predicted structures for a given CASP target should incorporate statistics, such as the GDT-TS,[64] quantifying how accurately a structure models a target, as well as measures of biophysical structure quality such as the MolProbity score.[65] Incorporation of multiple measures of structure accuracy and quality into a single ranking involves either folding multiple statistics into a single composite score or using consensus methods to combine rankings based on multiple measures of structure quality and accuracy into a single composite ranking. As in previous template-based modeling CASP experiments,[43,62,63] our final ranking of NMR data assisted predictions in CASP13 combines multiple structure evaluation statistics using a weighted sum. Thus, the question of how to rank predictions of a given target reduces to a question of finding appropriate weights to use in adding together a selection of measurements of structure quality and accuracy.

We chose to use for evaluation of NMR-data assisted structure predictions superposition-dependent global measures of structure accuracy (GDT-HA and GDT-SC), a local superimposition dependent local measure of structure accuracy (SphereGrinder), measurements of the accuracy of interatomic contact areas (CAD-AA) and contact distances (RPF), and a measurement of the physical reasonableness of the structure (MolProbity score).

A recent analysis[44] of protein structure evaluation scores indicates that most methods for structure evaluation are highly correlated. We observed this also for the set of metrics we used for assessing NMR-assisted structure predictions (Table S2). The MolProbity score, however, is less correlated to the other scores, and provides complementary information. While the MolProbity score was distinct, none of the scores was inconsistent with each other according to Friedman's test. In general, models with reasonable accuracy were (as judged by MolProbity score) physically reasonable structures, although some models with good Molprobity scores were not particularly accurate. Inaccurate models with good MolProbity scores have also been observed in assessments of incorrect homology models[66] and of inaccurate CASD-NMR experimental NMR structures.[18,19]

The high correlation between structure evaluation statistics suggests that Principal Component Analysis (PCA) may be a useful ad hoc method to calculate weights for summing multiple measurements of structure quality and accuracy. PCA identified that a composite score of 0.442*Z_GDT_HA + 0.449*Z_GDT_SC + 0.425*Z_RPF + 0.428*Z_SphGrdr + 0.433*Z_CAD_AA + 0.227*Z_MolProbity (where Z_[X] indicates the *z*-score calculated on a per target basis, using the first model provided by each predictor, from quality measure [X]) explains approximately 87% of the variance in structure evaluation scores (Table S3). Setting all *Z*-scores below a certain threshold yields a similar composite statistic via PCA. We rounded the coefficients of the PCA score explaining the highest amount of variance down to the nearest tenths to create the following linear regression: 0.40 for GDT-HA, GDC-SC, RPF, SphereGrinder and CAD-AA and 0.20 for MolProbity. Note that while PCA is a useful tool for constructing a composite metric for assessing structure prediction quality, we do not necessarily expect PCA-derived weights will be similar from one CASP dataset to the next.

3.11 | Overall ranking based on PCA-defined combination of scores

The final ranking of NMR-assisted predictions, using the combined weighted *Z* scores of GDT-HA (0.4), GDT-SC (0.4), RPF(0.4), SphereGrinder (0.4), CAD-AA (0.4), and MolProbity (0.2) is illustrated in Figure 3B for the predictor-designated "first" models. Again, in this analysis *Z* score is set to −2 for any model with *Z* score ≤ −2. Similar results were obtained by selecting the best-scoring model out of the five submitted. These rankings are essentially the same as those obtained using the GDT-TS (or other individual metrics) alone. Relative to the baseline groups, two prediction groups (Laufer and

Meilerlab) generally provided more accurate models, while the remaining four groups had somewhat poorer accuracy performance. For the 14 assessment units (AUs, listed in column 3 of Table 1), the top prediction groups were Laufer 431 for seven AUs, Meilerlab 250 for three AUs, ASDP Baseline_Group 321 (without ECs) for two AUs, and ASDP Baseline_Group 459 (with PSI COV ECs) for two AUs, respectively. Groups wf-Baker-UNRES and KIAS-Gdansk also outperformed the ASDP Baseline_Groups on 3 and 1 AU, respectively. These results demonstrate that for 10 of 14 AUs, two CASP13 predictor groups—Laufer and Meilerlab—submitted first-ranked models for many targets that are more accurate than those generated using our conventional automated *ASDP* modeling protocol.

3.12 | Target N1008: real NMR data with backbone assignments only

Two real NMR data sets (N1008 and n1008) were provided for the data-guided prediction program of CASP13. Both data sets were for the CASP COMMONS target T1008 (foldit3), proposed by Brian Koepnick and David Baker as part of their project assessing de novo protein design by citizen scientists in the online protein-folding game Foldit.[26] Foldit players were provided a set of general principles for protein design in the form of Foldit rules, and the resulting designs were assessed by Rosetta stability calculations. One hundred and fifty-six designs were encoded in synthetic genes, which were expressed, screened for stability, and (in four cases) experimental structures were determined.[26] One of the protein designs, the 80-residue foldit3 protein, was produced for CASP13 with uniform ^{15}N,^{13}C-enrichment, and its structure was determined by conventional triple-resonance NMR in the context of this project. The structural statistics and global structure quality factors including Verify3D,[67] ProsaII,[68] PROCHECK,[69] and MolProbity[65] raw and statistical *Z*-scores were computed using the Protein Structure Validation Suite Software PSVS 1.5[66] and PDBStat[42] software packages. The global goodness-of-fit of the final structure ensembles with the NOESY peak list data, the NMR DP score, was determined using the *RPF* analysis program. The resulting reference 3D structure of foldit3, CASP target 1008, exhibits excellent convergence and structure quality statistics (Table S4). This structure and the associated data have been deposited in the Protein Data Bank (PDB id 6msp) and chemical shifts have been deposited in the BioMagResDataBase (BMRB id 30527).

Ambiguous contact lists for target 1008 were provided to CASP13 predictors as two distinct NMR-assisted targets. In the first cycle, NMR-assisted target N1008, backbone resonance assignments (only) were combined with complete ^{15}N- and ^{13}C edited NOESY spectra and Talos_N backbone dihedral restraints (derived from backbone chemical shift data) as input to the program *ASDP*. Following one cycle of analysis with *ASDP*, the structure-independent NOESY peak assignments (most of which are assigned to multiple possible ^{1}H-^{1}H interactions) were used to generate the ambiguous contact list for target N1008. This CASP13 target explores a novel approach to NMR structure determination, in that ^{1}H-^{1}H NOE interactions due to backbone-sidechain and sidechain-sidechain contacts are present in the NOESY peak list, but cannot be correctly assigned as the sidechain resonances are not present in the chemical shift list. For the second target data set, NMR-assisted target n1008, the nearly complete backbone and sidechain resonance assignments were combined with complete ^{15}N- and ^{13}C-edited NOESY data and Talos_N backbone dihedral restraints, as input to *ASDP*. Following one cycle of analysis with *ASDP*, the structure-independent NOESY peak assignments (most of which, again, are assigned to multiple possible ^{1}H-^{1}H interactions) were used to generate ambiguous contact list n1008. In this case, backbone-sidechain and sidechain-sidechain NOEs could generally be reliably assigned. Note that no RDC or EC data are available for target 1008. CASP13 predictor models provided using ambiguous contact lists N1008 and n1008 were all assessed against the final manually refined NMR structure.

Data set N1008 was designed to test the ability to combine backbone-only assignments with advanced structure prediction methods. Several studies have previously explored the combination of sparse NMR data obtained on perdeuterated protein samples with advanced molecular modeling methods.[7,8,20,70-76] For such data sets, the NOESY cross peak assignments are not complicated by the presence of unassigned sidechain resonances. However, in the case of N1008, the backbone resonances were assigned in a fully protonated ^{13}C,^{15}N-enriched protein sample; hence the NOEs may arise from backbone/backbone, backbone/sidechain, or sidechain/sidechain interactions. Since the sidechain resonances are not in the chemical shift list, many backbone/sidechain NOEs may be incorrectly assigned as unique backbone/backbone interactions; this is particularly problematic for NOEs involving resonances which are degenerate with assigned backbone H^{N} and H^{α} protons. The resulting falsely assigned backbone-backbone contacts might be expected to corrupt the structure. The goal of this experiment is to assess if data-guided predictions could overcome such corruption and provide an accurate 3D structure without sidechain assignments. In practice, the complicating effects of NOEs involving sidechain protons can be overcome using protein samples with perdeuterated H^{α} and sidechain resonances. However, here we explored the potential of avoiding perdeuteration, completing backbone assignments, and using such noisy NOESY data for accurate structure determination.

CASP13 results for target N1008 are illustrated in Figure 4. While these NOESY peak lists contain large numbers of NOE peaks which cannot be correctly assigned using the backbone chemical shift list, three prediction groups did very well with these data: Meilerlab 250 (GDT-TS 0.75), KIAS-Gdansk 208 (GDT-TS 0.73), and Laufer 431 (GDT-TS 0.68). The results of the ASDP baseline group 321 (GDT-TS 0.53) was significantly less accurate (note that Baseline group 459 did not contribute a distinct structure because no ECs are available for this de novo designed protein). For this N1008 data set, the first selected model of predictor groups were 15-22 GDT-TS points higher than the best models provided by the baseline methods. These results demonstrate the significant value of these modeling methods in obtaining accurate structures from sparse, noisy NOESY data.

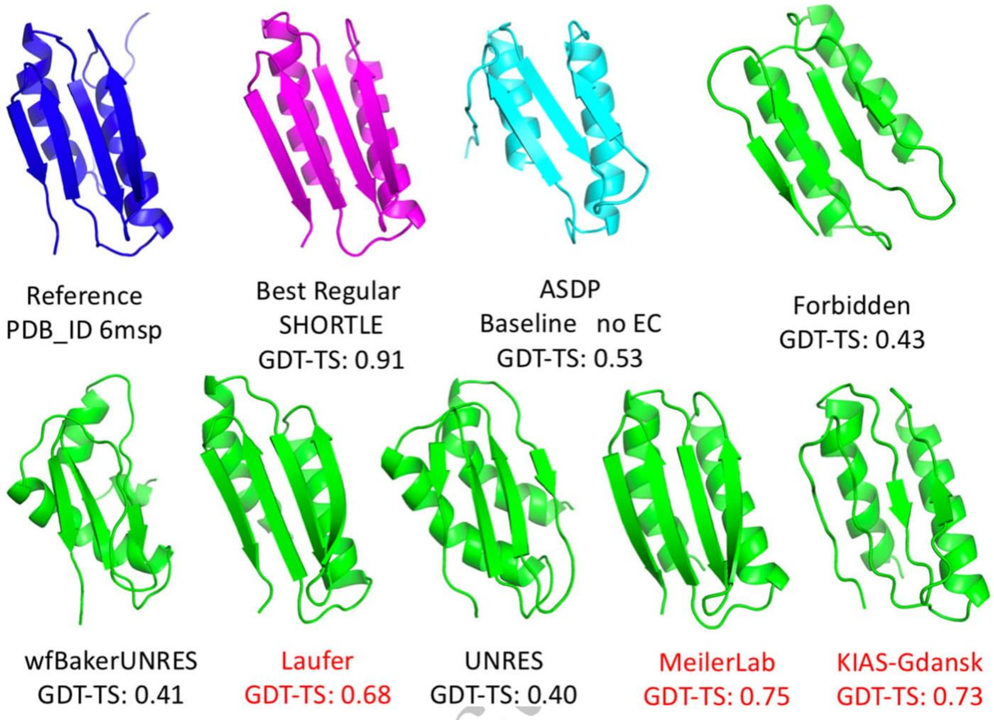


FIGURE 4 NMR-guided predictions for real NMR data target N1008. Target N1008 was produced with uniform ^{15}N-,^{13}C-enrichment (without perdeuteration). The backbone chemical shift assignments were used together with 3D ^{15}N-NOESY data to generate ambiguous contact list, which include many NOESY peaks involving sidechain resonances for which no assignment in list of possible assignments is correct. Blue, experimental solution NMR structure. Magenta, best nonassisted regular prediction. Lt blue, ASDP Baseline model. Green, NMR-assisted prediction results

Although target N1008 was a NMR-data assisted CASP target, some of the "regular predictions," which did not use any NMR data, were also quite good. In particular, SHORTLE 281 (GDT-TS 0.91) (Figure 4), A7D-DeepMind 043 (GDT-TS 0.81), and other regular predictor groups did remarkably well with this target, and significantly better than any group could do using these sparse NMR data. It should be noted that target 1008 is a de novo designed protein, and may be more amenable to accurate structure prediction compared to natural proteins. None the less, these results suggest hybrid methods in which models generated with regular methods are simply validated or refined against NMR data could be used for data sets like that provided for N1008.

3.13 | Target n1008: real NMR data with extensive backbone and sidechain assignments

Data set n1008 was provided as a control for performance with essentially complete NMR assignments. A reliable NMR-assisted prediction method should do well with these data. Because n1008 was among the very last data sets released for CASP13, only four predictor groups, plus Baseline_Group 321, ASDP without ECs, submitted models for n1008. For this data set, Baseline Group 321 provided the most accurate top ranked model, with GDT-TS 0.83. Laufer-431 also submitted a good first-ranked model (GDT-TS 0.57), followed by UNRES 288 (GDT-TS 0.41), Forbidden 122 (GDT-TS 0.40), and wf-Baker-UNRES 492 (GDT-TS 0.27). These results highlight the value of such control data in testing and developing NMR-assisted prediction methods, as the n1008 control data set is an important benchmark for testing various methods.

3.14 | Overall performance per target per group

Table 2 provides a summary of overall performance for the six predictor groups and two Baseline groups (321No EC or 459 Meta-PSI-COV EC), based on GDT-TS score of the top-ranked submission, per group, and per target. The table is color coded so that GDT-TS scores >0.50 (correct fold) are colored in shades of green, and scores <0.50 in shades of red. Certain targets (eg, N0981-D4) appear to be relatively

TABLE 2 GDT-TS scores for first ranked models for each target and each NMR-assisted predictor group

Target	Nres	Best regular prediction	492 wfBaker UNRES	431 Laufer	288 UNRES	250 Meilerlab	208 KIAS-Gdansk	122 Forbidden	ASDP baseline no EC	ASDP baseline with EC
N0957s1	162	45.2 A7D	31.9	52.9	28.8	56.0	NA	20.5	32.2	30.2
N0989	246	31.3 Zhang	12.7	23.5	13.9	17.5	NA	10.7	15.4	16.6
N0968s1	123	71.4 Elofsson	64.6	59.5	45.3	69.0	NA	31.5	59.7	54.6
N0968s2	115	78.7 A7D	60.0	73.7	55.4	43.2	NA	30.4	33.7	49.5
N0980s1	105	54.8 Multicom	29.8	67.7	25.0	59.8	NA	28.6	62.0	72.1
N0981-D1	86	66.2 slbio_server	49.4	58.4	53.7	55.5	61.0	NA	70.3	69.7
N0981-D2	80	34.0 Venclovas	NA	40.0	33.7	34.0	42.1	NA	64.3	67.5
N0981-D3	203	55.1 A7D	37.9	41.0	39.0	17.4	49.3	NA	55.7	55.1
N0981-D4	111	65.9 Multicom	50.6	65.7	47.7	61.7	59.6	NA	60.1	58.1
N0981-D5	127	72.8 A7D	53.1	76.5	39.7	59.8	40.9	NA	38.9	25.9
N1005	326	56.3 A7D	28.9	49.8	26.4	36.2	29.2	NA	33.9	29.4
N1008	97	91.2 SHORTLE	40.5	68.1	40.2	75.0	73.0	42.8	52.9	NA
n1008	97	91.2 SHORTLE	27.2	57.4	41.5	NA	NA	40.2	82.7	NA

easy, as most groups submitted good models with GDT-TS > 0.50 while other targets (eg, N0989, N0981-D2, N0981-D3, and N1005) were more difficult. Not surprisingly, the largest targets (>200 residues) were all among the most difficult targets. These results suggest that more efforts are needed even by the best NMR-assisted prediction methods, for addressing larger perdeuterated proteins where only sparse NMR data can be collected.

3.15 | Sidechain rotamer metrics

Another valuable structure quality metric involves comparing the sidechain conformations of buried residues in predictor models vs the reference structure.[49] Both $\chi 1$ and $\chi 2$ sidechain rotamer states for residues with buried side chains were compared between the first-ranked predicted models and the corresponding reference structure using the *PDBStat* program,[42] as described elsewhere.[49] Groups Laufer 431, Meilerab 250, and KIAS-Gdansk 208, well as ASDP Baseline_Groups 321 and 459, provided models with significantly better than average $\chi 1$ and $\chi 2$ sidechain rotamer agreement with reference structures, compared with the other predictor groups (Figure 5). However, all of the predictor groups have average $\chi 1$ and $\chi 2$ rotamer agreement of only 30%-60%, indicating that this is a valuable metric which should be focused on in future CASP experiments.

3.16 | RDC *Q*-scores for NMR-assisted prediction models

The RDC *Q*-score is a measure of the agreement between RDC values calculated from the model, and the RDC data. The *Q*-score ranges from 1 to 0, with lower values indicating better agreement between calculated and observed RDCs. The average ^{15}N-^{1}H RDC *Q*-scores for each of the six predictor groups ranged from 0.49 to 0.83 (Figure S3). These scores are significantly poorer than those of the baseline groups, which range from 0.19 to 0.21. Among the predictor groups, the best average RDC *Q*-scores were for models submitted by Meilerlab (average RDC *Q*-score 0.49) and Laufer (average RDC *Q*-score 0.63). The submitted models of the remaining four predictor groups have poor RDC *Q*-scores. These results demonstrate that all of the data-guided predictor groups could improve model accuracy by better consideration of RDC data in their prediction algorithms.

3.17 | DP scores for NMR-assisted prediction models

The NMR DP score[37,38] is a "NMR R factor", comparing the short ^{1}H-^{1}H distances in a protein structure model with all possible assignments of peaks in the NOESY peak list, considering the available

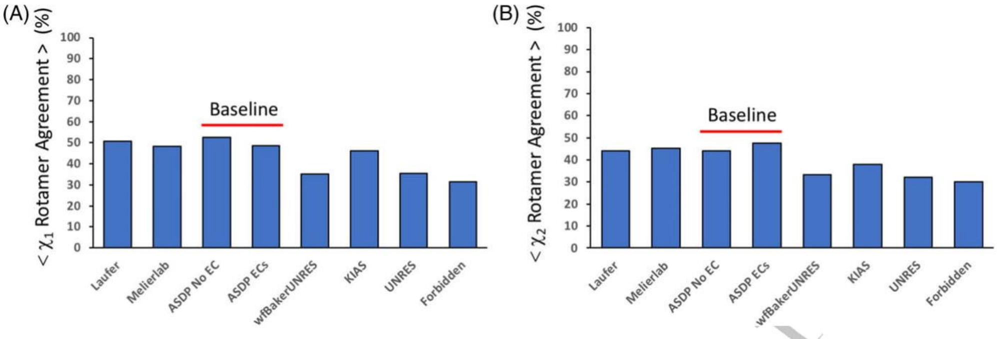


FIGURE 5 Sidechain χ_1 and χ_2 rotamer states. Percent of (A) χ_1 or (B) χ_2 rotamer states for buried residues in each assessment unit that match the corresponding dihedral angle of the reference structure in the best prediction models submitted by each prediction group

chemical shift data. The NMR DP score ranges from 0 to 1, and is correlated with structural accuracy. Correct structures generally have DP scores >0.6.[37,38,54] The DP scores for each of the six predictor groups ranged from 0.47 to 0.69 (Figure S4). These scores are generally lower than those of the baseline groups, which range from 0.73 to 0.75. However, ASDP uses the DP score to guide the automated NOESY peak analysis process. Among the predictor groups, the best average DP scores were for NMR-assisted models submitted by Laufer (average DP score 0.69) and Meilerlab (average DP score 0.63). DP scores for each reference X-ray crystal and NMR structure, and for the best regular and NMR-assisted models submitted for each target are summarized in Figure S5. These are generally consistent with the corresponding GDT-TS scores. Overall, the NMR-assisted predictor rankings based on DP scores are consistent with the rankings based on GDT-TS and other conventional CASP metrics, placing Laufer and Meilerlab as the best performing NMR-assisted prediction groups in CASP13.

4 | DISCUSSION

4.1 | Simulations of NOESY data for CASP targets

Simulated data provide an important tool for computational methods development. Although it is challenging to accurately simulate something as complex as a protein NOESY peak list, the powerful advantage of such simulated data is that the ground truth structure is known. Using real data has the advantage of including effects which are not captured in simulated data. For example, protein dynamics and signal overlap are primary causes of errors in the conversion of NMR observables into structural restraints, and may be difficult to account for in simulating NOESY data. However, with real data the "true" structural distribution from which these data arise is generally not known.

Normally, one cannot assign the frequencies of resonance of all nuclei in a protein. In practice, two scenarios may occur: either there are no observable peaks involving a given nucleus, because of local dynamics preventing their detection, or the resonance cannot be assigned to a unique atom with confidence, for example, due to accidental degeneracies. In the latter case, peaks are observable but cannot be converted into the appropriate structural restraints. For this work, we introduced both types of problems in our simulations, by manually selecting for each target regions of protein sequence from which resonance assignment were deleted. To simulate missing NOESY data for assigned resonances, removed some resonance assignments before simulating the NOESY peak list. To simulate missing assignments of resonances which do provide NOESY peaks, we removed resonance assignments for some residues after simulating the NOESY spectrum, while retaining the NOESY peaks. This second situation is actually common in sparse NMR data sets, and can lead to restraints that are incorrect. For example, mis-assignments of NOESY peaks due to missing resonance assignments are particularly extensive for the real NMR data set N1008, in which only backbone resonance assignments were provided while the NOESY peak list includes NOEs with unassigned sidechain resonances. Importantly, some of the modeling methods used in this data assisted CASP13 experiment were able to overcome the challenges of these real data provided for target N1008, to provide accurate structures without the need for sidechain resonance assignments or perdeuteration of the protein sample. This represents a novel approach to small protein structure determination by NMR.

In CASP11, one successful strategy used by the Baker lab was to focus their initial NMR-guided predictions on the uniquely assigned NOE-based contacts provided in the ambiguous contact lists (7). Table S5 provides an analysis of the unique long-range, and unique H^N-H^N long-range, contacts in the ambiguous contact lists provided to CASP13 predictors. These ranged from 0.8 to 1.6 uniquely assigned long-range contacts per residue, and from 0.24 to 0.58 H^N-H^N long-range contacts per residue, similar to the distributions provided in the CASP11 ambiguous contact lists, which were also based on simulated NOESY peak lists. For the real sparse NMR data set, target N1008, the ambiguous contact list has 0.54 and 0.28 unique long-range and unique H^N-H^N long-range contacts, respectively. These densities of long-range contacts in this real NMR data set are similar to, but at the lower-end, of the ranges provided in the simulated ambiguous contact lists.

4.2 | Impact of NMR data in improving regular predictions

An important question to be addressed in this NMR data guided prediction component of CASP13 is whether incorporation of sparse experimental data can improve the accuracy of prediction. To assess this, we compared the best "regular prediction" model with the best "NMR data assisted" model, where both the regular and assisted models were provided by the same predictor group; that is, for Forbidden, KIAS-Gdansk, UNRES, Laufer, and wf-Baker-UNRES. Meilerlab did not provide a "regular prediction," precluding this analysis. The CASP organizers recognize that predictor groups may have utilized different modeling methods for their regular and "NMR assisted" predictions, or even used models submitted by other predictor groups, which became available between the release of "regular" and "data assisted" targets, making these comparisons not as rigorous as we might like.

Our results show that, in most cases, incorporation of NMR data results in models with much higher accuracy predictions. NMR-assisted prediction models are, on average, more accurate than the corresponding regular prediction by the same group (Figure 6A). Modeling methods that used NMR data generally improved accuracy of prediction over modeling methods used by the same groups without NMR data. In some cases, the improvement for particular targets was as much as 40 GDT-TS points (Figure 6B). The improvement was particularly dramatic for groups Laufer (average improvement across all targets of 25 GDT-TS points; maximum improvement on a specific target of 42 GDT-TS points) and wf-Baker-UNRES (average improvement across all targets of five GDT-TS points; maximum improvement on a specific target of 39 GDT-TS points). Hence, we conclude that sparse, noisy NMR data can generally improve model prediction accuracy. In some cases, however, incorporating these simulated or real NMR data resulted in reduced model accuracy for some targets; for example, most of the predictor groups submitted first-ranked "data assisted" targets which are less accurate than their corresponding "regular" predictions (Figure 6B). For group KIAS-Gdansk 208, this results in an average reduction in GDT-TS score across all targets of—five GDT-TS points (Figure 6A). These results suggest that more efforts are needed by predictor in implementing sparse NMR data in their data-guided prediction algorithms.

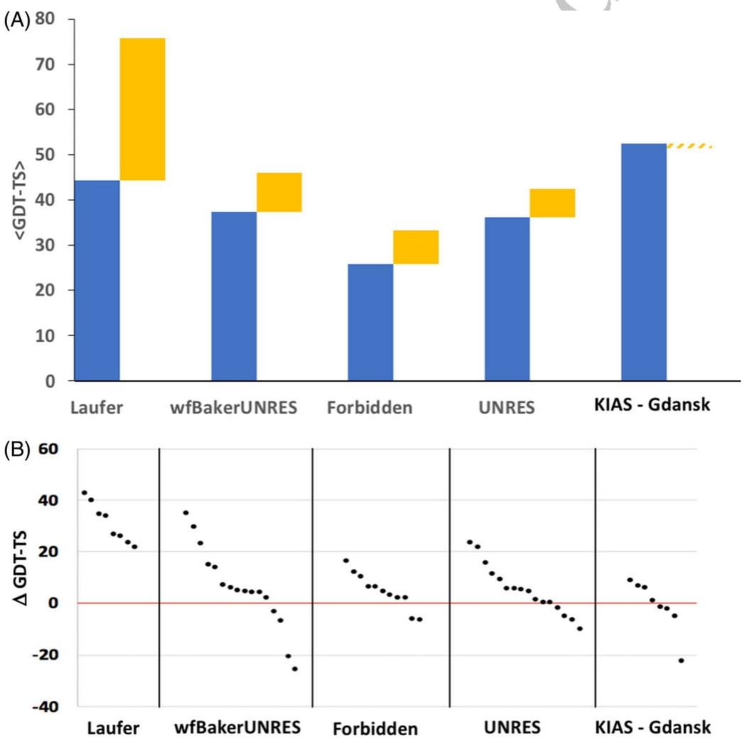


FIGURE 6 Impact of sparse NMR data on modeling accuracy: a NMR-assisted prediction group also submitted regular prediction models, and the corresponding best regular and best NMR-assisted predictions were compared. (A) Stacked column plot showing the average GDT-TS scores for best models submitted for each assessment unit as a regular prediction (blue) or NMR-assisted prediction (yellow). (B) Δ GDT-TS = GDT-TS$_{assisted}$ - GDT-TS$_{regular}$ for the best NMR-assisted and best regular model submitted for each target by each group providing both regular and assisted predictions

4.3 | NMR-assisted predictions of larger proteins

None of the NMR-assisted groups did particularly well with the three larger (>200 residue) targets (Table 2). For these targets, most NMR-assisted models have GDT-TS < 0.50. The most accurate predictions were those of the Laufer (two of three targets) and ASDP baseline (one of three targets) groups. This contrasts the results in CASP11, where two predictor groups, Lee and Baker, were particularly outstanding in modeling larger (>200 residues) proteins more accurately than baseline methods using sparse NMR data. Regrettably, neither these Lee or Baker groups of CASP11 participated in the NMR-assisted prediction component of CASP13.

4.4 | The best "regular" prediction for a target was often more accurate than the best "data assisted" prediction

A second key question we wanted to address involves comparing the accuracy of all regular prediction methods with NMR data assisted predictions. For six of the 13 targets, the best NMR-assisted models were more accurate than any regular prediction (solid yellow histogram bars in Figure 7A). This improvement using the NMR data was particularly dramatic for target N0981-D2 (Figure 8). For this target, the best NMR-assisted model (Baseline_Group 313 ASDP No EC, GDT-TS 0.68) is significantly more accurate than the best unassisted regular model (Venclovas 366; GDT-TS 0.35). Interestingly, all 15 of the top-ranked assisted models for target N0981-D2 were from the Baseline groups 321, 459, and 313. This may reflect the nature of this fold, since the *ASDP* program uses algorithms designed to address the unique features of beta sheets.[24] The best NMR-assisted models submitted by predictor groups Laufer (GDT-TS 0.46) and Meilerlab (GDT-TS 0.36) were also more accurate than the best model from regular prediction groups. These results confirm the expectation that, generally speaking, inclusion of sparse NMR data improves the accuracy of predictions. Detailed descriptions of the methods used by Laufer[77] and Meilerlab[78] are presented in their own papers on NMR-assisted prediction in CASP13.

Although NMR-assisted modeling provide the best models generated by any methods for several targets, for seven out of 13 targets used in the NMR-data assisted component of CASP13, the most accurate (best) model provided by a regular prediction group was actually more accurate than the most accurate model provided by any NMR-data assisted prediction group (hashed yellow histogram bars in Figure 7A). This is also evident by plotting the GDT-TS score for best model submitted by the NMR-assisted groups against the GDT-TS score for the best model of the corresponding target by any regular prediction group (Figure 7B); many of these comparisons fall below the diagonal indicating that at least one regular prediction group provided a more accurate model than the corresponding NMR-assisted model. Although in these cases, the improved accuracy of the non-assisted group is only marginal, they are non-the-less impressive because no sample-specific experimental data was used. The regular prediction groups providing these highly accurate "regular predictions" include groups A7D 043 (Deep Mind), Zhang 322, Venclovas 366, slbio_serve 266s, SHORTLE 281, and MULTICOM 083. Several

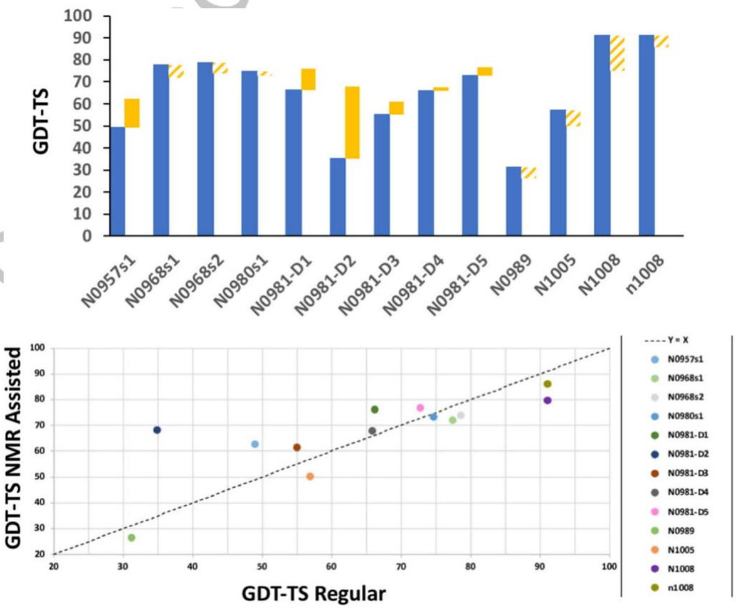


FIGURE 7 Best NMR assisted vs best regular predictions. (A) GDT-TS scores for the best model submitted by any NMR-assisted prediction group (blue bars) compared with the best model submitted by any regular prediction group (yellow bars). The solid yellow bars indicate targets for which the best model submitted by any regular prediction was more accurate than the best model submitted by any NMR-assisted prediction group. The hashed-yellow bars indicate targets for which the best model submitted by any regular prediction was less accurate than the best model submitted by any NMR-assisted prediction group. (B) Scatter plot comparing the GDT-TS scores for the best model submitted by any regular prediction vs the best model submitted by any NMR-assisted prediction

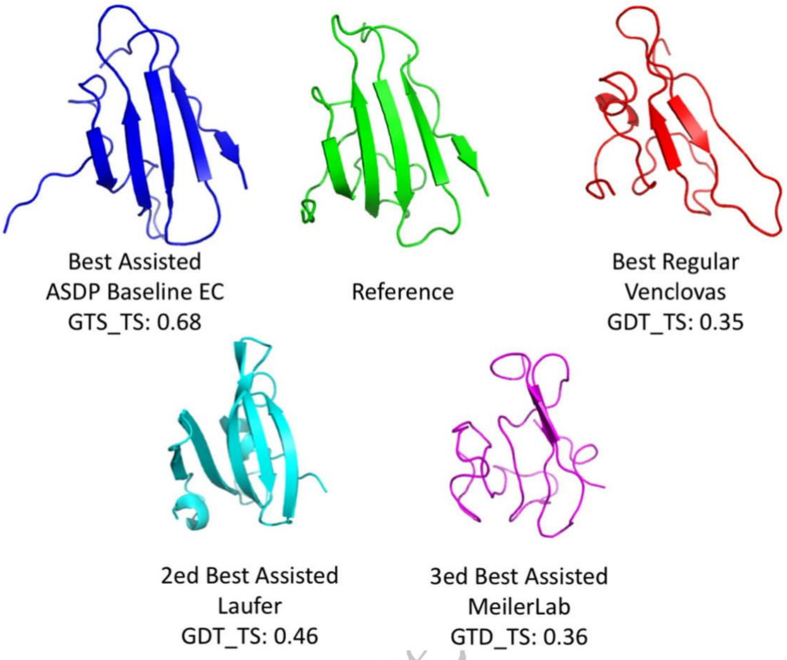


FIGURE 8 Comparison of best assisted and best regular prediction models submitted for target N0981-D2. The best NMR-assisted models of the baseline ASDP (with EC) and Laufer groups are more accurate than the best models provided by any regular prediction group (Venclovas)

of these regular prediction groups utilize novel machine learning methods to guide structural modeling. For the NMR-guided targets, group A7D 043 (Deep Mind) provided the most accurate models for six of 13 targets, and 10 of 16 assessment units; they also contributed 27 of 48 (ie, 3 × 16 = 48) top three most accurate models. These remarkable results suggest a novel approach for structure determination using sparse NMR data, in which pure prediction methods, like the machine leaning methods being developed by Deep Mind, Zhang, MULTICOM, and other groups, are first used to generate structural models, and the sparse NMR data is then used validate and/or refine these models.

ACKNOWLEDGMENTS

This work was supported by NIH NIGMS grants 1R01GM120574 (to G.T.M.), R01GM100482 (to K.F.), and P20 RR-016461 (to H.V.).

CONFLICT OF INTEREST

G. L. is an officer of Nexomics Biosciences, Inc. M. I. is a scientific advisor of Nexomics Biosciences, Inc. G. T. M. is a founder of Nexomics Biosciences, Inc.

ORCID

David Snyder https://orcid.org/0000-0001-6608-2975
Krzysztof Fidelis https://orcid.org/0000-0002-8061-412X
Andriy Kryshtafovych https://orcid.org/0000-0001-5066-7178
Gaetano T. Montelione https://orcid.org/0000-0002-9440-3059

REFERENCES

1. Lesk AM. CASP2: report on ab initio predictions. *Proteins*. 1997;suppl 1:151-166.
2. Monastyrskyy B, Fidelis K, Tramontano A, Kryshtafovych A. Evaluation of residue-residue contact predictions in CASP9. *Proteins*. 2011; 79(suppl 10):119-125.
3. Monastyrskyy B, D'Andrea D, Fidelis K, Tramontano A, Kryshtafovych A. Evaluation of residue-residue contact prediction in CASP10. *Proteins*. 2014;82(suppl 2):138-153.
4. Monastyrskyy B, D'Andrea D, Fidelis K, Tramontano A, Kryshtafovych A. New encouraging developments in contact prediction: assessment of the CASP11 results. *Proteins*. 2016;84(suppl 1): 131-144.
5. Schaarschmidt J, Monastyrskyy B, Kryshtafovych A, Bonvin A. Assessment of contact predictions in CASP12: co-evolution and deep learning coming of age. *Proteins*. 2018;86(suppl 1):51-66.
6. Kinch LN, Li W, Monastyrskyy B, Kryshtafovych A, Grishin NV. Assessment of CASP11 contact-assisted predictions. *Proteins*. 2016; 84(suppl 1):164-180.

7. Ovchinnikov S, Park H, Kim DE, Liu Y, Wang RY, Baker D. Structure prediction using sparse simulated NOE restraints with Rosetta in CASP11. *Proteins*. 2016;84(suppl 1):181-188.
8. Joo K, Joung I, Cheng Q, Lee SJ, Lee J. Contact-assisted protein structure modeling by global optimization in CASP11. *Proteins*. 2016;84 (suppl 1):189-199.
9. Morcos F, Pagnani A, Lunt B, et al. Direct-coupling analysis of residue coevolution captures native contacts across many protein families. *Proc Natl Acad Sci USA*. 2011;108:E1293-E1301.
10. Marks DS, Colwell LJ, Sheridan R, et al. Protein 3D structure computed from evolutionary sequence variation. *PLoS One*. 2011;6: e28766.
11. Marks DS, Hopf TA, Sander C. Protein structure prediction from sequence variation. *Nat Biotechnol*. 2012;30:1072-1080.
12. Ovchinnikov S, Kinch L, Park H, et al. Large-scale determination of previously unsolved protein structures using evolutionary information. *Elife*. 2015;4:e09248.
13. Ovchinnikov S, Park H, Varghese N, et al. Protein structure determination using metagenome sequence data. *Science*. 2017;355: 294-298.
14. Buchan DWA, Jones DT. Improved protein contact predictions with the MetaPSICOV2 server in CASP12. *Proteins*. 2018;86(suppl 1): 78-83.
15. Ogorzalek TL, Hura GL, Belsom A, et al. Small angle X-ray scattering and cross-linking for data assisted protein structure prediction in CASP 12 with prospects for improved accuracy. *Proteins*. 2018;86 (suppl 1):202-214.
16. Rosato A, Aramini JM, Arrowsmith C, et al. Blind testing of routine, fully automated determination of protein structures from NMR data. *Structure*. 2012;20:227-236.
17. Rosato A, Bagaria A, Baker D, et al. CASD-NMR: critical assessment of automated structure determination by NMR. *Nat Methods*. 2009;6: 625-626.
18. Rosato A, Vranken W, Fogh RH, et al. The second round of critical assessment of automated structure determination of proteins by NMR: CASD-NMR-2013. *J Biomol NMR*. 2015;62:413-424.
19. Ragan TJ, Fogh RH, Tejero R, et al. Analysis of the structural quality of the CASD-NMR 2013 entries. *J Biomol NMR*. 2015;62:527-540.
20. Lange OF, Rossi P, Sgourakis NG, et al. Determination of solution structures of proteins up to 40 kDa using CS-Rosetta with sparse NMR data from deuterated samples. *Proc Natl Acad Sci USA*. 2012; 109:10873-10878.
21. Rossi P, Monneau YR, Xia Y, Ishida Y, Kalodimos CG. Toolkit for NMR studies of methyl-labeled proteins. *Methods Enzymol*. 2019;614: 107-142.
22. Rosen MK, Gardner KH, Willis RC, Parris WE, Pawson T, Kay LE. Selective methyl group protonation of perdeuterated proteins. *J Mol Biol*. 1996;263:627-636.
23. Gardner KH, Rosen MK, Kay LE. Global folds of highly deuterated, methyl-protonated proteins by multidimensional NMR. *Biochemistry*. 1997;36:1389-1401.
24. Huang YJ, Tejero R, Powers R, Montelione GT. A topology-constrained distance network algorithm for protein structure determination from NOESY data. *Proteins*. 2006;62:587-603.
25. Huang YJ, Mao B, Xu F, Montelione GT. Guiding automated NMR structure determination using a global optimization metric, the NMR DP score. *J Biomol NMR*. 2015;62:439-451.
26. Koepnick B, Flatten J, Husain T, et al. De novo protein design by citizen scientists. *Nature*. 2019;570:390-394.
27. Suzuki M, Zhang J, Liu M, Woychik NA, Inouye M. Single protein production in living cells facilitated by an mRNA interferase. *Mol Cell*. 2005;18:253-261.
28. Schneider WM, Inouye M, Montelione GT, Roth MJ. Independently inducible system of gene expression for condensed single protein production (cSPP) suitable for high efficiency isotope enrichment. *J Struct Funct Genomics*. 2009;10:219-225.
29. Acton TB, Xiao R, Anderson S, et al. Preparation of protein samples for NMR structure, function, and small-molecule screening studies. *Methods Enzymol*. 2011;493:21-60.
30. Delaglio F, Grzesiek S, Vuister GW, Zhu G, Pfeifer J, Bax A. NMRPipe: a multidimensional spectral processing system based on UNIX pipes. *J Biomol NMR*. 1995;6:277-293.
31. Goddard TD, Kneller DG. *Sparky 3*. San Francisco, CA: University of California; 2000.
32. Bartels C, Xia TH, Billeter M, Guntert P, Wuthrich K. The program XEASY for computer-supported NMR spectral analysis of biological macromolecules. *J Biomol NMR*. 1995;6:1-10.
33. Zimmerman DE, Kulikowski CA, Huang Y, et al. Automated analysis of protein NMR assignments using methods from artificial intelligence. *J Mol Biol*. 1997;269:592-610.
34. Moseley HN, Monleon D, Montelione GT. Automatic determination of protein backbone resonance assignments from triple resonance nuclear magnetic resonance data. *Methods Enzymol*. 2001;339: 91-108.
35. Güntert P, Mumenthaler C, Wüthrich K. Torsion angle dynamics for NMR structure calculation with the new program DYANA. *J Mol Biol*. 1997;273:283-298.
36. Guntert P. Automated NMR structure calculation with CYANA. *Methods Mol Biol*. 2004;278:353-378.
37. Huang YJ, Powers R, Montelione GT. Protein NMR recall, precision, and F-measure scores (RPF scores): structure quality assessment measures based on information retrieval statistics. *J Am Chem Soc*. 2005;127:1665-1674.
38. Huang YJ, Rosato A, Singh G, Montelione GT. RPF: a quality assessment tool for protein NMR structures. *Nucleic Acids Res*. 2012;40: W542-W546.
39. Linge JP, Williams MA, Spronk CA, Bonvin AM, Nilges M. Refinement of protein structures in explicit solvent. *Proteins*. 2003;50:496-506.
40. Brunger AT, Adams PD, Clore GM, et al. Crystallography & NMR system: a new software suite for macromolecular structure determination. *Acta Crystallogr*. 1998;54:905-921.
41. Bhattacharya A, Tejero R, Montelione GT. Evaluating protein structures determined by structural genomics consortia. *Proteins*. 2007;66: 778-795.
42. Tejero R, Snyder D, Mao B, Aramini JM, Montelione GT. PDBStat: a universal restraint converter and restraint analysis software package for protein NMR. *J Biomol NMR*. 2013;56:337-351.
43. Cozzetto D, Kryshtafovych A, Fidelis K, Moult J, Rost B, Tramontano A. Evaluation of template-based models in CASP8 with standard measures. *Proteins*. 2009;77(suppl 9):18-28.
44. Olechnovic K, Monastyrskyy B, Kryshtafovych A, Venclovas C. Comparative analysis of methods for evaluation of protein models against native structures. *Bioinformatics*. 2019;35:937-944.
45. Kryshtafovych A, Monastyrskyy B, Fidelis K. CASP prediction center infrastructure and evaluation measures in CASP10 and CASP ROLL. *Proteins*. 2014;82(suppl 2):7-13.
46. Kryshtafovych A, Monastyrskyy B, Fidelis K. CASP11 statistics and the prediction center evaluation system. *Proteins*. 2016;84(suppl 1): 15-19.
47. Clore GM, Schwieters CD. How much backbone motion in ubiquitin is required to account for dipolar coupling data measured in multiple alignment media as assessed by independent cross-validation? *J Am Chem Soc*. 2004;126:2923-2938.
48. Valafar H, Prestegard JH. REDCAT: a residual dipolar coupling analysis tool. *J Magn Reson*. 2004;167:228-241.
49. Tang Y, Huang YJ, Hopf TA, Sander C, Marks DS, Montelione GT. Protein structure determination by combining sparse NMR data with evolutionary couplings. *Nat Methods*. 2015;12:751-754.

50. Koradi R, Billeter M, Wuthrich K. MOLMOL: a program for display and analysis of macromolecular structures. *J Mol Graph*. 1996;14:51-55. 29-32.
51. Struyf A, Hubert M, Peter R. Clustering in an object-oriented environment. *J Stat Softw*. 1997;1:1-30.
52. Davis IW, Leaver-Fay A, Chen VB, et al. MolProbity: all-atom contacts and structure validation for proteins and nucleic acids. *Nucleic Acids Res*. 2007;35:W375-W383.
53. Han B, Liu Y, Ginzinger SW, Wishart DS. SHIFTX2: significantly improved protein chemical shift prediction. *J Biomol NMR*. 2011;50:43-57.
54. Rosato A, Tejero R, Montelione GT. Quality assessment of protein NMR structures. *Curr Opin Struct Biol*. 2013;23:715-724.
55. Shen Y, Bax A. Protein structural information derived from NMR chemical shift with the neural network program TALOS-N. *Methods Mol Biol*. 2015;1260:17-32.
56. Valafar H, Prestegard JH. REDCAT: a residual dipolar coupling analysis tool. *J Magn Reson (San Diego Calif)*. 2004;1997(167):228-241.
57. Schmidt C, Irausquin SJ, Valafar H. Advances in the REDCAT software package. *BMC Bioinformatics*. 2013;14:302.
58. Zweckstetter M. NMR: prediction of molecular alignment from structure using the PALES software. *Nat Protoc*. 2008;3:679-690.
59. Mao B, Tejero R, Baker D, Montelione GT. Protein NMR structures refined with Rosetta have higher accuracy relative to corresponding X-ray crystal structures. *J Am Chem Soc*. 2014;136:1893-1906.
60. Sheridan R, Fieldhouse RJ, Hayat S, Sun Y, Antipin Y, Yang L, Hopf T, Marks DS, Sander C. EVfold.org: evolutionary couplings and protein 3D structure prediction. bioRxiv, 2015; bioRxiv 021022.
61. Eddy SR. Accelerated profile HMM searches. *PLoS Comput Biol*. 2011; 7:e1002195.
62. Mariani V, Kiefer F, Schmidt T, Haas J, Schwede T. Assessment of template based protein structure predictions in CASP9. *Proteins*. 2011;79(suppl 10):37-58.
63. Huang YJ, Mao B, Aramini JM, Montelione GT. Assessment of template-based protein structure predictions in CASP10. *Proteins*. 2014;82(suppl 2):43-56.
64. Zemla A. LGA: a method for finding 3D similarities in protein structures. *Nucleic Acids Res*. 2003;31:3370-3374.
65. Chen VB, Arendall WB 3rd, Headd JJ, et al. MolProbity: all-atom structure validation for macromolecular crystallography. *Acta Crystallogr D Biol Crystallogr*. 2010;66:12-21.
66. Bhattacharya A, Wunderlich Z, Monleon D, Tejero R, Montelione GT. Assessing model accuracy using the homology modeling automatically software. *Proteins*. 2008;70:105-118.
67. Luthy R, Bowie JU, Eisenberg D. Assessment of protein models with three-dimensional profiles. *Nature*. 1992;356:83-85.
68. Sippl MJ. Recognition of errors in three-dimensional structures of proteins. *Proteins*. 1993;17:355-362.
69. Laskowski RA, MacArthur MW, Moss DS, Thornton JM. PROCHECK - a program to check the stereochemical quality of protein structures. *J Appl Cryst*. 1993;26:283-291.
70. Zheng D, Huang YJ, Moseley HN, et al. Automated protein fold determination using a minimal NMR constraint strategy. *Protein Sci*. 2003; 12:1232-1246.
71. Ramelot TA, Raman S, Kuzin AP, et al. Improving NMR protein structure quality by Rosetta refinement: a molecular replacement study. *Proteins*. 2009;75:147-167.
72. Raman S, Lange OF, Rossi P, et al. NMR structure determination for larger proteins using backbone-only data. *Science*. 2010;327:1014-1018.
73. Li W, Zhang Y, Kihara D, et al. TOUCHSTONEX: protein structure prediction with sparse NMR data. *Proteins*. 2003;53:290-306.
74. Shealy P, Simin M, Park SH, Opella SJ, Valafar H. Simultaneous structure and dynamics of a membrane protein using REDCRAFT: membrane-bound form of Pf1 coat protein. *J Magn Reson*. 1997;2010 (207):8-16.
75. Cole CA, Ishimaru D, Hennig M, Valafar H. An investigation of minimum data requirement for successful structure determination of Pf2048.1 with REDCRAFT. Proceedings of the International Conference on Bioinformatics & Computational Biology (BIOCOMP): The Steering Committee of the World Congress in Computer Science, Computer Engineering and Applied Computing (WorldComp); 2015: 17–24.
76. Timko E, Shealy P, Bryson M, Valafar H. Minimum data requirements and supplemental angle constraints for protein structure prediction with REDCRAFT. BIOCOMP, 2008: 738–744.
77. Robertson JC, Nassar R, Liu C, Brini E, Dill KA, Perez A. NMR-assisted protein structure prediction with MELDxMD. *Proteins*. 2019.
78. Kuenze G, Meiler J. Protein structure prediction using sparse NOE and RDC restraints with Rosetta in CASP13. *Proteins*. 2019.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of this article.

How to cite this article: Sala D, Huang YJ, Cole CA, et al. Protein structure prediction assisted with sparse NMR data in CASP13. *Proteins*. 2019;1–18. https://doi.org/10.1002/prot.25837

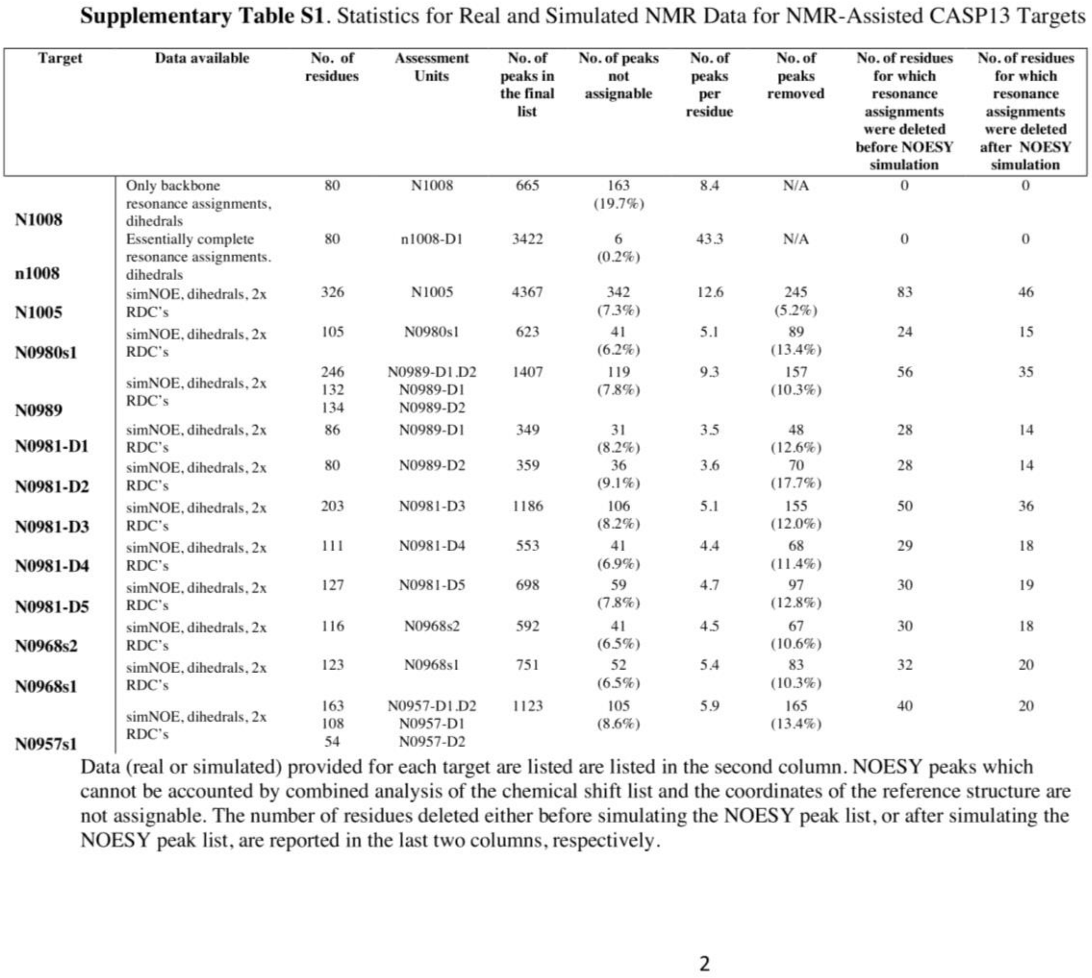
Supplementary Material for

Protein Structure Prediction Assisted with Sparse NMR Data in CASP13

Davide Sala#, Yuanpeng Janet Huang#, Casey A. Cole, David Snyder, Gaohua Liu, Yojiro Ishida, G.V.T. Swapna, Kelly P. Brock, Chris Sander, Krzysztof Fidelis, Andriy Kryshtafovych, Masayori Inouye, Roberto Tejero, Homayoun Valafar, Antonio Rosato*, Gaetano T. Montelione*
# co-first author
* co-submitting authors

Production and Solution NMR Structure Determination of Target 1008 (foldit3).

The synthetic gene for foldit3 [26] without ACA sequences [27, 28] was obtained from Genscript already incorporated into plasmid pET15TEV_NESG, which includes a N-terminal 6xHis purification tag, followed by a TEV protease cleavage site (sequence 'MGHHHHHHGWSENLYFQGS'). For these NMR studies, this affinity purification tag was not removed. Sample preparation followed standard protocols, as outlined in the previous publication on foldit3 [26]. *E. coli* BL21(DE3) cells harboring plasmid pET15TEV_NESG-foldit3 were grown in 1 L MJ9 minimal media [78], supplemented with 100 μg/ml ampicillin at 37 ºC. In order to produce uniformly ^{15}N and ^{13}C enriched protein samples, 1g / L ^{15}NH$_4$-salts and 2g / L U-^{13}C glucose were added as sole a nitrogen and a carbon sources, respectively. When O.D.$_{600}$ reached around 0.5 units, the culture was transferred to 18 ºC, and the protein production was induced by addition of 1 mM IPTG. After overnight incubation, the cells were collected and resuspended in 20 ml binding buffer, containing 20 mM Tris-HCl pH 8.0, 500 mM NaCl and 20 mM imidazole. After passing the cells through a 16,000-17,000 psi French press twice, cell debris were removed by 10,000 rpm for 30 min. The supernatant was further spun down at 40,000 rpm for 1 hr. The obtained supernatant (soluble fraction) was mixed with 1 mL of Ni-resin and incubated at 4 ºC for 1 hr. The non-specific binding proteins were removed by 20 mL binding buffer and washing buffer, containing 20 mM Tris-HCl pH 8.0, 500 mM NaCl and 50 mM imidazole, and the target protein was eluted by 5 mL elution buffer (20 mM Tris-HCl pH 8.0, 500 mM NaCl and 300 mM imidazole). The protein was dialyzed against gel filtration buffer, containing 20 mM Tris-HCl pH 8.0, 100 mM NaCl), overnight, and gel filtration was carried out using AKTA Express purification system with high-load 26/600 Superdex 200 pg column. Homogeneity (> 97%) was validated by SDS polyacrylamide gel electrophoresis. The purified protein was dialyzed against 20 mM potassium phosphate (pH 6.5), and the protein concentration was adjusted to between 0.3-0.4 mM for NMR studies.



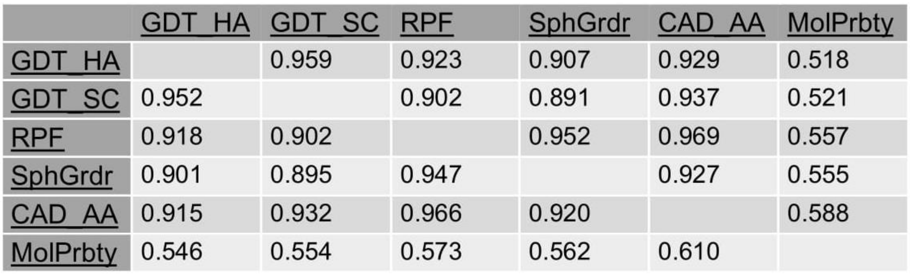
Supplementary Table S1. Statistics for Real and Simulated NMR Data for NMR-Assisted CASP13 Targets

Target	Data available	No. of residues	Assessment Units	No. of peaks in the final list	No. of peaks not assignable	No. of peaks per residue	No. of peaks removed	No. of residues for which resonance assignments were deleted before NOESY simulation	No. of residues for which resonance assignments were deleted after NOESY simulation
N1008	Only backbone resonance assignments, dihedrals	80	N1008	665	163 (19.7%)	8.4	N/A	0	0
n1008	Essentially complete resonance assignments. dihedrals	80	n1008-D1	3422	6 (0.2%)	43.3	N/A	0	0
N1005	simNOE, dihedrals, 2x RDC's	326	N1005	4367	342 (7.3%)	12.6	245 (5.2%)	83	46
N0980s1	simNOE, dihedrals, 2x RDC's	105	N0980s1	623	41 (6.2%)	5.1	89 (13.4%)	24	15
N0989	simNOE, dihedrals, 2x RDC's	246 132 134	N0989-D1.D2 N0989-D1 N0989-D2	1407	119 (7.8%)	9.3	157 (10.3%)	56	35
N0981-D1	simNOE, dihedrals, 2x RDC's	86	N0989-D1	349	31 (8.2%)	3.5	48 (12.6%)	28	14
N0981-D2	simNOE, dihedrals, 2x RDC's	80	N0989-D2	359	36 (9.1%)	3.6	70 (17.7%)	28	14
N0981-D3	simNOE, dihedrals, 2x RDC's	203	N0981-D3	1186	106 (8.2%)	5.1	155 (12.0%)	50	36
N0981-D4	simNOE, dihedrals, 2x RDC's	111	N0981-D4	553	41 (6.9%)	4.4	68 (11.4%)	29	18
N0981-D5	simNOE, dihedrals, 2x RDC's	127	N0981-D5	698	59 (7.8%)	4.7	97 (12.8%)	30	19
N0968s2	simNOE, dihedrals, 2x RDC's	116	N0968s2	592	41 (6.5%)	4.5	67 (10.6%)	30	18
N0968s1	simNOE, dihedrals, 2x RDC's	123	N0968s1	751	52 (6.5%)	5.4	83 (10.3%)	32	20
N0957s1	simNOE, dihedrals, 2x RDC's	163 108 54	N0957-D1.D2 N0957-D1 N0957-D2	1123	105 (8.6%)	5.9	165 (13.4%)	40	20

Data (real or simulated) provided for each target are listed are listed in the second column. NOESY peaks which cannot be accounted by combined analysis of the chemical shift list and the coordinates of the reference structure are not assignable. The number of residues deleted either before simulating the NOESY peak list, or after simulating the NOESY peak list, are reported in the last two columns, respectively.

2

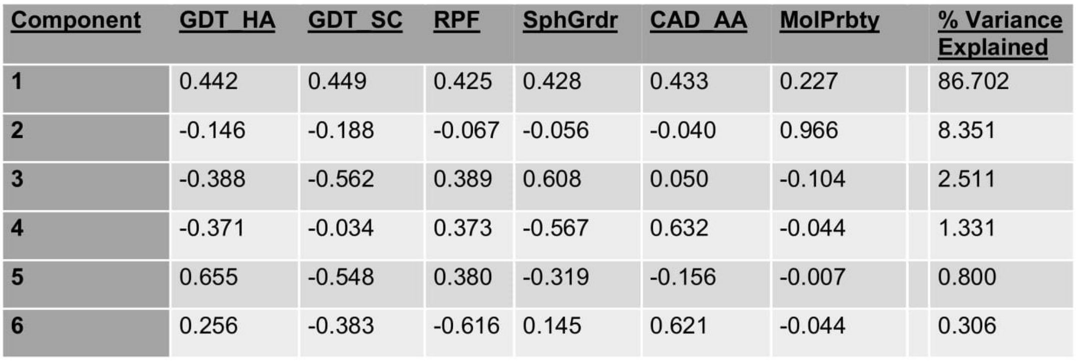
Supplementary Table S2. Correlation Coefficients Between Various CASP13 Metrics.



	GDT_HA	GDT_SC	RPF	SphGrdr	CAD_AA	MolPrbty
GDT_HA		0.959	0.923	0.907	0.929	0.518
GDT_SC	0.952		0.902	0.891	0.937	0.521
RPF	0.918	0.902		0.952	0.969	0.557
SphGrdr	0.901	0.895	0.947		0.927	0.555
CAD_AA	0.915	0.932	0.966	0.920		0.588
MolPrbty	0.546	0.554	0.573	0.562	0.610	

Friedman’s Test indicates, aside from the MolProbity packing metric, different scoring techniques do not give significantly different rankings. Upper right – Pearson coefficient.
Lower left – Spearman coefficient.

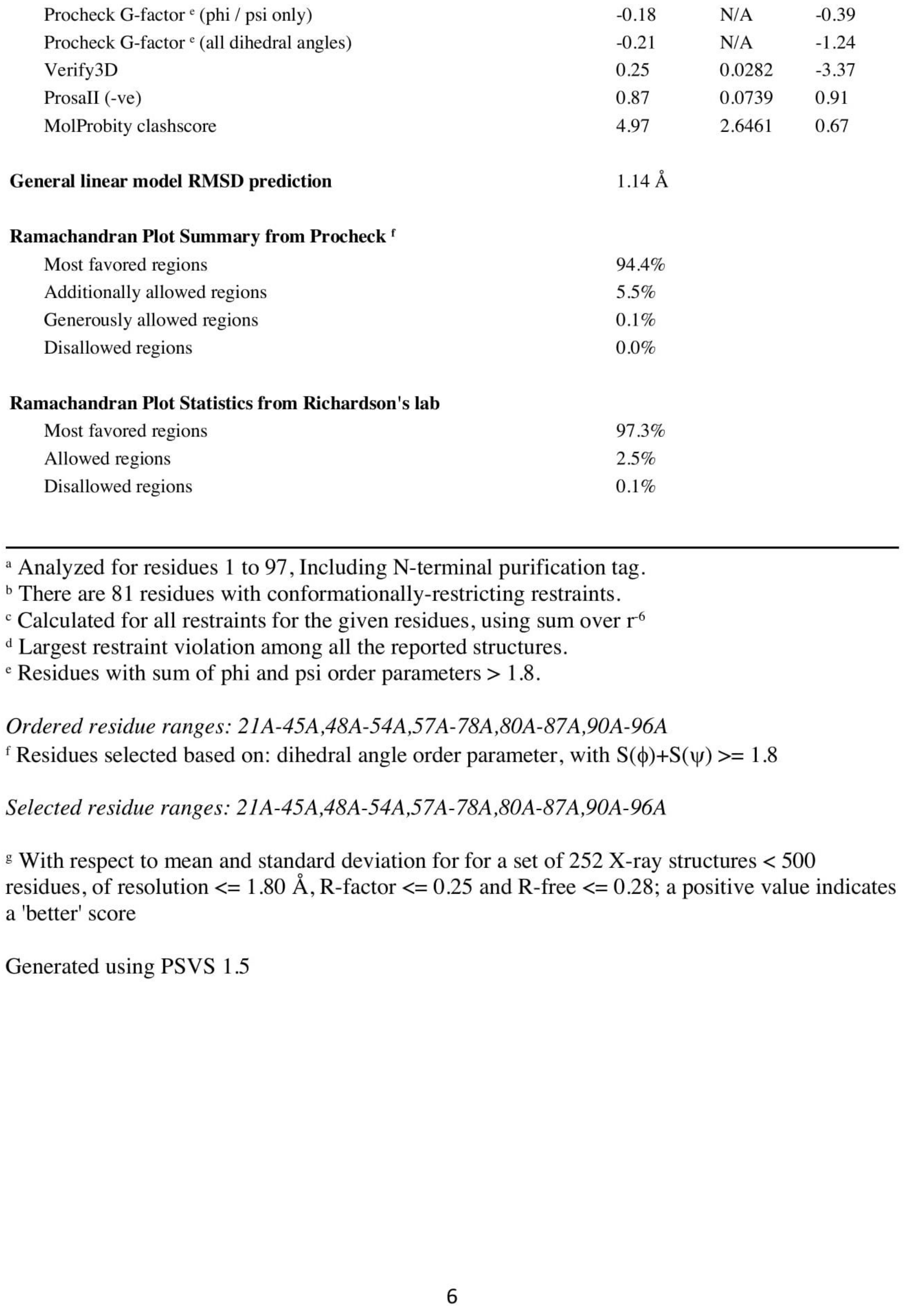
Supplementary Table S3. Principal Component Analysis of Key Structure Assessment Metrics



Component	GDT_HA	GDT_SC	RPF	SphGrdr	CAD_AA	MolPrbty		% Variance Explained
1	0.442	0.449	0.425	0.428	0.433	0.227		86.702
2	-0.146	-0.188	-0.067	-0.056	-0.040	0.966		8.351
3	-0.388	-0.562	0.389	0.608	0.050	-0.104		2.511
4	-0.371	-0.034	0.373	-0.567	0.632	-0.044		1.331
5	0.655	-0.548	0.380	-0.319	-0.156	-0.007		0.800
6	0.256	-0.383	-0.616	0.145	0.621	-0.044		0.306

Supplemental Table S4: NMR Data and Refinement Statistics for Foldit3 [26]

Summary of conformationally-restricting experimental restraints [a]				
NOE-based distance restraints:				
Total		1725		
intra-residue [i = j]		448		
sequential [| i - j | = 1]		441		
medium range [1 < | i - j | < 5]		344		
long range [| i - j | ≥ 5]		492		
NOE restraints per restrained residue [b]		21.3		
Hydrogen bond restraints:				
Total		66		
long range [| i - j | ≥ 5]		22		
Dihedral-angle restraints:		118		
Total number of restricting restraints [b]		1909		
Total number of restricting restraints per restrained residue [b]		23.6		
Restricting long-range restraints per restrained residue [b]		6.3		
Total structures computed		100		
Number of structures used		20		
Residual constraint violations [a,c]				
Distance violations / structure				
0.1 - 0.2 Å		10.05		
0.2 - 0.5 Å		2.35		
> 0.5 Å		0		
RMS of distance violation / restraint		0.01 Å		
Maximum distance violation [d]		0.42 Å		
Dihedral angle violations / structure				
1 - 10 °		17		
> 10 °		0		
RMS of dihedral angle violation / restraint		1.12 °		
Maximum dihedral angle violation [d]		8.40 °		
RPF scores				
	Recall	Precision	F-measure	DP-score
	0.945	0.956	0.95	0.842
RMSD Values				
		all	ordered[e]	Selected[f]
All backbone atoms		6.9 Å	0.6 Å	0.6 Å
All heavy atoms		7.7 Å	1.1 Å	1.1 Å
Structure Quality Factors				
		Mean score	SD	Z-score [g]



Procheck G-factor [e] (phi / psi only)	-0.18	N/A	-0.39
Procheck G-factor [e] (all dihedral angles)	-0.21	N/A	-1.24
Verify3D	0.25	0.0282	-3.37
ProsaII (-ve)	0.87	0.0739	0.91
MolProbity clashscore	4.97	2.6461	0.67
General linear model RMSD prediction	1.14 Å		
Ramachandran Plot Summary from Procheck [f]			
Most favored regions	94.4%		
Additionally allowed regions	5.5%		
Generously allowed regions	0.1%		
Disallowed regions	0.0%		
Ramachandran Plot Statistics from Richardson's lab			
Most favored regions	97.3%		
Allowed regions	2.5%		
Disallowed regions	0.1%		

[a] Analyzed for residues 1 to 97, Including N-terminal purification tag.

[b] There are 81 residues with conformationally-restricting restraints.

[c] Calculated for all restraints for the given residues, using sum over r^{-6}

[d] Largest restraint violation among all the reported structures.

[e] Residues with sum of phi and psi order parameters > 1.8.

Ordered residue ranges: 21A-45A,48A-54A,57A-78A,80A-87A,90A-96A

[f] Residues selected based on: dihedral angle order parameter, with $S(\phi)+S(\psi) >= 1.8$

Selected residue ranges: 21A-45A,48A-54A,57A-78A,80A-87A,90A-96A

[g] With respect to mean and standard deviation for for a set of 252 X-ray structures < 500 residues, of resolution <= 1.80 Å, R-factor <= 0.25 and R-free <= 0.28; a positive value indicates a 'better' score

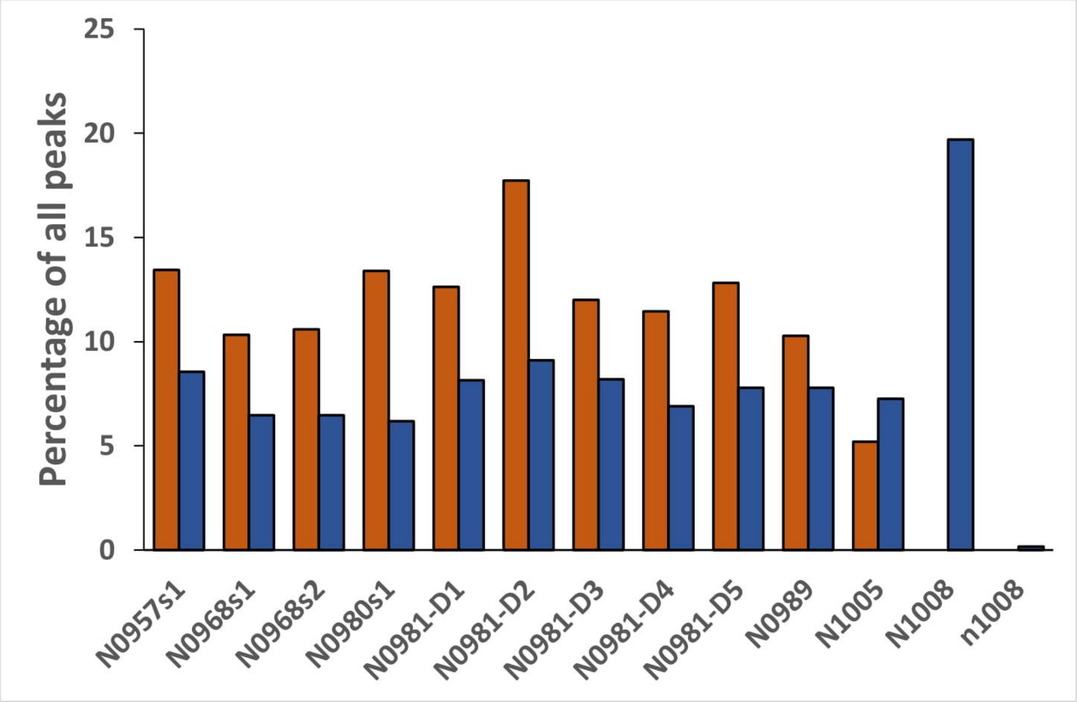
Generated using PSVS 1.5

6

Supplementary Table S5. Assessment of Contact Ambiguity

	No. of Residues	No. of Possible Contacts	Average Ambiguity per Contact	Maximum Ambiguity per Contact	Unique Long-range Contacts Total / Per Residue	Unique Long-range H^N-H^N Contacts Total / Per Residue
<u>Simulated NMR Data</u>						
N0957s1	163	5582	5	50	110 / 0.67	39 / 0.24
N0968s1	123	1506	2	16	138 / 1.12	29 / 0.24
N0968s2	115	2088	4	32	93 / 0.81	51 / 0.44
N0980s1	105	1489	3	18	92 / 0.88	34 / 0.32
N0981-D1	86	538	2	10	126 / 1.47	44 / 0.51
N0981-D2	80	504	2	8	127 / 1.59	60 / 0.75
N0981-D3	203	4701	4	32	193 / 0.95	67 / 0.33
N0981-D4	111	1093	2	10	100 / 0.90	49 / 0.44
N0981-D5	127	1983	3	21	135 / 1.06	74 / 0.58
N0989	246	7095	5	90	200 / 0.81	91 / 0.37
N1005	326	49,887	11	92	263 / 0.81	90 / 0.28
<u>Real NMR Data</u>						
N1008	97[a]		5	54	53 / 0.54	27 / 0.28
n1008	97[a]		9	169	200 / 2.06	19 / 0.20

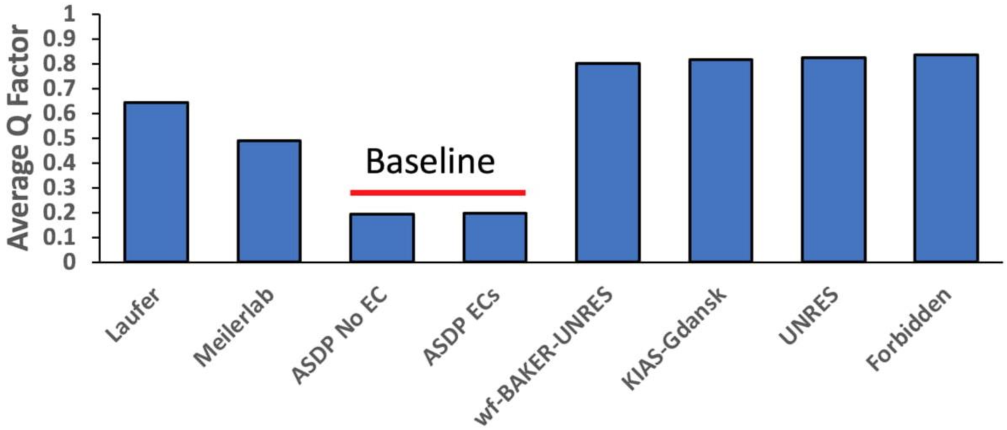
[a]Statistics for target 1008 include the N-terminal 17-residue polypeptide tail.



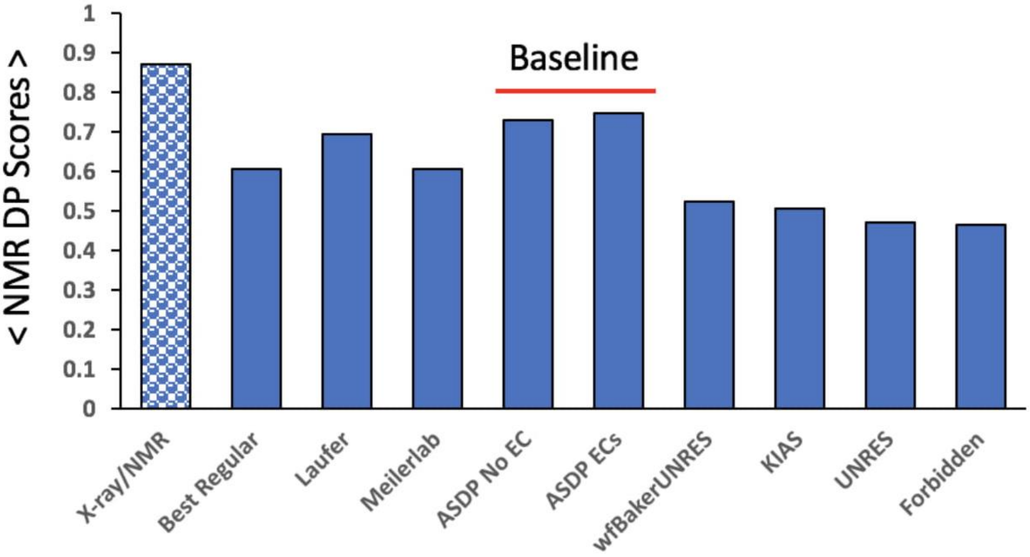
Supplementary Fig. S1. Analysis of NOESY peak lists against the reference atomic coordinates. Orange bars – percentage of all possible NOESY peaks that are removed by simulated deletions of "exchange broadened" resonances from the resonance assignment list. Blue bars – Percentage of all NOESY peaks in the real or simulated spectra that cannot be correctly assigned based on the information provided in the Ambiguous Contact Lists. "Unassignable peaks" arise either from noise peaks, which do not correspond to a true NOE interaction, or for real NOESY peaks when the true resonance that gives rise the cross peak is not assigned in the chemical shift list, leading to erroneous assignments of the NOESY cross peak. This problem is particularly severe for data set N1008 in which many sidechain-backbone NOEs are present in the NOESY peak list, but no sidechain assignments are available in the chemical shift list.

Residue 1	Residue 2	Peak No.	Upper-bound		Atom 1	Atom 2
R1	**R2**	**P#**	**UPL**	**Confid**	**A1**	**A2**
79	77	17	5.0	0.95	H	H
79	177	20	6.0	0.67	H	HD2
79	135	20	6.0	0.97	H	HD1
79	249	20	6.0	0.96	H	HD1
79	50	20	6.0	0.81	H	HD2
79	217	23	5.0	0.68	H	H
79	230	23	5.0	0.75	H	H
79	232	23	5.0	0.72	H	H
79	106	23	5.0	0.76	H	H
79	166	23	5.0	0.83	H	H
79	100	23	5.0	0.83	H	H
79	82	23	5.0	0.74	H	H
79	246	23	5.0	0.71	H	H
79	216	23	5.0	0.67	H	H
45	37	28	7.5	0.84	HD2	HG1

Supplementary Fig. S2. Format of Ambiguous Contact Lists. These data were provided in place of NOESY peak list data to CASP13 predictors. For each peak in the ^{15}N-edited or ^{13}C-edited 3D NOESY peak list (column P#), a set of ambiguous contacts were determined based on the simulated chemical shift list, using the Cycle 0 protocol of the NOESY peak assignment program *ASDP*. Possible contacts are listed between H atom 1 (Residue number R1, and Atom A1), and H atom 2 (Residue number R2, and Atom A2), together with an upper bound distance (UPL) in Å. Early Ambiguous Contact Lists included an assignment confidence score (Confid) ranging from 0 to 1, based on the quality of the match between the chemical shift values of the NOESY peak and the chemical shift values of candidate interacting atoms in the resonance assignment list. Since the Confid score was not used in CASP11, it was phased out of use during CASP13. Atom types include amide H^{N} protons (H) and various methyl proton groups (HB, HG1, HG2, HD1, HD2, etc). In this example, ASDP has uniquely assigned peak P# 20 to an interaction between the amide H^{N} of residue 77 and the amide H^{N} of residue 79, while peak P# 20 has four ambiguous assignments, H^{N} of residue 79 and methyl resonances of residues 177, 135, 249, and 50

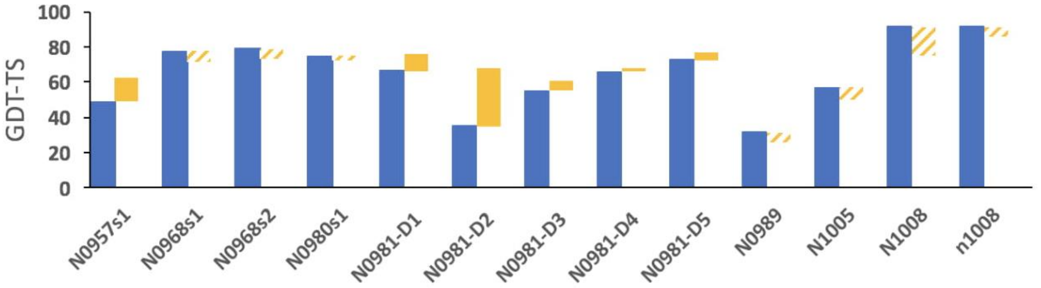


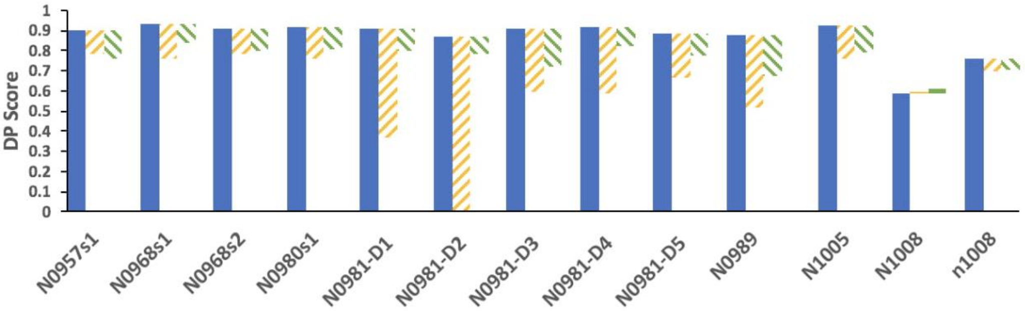
Supplemental Fig. S3. Average ^{15}N-^{1}H RDC Q-Factors, averaged over submitted evaluation units, for each predictor group.



Supplemental Fig. S4. NMR DP scores, averaged over submitted first-ranked models, for each predictor group.

10





Supplementary Fig. S5. **NMR DP Scores for X-ray Crystal Structures, Experimental NMR Structures Compared with Best Regular or Best NMR-Assisted Models**. Top (same data as main text Fig 7A): GDT-TS scores for the "best" model submitted by any NMR-assisted prediction group (blue bars) compared with the "best" model submitted by any regular prediction group (yellow solid bars show improved accuracy, and hashed yellow bars show average accuracy, due to addition of sparse NMR data). Bottom: DP scores for experimental structures determined by X-ray or NMR (blue bars) compared with the "best" model submitted by any regular prediction group (yellow bars) and the "best" model submitted by any NMR-assisted prediction group (green bars). Hashed yellow or green bars indicate targets are less accurate than the experimental structures, while solid yellow or green bars indicate targets are more accurate.

2.3.3 A protocol to automatically calculate homo-oligomeric protein structures through the integration of evolutionary constraints and ambiguous contacts derived from solid- or solution-state NMR

Davide Sala[†], *Linda Cerofolini* [‡], *Marco Fragai*[†,§], *Andrea Giachetti*[‡], *Claudio Luchinat*[†,§] *and Antonio Rosato*[†,§,*]

[†]Magnetic Resonance Center (CERM), University of Florence, Via Luigi Sacconi 6, 50019 Sesto Fiorentino, Italy.

[‡]Consorzio Interuniversitario di Risonanze Magnetiche di Metallo Proteine, Via Luigi Sacconi 6, 50019 Sesto Fiorentino, Italy.

[§]Department of Chemistry, University of Florence, Via della Lastruccia 3, 50019 Sesto Fiorentino, Italy.

Submitted – Preprint doi: https://doi.org/10.1101/714857

ABSTRACT

Protein assemblies are involved in many important biological processes. Solid-state NMR (SSNMR) spectroscopy is a technique suitable for the structural characterization of samples with high molecular weight and thus can be applied to such assemblies. A significant bottleneck in terms of both effort and time required is the manual identification of unambiguous intermolecular contacts. This is particularly challenging for homo-oligomeric complexes, where simple uniform labeling may not be effective. We tackled this challenge by exploiting coevolution analysis to extract information on homo-oligomeric interfaces from NMR-derived ambiguous contacts. After removing the evolutionary couplings (ECs) that are already satisfied by the 3D structure of the monomer, the predicted ECs are matched with the automatically generated list of experimental contacts. This approach provides a selection of potential interface residues that is used directly in monomer-monomer docking calculations. We validated the protocol on tetrameric L-asparaginase II and dimeric Sod1.

INTRODUCTION

Many proteins carry out their functional role acting as part of protein assemblies, i.e. a combination of different proteins (hetero-complexes) or of multiple copies of the same monomeric unit (homo-complexes). The assembly of the correct biological complex strongly depends upon specific protein-protein interactions (PPIs) that often are conserved among species (Qian et al., 2011; Sun and Kim, 2011). Frequently, an initial step in the study of an assembly is to characterize the three-dimensional structure of its individual subunit components either by X-ray crystallography or NMR spectroscopy. Among NMR techniques, solid-state NMR (SSNMR) has been receiving increasing attention because it is not limited by protein size, solubility, crystallization problems, presence of inorganic/organic matrices or lack of long-range order that often make the application of other structural biology methods unsuitable. In particular, it is straightforward to extend SSNMR experiments designed for individual proteins to the investigation of protein assemblies (Demers et al., 2018), as the quality of SSNMR spectra does not decrease with increasing molecular weight, as happens for solution NMR.

A crucial step in the application of SSNMR to structure determination is the identification and assignment of through-space nucleus-nucleus interactions. DARR (Dipolar Assisted Rotational Resonance) is a commonly used pulse sequence for this purpose, which is based on ^{13}C-^{13}C magnetization transfer through proton-driven spin diffusion (Takegoshi et al., 2001). Tuning of experimental DARR parameters allows users to select the range of distances at which inter-nuclear interactions are sampled. Although solid-state resonance lines of protein complexes are narrow, spectral congestion makes the assignment of DARR peaks a challenging task. In practice, DARR experiments yield a list of ambiguous contacts in which the quaternary contacts must be separated from intra-monomeric contacts to determine the 3D structure of the complex. In hetero-complexes this problem can be alleviated by using different schemes for enrichment in stable NMR-active isotopes (^{13}C, ^{15}N) in the various subunits of the complex (Göbl et al., 2014); for instance, one subunit can be uniformly enriched while all other subunits are not. This approach may not be very effective for homo-complexes, and more complex and labor intensive strategies for the asymmetric enrichment of all subunits have been proposed (Traaseth et al., 2008). Thus, the investigation of homo-complexes by SSNMR often remains a manual task, especially with respect to the identification of inter-subunit contacts.

Coevolution analysis assumes that evolutive pressure favors the preservation of protein function through the conservation of fundamental residue interactions (Salinas and Ranganathan, 2018). This concept has been implemented, among others, in global coevolutionary or direct coupling analysis (DCA) methods (Morcos et al., 2011; Weigt et al., 2008). These methods differ for the types of approximation used, from dimensional reduction (Cocco et al., 2013) to pseudo-likelihood maximization (Ekeberg et al., 2013) and others (Burger and van Nimwegen, 2010; Jones et al., 2012; Skwark and Elofsson, 2013). The information derived allows the identification of multiple protein conformational states (Morcos et al., 2013; Sutto et al., 2015) and the prediction of tertiary protein structures, either alone or in combination with experimental data (Anishchenko et al., 2017; Dago et al., 2012; Marks et al., 2012, 2011; Tang et al., 2015). Coevolution analysis can detect also ECs corresponding to inter-subunit contacts (Hopf et al., 2014; Ovchinnikov et al., 2014; Rodriguez-Rivas et al., 2016; Schug et al., 2009; Szurmant and Weigt, 2018). The identification of ECs consistent with PPIs for hetero-complexes requires the creation of a *joint* multiple sequence alignment (MSA) in which each line corresponds to an interacting protein pair (Bitbol et al., 2016; Burger and van Nimwegen, 2008; Cheng et al., 2014; Procaccini et al., 2011). This is a relatively complex task, especially due to the analysis required for the separation of orthologs and paralogs, prior to the construction of the MSA. Instead, the coevolution analysis of homo-complexes is based on a single protein sequence and thus on a single MSA. While this simplifies the construction of the alignment, it makes the identification of ECs belonging to inter-molecular contacts much more complicated because such information is hidden among hundreds or thousands of ECs of which the majority are tertiary contacts (dos Santos et al., 2015; Uguzzoni et al., 2017). The removal of tertiary contacts requires knowledge of the 3D structure of the monomeric protein. Notably, there is a relevant number (about 2000) of protein families annotated as forming homo-oligomeric assemblies *in vivo* with a deposited monomeric structure in the Protein Data Bank (PDB) (El-Gebali et al., 2019; Rose et al., 2015). These families potentially constitute an interesting target for homo-oligomeric structural predictions, also in the frame of drug discovery (Bai et al., 2016).

In the present work we developed a protocol to extract information on the protein-protein interface of homo-complexes from SSNMR-derived ambiguous contact lists, which can be automatically generated, using coevolution analysis. All the ECs with a relevant probability to be true residue interactions in either the monomer (intra-monomeric contacts) or in the homo-oligomerization interface (inter-monomeric contacts) are considered. The removal of intra-monomeric ECs requires the availability of the structure of the monomer. The predicted ECs with possible matches to experimental peaks are used to identify candidate interface residues. The final list of such residues is used directly in protein-protein docking calculations. The same protocol can be also applied using only solution-state NMR data.

RESULTS

Our protocol aims to predict the structure of homo-oligomeric complexes by using ambiguous NMR contacts to identify reliable inter-monomeric contacts within the list of ECs. The whole procedure, which is described in detail in the next section, can be divided in two main parts. First, intra-monomeric evolutionary couplings (ECs) are removed from the list of ECs based on the 3D structure of the monomer. Second, the list of ECs predicted to potentially be at the complex interface is compared with the list of ambiguous NMR contacts to extract all residue pairs matching both the predicted and the experimental dataset. The protocol was validated by predicting the tetrameric structure of *Escherichia coli* L-asparaginase II (Cerofolini et al., 2019) (PDB ID: 6EOK), in which two distinct dimeric conformations must be recognized to reconstruct the functional complex (Fig. 1). Furthermore, the robustness of the procedure in the identification of complexes with small interface regions was tested by predicting the structure of dimeric human apo Sod1 (Bertini et al., 2009) (PDB ID: 3ECU) (Fig. 1). For L-asparaginase II we used solid-state NMR data (Cerofolini et al., 2019), whereas for Sod1 we used solution NMR data (Bertini et al., 2009).

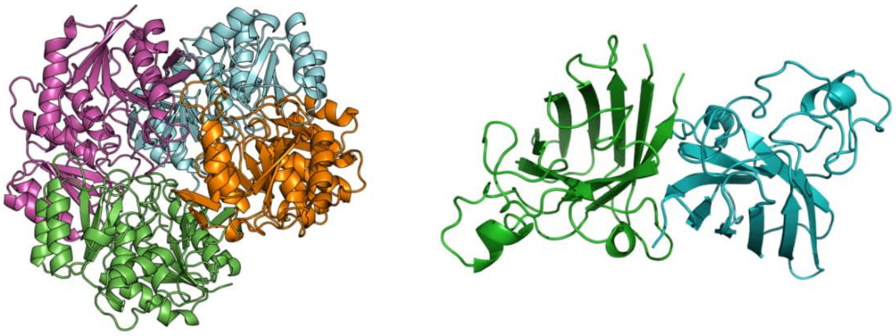


Figure 1. Crystal structures of the tetrameric L-asparaginase II and the dimeric apo Sod1.

Description and application of the protocol

This protocol calculates a list of putative interface residues to be used as input to HADDOCK for docking calculations. It needs four inputs (Fig. 2): one or more files with the list of ECs, the structure of the monomer, the experimental NMR-derived list of ambiguous contacts and the Naccess file (rsa format) with the per-residue relative solvent accessible area. The ECs of the target protein are obtained from so-called coevolution analysis. A number of servers performing coevolution analysis are available online (see *Methods*). In general, they need the protein sequence as input to predict a contact map from multiple sequence alignments (MSAs). The output is a list of residue pairs scored for the probability that they are actually in contact in the monomeric or oligomeric structure. We apply a probability cutoff P to remove ECs with low probability of being true interactions. Coevolution analysis usually outputs from hundreds to thousands of ECs that cannot be assigned as intra-monomeric or inter-monomeric contacts without any structural information. As a consequence, our protocol calculates for each EC the corresponding Cα-Cα distance in the 3D structure of the monomer and all the ECs below the distance cutoff of 12 Å are classified as intra-monomeric and removed .

After the removal of intra-monomeric ECs, the resulting list is enriched in contacts across the interaction interface (inter-monomeric ECs). Nevertheless, it still contains a relevant number of false-positives. False-positives can be either ECs that do not correspond to a true residue-residue interaction or ECs that correspond to intra-monomeric interactions that occur in conformations sampled during the physiological conformational dynamics of the protein. The EC list thus cannot be used directly in docking calculations. We thought that the rate of false positives could be reduced by leveraging the information present in the list(s) of ambiguous contacts provided by NMR experiments. Indeed, NMR-derived contacts list of protein complexes are affected by a high level of ambiguity caused by the accidental overlap of NMR resonances, making the extraction of reliable inter-monomeric contacts an arduous task. Our protocol overcomes this bottleneck by matching the predicted inter-monomeric ECs with the experimental list to extract information present in both the datasets. In practice, residue pairs in the predicted inter-monomeric EC list are matched to ambiguous assignments in the experimental list, providing a list of interface residue pairs.

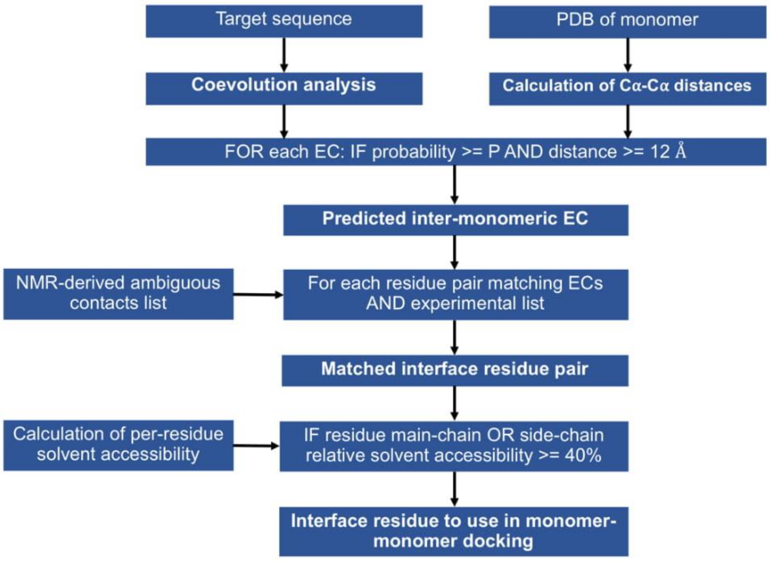

Target sequence
PDB of monomer
Coevolution analysis
Calculation of Cα-Cα distances
FOR each EC: IF probability >= P AND distance >= 12 Å
Predicted inter-monomeric EC
NMR-derived ambiguous contacts list
For each residue pair matching ECs AND experimental list
Matched interface residue pair
Calculation of per-residue solvent accessibility
IF residue main-chain OR side-chain relative solvent accessibility >= 40%
Interface residue to use in monomer-monomer docking

Figure 2. Scheme of the protocol adopted to predict the structure of homo-oligomeric complexes using coevolution analysis and ambiguous NMR contacts.

The number of residual false-positives in the matched list is further decreased by removing all the residues with a relative solvent accessibility lower than 40% in both main-chain and side-chain (i.e. buried residues). The remaining residues constituting the output list from our protocol can be used directly as ambiguous interaction restraints (AIRs) in monomer-monomer docking calculations with HADDOCK. The protocol can be run using the python script provided as supplementary material (*SI Appendix*).

We assessed the accuracy of the protocol in predicting residues at the homo-oligomeric interface for different probability cutoffs (Tables 1 and 2). Furthermore, we evaluated the NMR data contribution to the prediction accuracy by comparing the results obtained with or without ("ECs + NMR" and "ECs only", respectively) matching with the NMR data. A residue accurately predicted at the complex interface is defined as a true-positive (TP) residue. More in detail, we defined a true-positive (TP) residue as having at least one atom with a distance < 7 Å from any atom located on a different chain in the crystal structure of the complex.

In the case of the L-asparaginase II protein, the crystallographic complex is formed by four subunits with a D_2 symmetry. Thus, the ensemble of all TP residues contains the amino acids at both dimeric interfaces. For this system, the inclusion of NMR data enhances the positive predictive value (PPV), defined as true-positive (TP) residue predictions over all predictions [TP/(TP+FP)], at all the probability cutoffs assessed (Table 1). In fact, on the basis of the "ECs only" analysis the absolute number of TP residues present in the prediction is significantly higher than the number of TP obtained after the match with NMR data. However, the same analysis also outputs a much greater number of FPs. Consequently, the "ECs + NMR" analysis features a PPV of 100% for P >= 0.35; the PPV remains very high (>= 80%) even at low probabilities (P < 0.35) and the number of predicted interface residues is sufficient to successfully drive docking calculations (see next section).

Table 1. Number of residues predicted to make contacts across the L-asparaginase II homomeric interface. The protocol was applied as depicted in figure 2 with the ECs matched with the NMR data "ECs + NMR" and without the matching step with NMR data "ECs only". P indicates the probability threshold used to accept ECs. PPV = TP/(TP+FP).

	L-asparaginase II					
	ECs only			**ECs + NMR**		
P	**TP+FP**	**TP**	**PPV**	**TP+FP**	**TP**	**PPV**
0.90	13	10	0.8	3	3	1.0
0.85	23	20	0.9	3	3	1.0
0.80	30	21	0.8	3	3	1.0
0.75	34	24	0.8	3	3	1.0
0.70	38	27	0.8	4	4	1.0
0.65	41	30	0.8	4	4	1.0
0.60	47	31	0.7	4	4	1.0
0.55	51	33	0.7	4	4	1.0
0.50	60	36	0.7	4	4	1.0
0.45	73	42	0.7	4	4	1.0
0.40	84	47	0.6	5	5	1.0
0.35	97	52	0.6	7	7	1.0
0.30	105	54	0.6	9	8	0.9
0.25	121	60	0.6	19	16	0.8
0.20	128	60	0.6	34	28	0.8

Instead, the Sod1 complex contains two subunits with a C_2 symmetry and a small protein-protein interface. As a consequence, in the central part of the interface the inter-monomeric contacts involve residue pairs that also are at intra-monomer distance smaller than the 12 Å threshold that we used to remove intra-monomeric ECs. In practice, this structural organization significantly reduces the number of detectable TPs because the aforementioned inter-monomeric contacts are discarded. Furthermore, small interfaces are harder to predict computationally and also provide a lower number of NMR-detectable contacts. All these features make the Sod1 system challenging but useful to test the limits of the protocol. When considering the Sod1 protein, the "ECs only" protocol yielded a reasonable PPV for P >= 0.55, but with only a handful of TPs in the prediction (Table 2). Instead, the match with NMR data removed the signal for P>= 0.45 while retaining information at lower P values, especially for P = 0.30.

Table 2. Number of residues predicted to make contacts across the Sod1 homomeric interface.

	Sod1					
	ECs only			**ECs + NMR**		
P	**TP+FP**	**TP**	**PPV**	**TP+FP**	**TP**	**PPV**
0.90	0	0	NA	0	0	NA
0.85	0	0	NA	0	0	NA
0.80	0	0	NA	0	0	NA
0.75	4	3	0.7	0	0	NA
0.70	4	3	0.7	0	0	NA
0.65	4	3	0.7	0	0	NA
0.60	8	4	0.6	2	0	NA
0.55	10	4	0.4	2	0	0.0
0.50	17	5	0.2	4	0	0.0
0.45	23	7	0.3	5	1	0.2
0.40	29	9	0.3	5	1	0.2
0.35	38	12	0.3	9	3	0.3

0.30	50	14	0.3	18	7	0.4
0.25	68	17	0.2	27	7	0.3
0.20	74	17	0.2	48	9	0.2

These results suggest that the quality of the initial EC prediction is quite important for the performance of our protocol, leading to a larger enhancement of the PPV when the prediction includes a larger number of TPs. When the EC data yielded is weaker and mixed with noise, our protocol retains a good part of the available information but the PPV is mostly unchanged.

HADDOCK calculations for L-asparaginase II

The ECs at the P cutoff of 0.25 were matched with a solid state 2D^{13}C-^{13}C DARR dataset (mixing time 200 ms) holding 4937 ambiguous assignments, resulting in 19 surface residues predicted to be at the protein-protein interface (corresponding to 14% of the whole protein surface). The final 200 water-refined models generated by HADDOCK were analyzed by measuring the RMSD from the structure with the lowest HADDOCK score. The clustering algorithm grouped the models in 7 clusters (Fig. 3A). The first cluster was the most populated and included the models with the lowest score. Indeed, the lowest HADDOCK score model of the first cluster was a dimer with an RMSD of 0.7 Å from the crystallographic dimer formed by chain A and chain C of the tetrameric protein (Fig. 3B). In addition to the HADDOCK score, the desolvation energy calculated using empirical atomic solvation parameters proved to be an useful scoring function(Fernández-Recio et al., 2004), allowing the identification of the correct A-C dimer (Fig. S1).

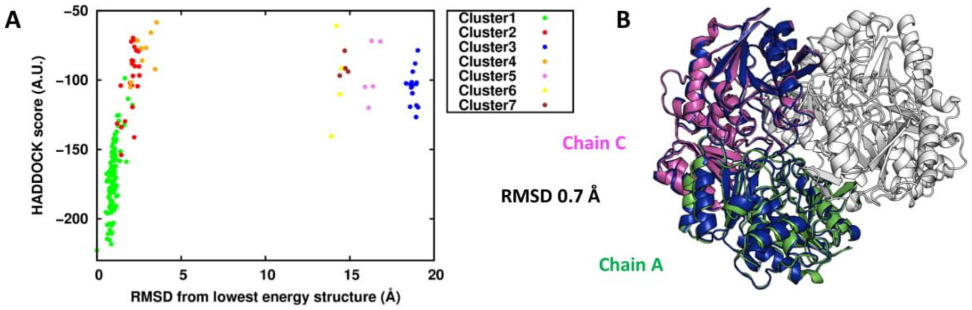


Figure 3. L-asparaginase II monomer-monomer docking. **A**) Plot of the HADDOCK score vs RMSD clusters distribution with respect to the lowest HADDOCK score model. **B**) Structural alignment between the lowest HADDOCK score model (in blue) of the first cluster and the crystal structure.

Both the predicted inter-monomeric ECs and the experimental NMR inter-monomeric contacts include residue pairs belonging to all the pairs of chains effectively in contact in the functional complex. In the case of the tetrameric L-asparaginase II, besides the largest A-C interface also chains A and D share a relevant number of contacts. According to this, in a single docking run one might expect to sample both relevant dimeric configurations (A-C and A-D) in two different clusters. Indeed, by checking the position of the 19 predicted interface residues within the crystal structure, it appears that the A-C and A-D interfaces were both mapped (Fig. 4). In fact, the largest portion of residues effectively in contact belonged to dimer A-C and the smallest portion to dimer A-D.

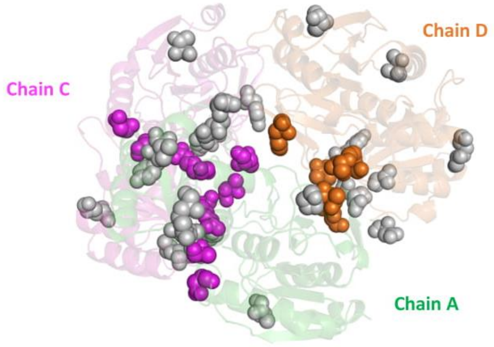


Figure 4. Projection on the crystal structure of the L-asparaginase II residues used to generate AIRs in the docking calculation. The residues making inter-monomeric contacts are shown as colored spheres (A-C interface in purple; A-D interface in orange).

However, the structural configuration present in the other clusters did not correspond to the A-D dimer. This could be easily verified by observing that the superimposition of the two dimers on the common chain A resulted in evident steric clashes between the subunits, as shown for the cluster 3 (Fig. 5). If the two dimers actually corresponded to the A-C and A-D dimers of the tetrameric structure, the superimposition on the A chain would have caused no significant clashes.

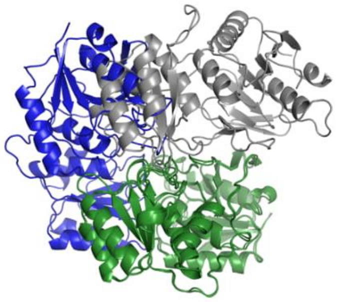


Figure 5. Superimposition on chain A (in green) of the third (in gray) and the best (in blue) dimer configurations in the first run.

In principle, the absence of the second compatible dimer in calculations can be due to two reasons. First, the interface residues belonging to the second configuration were not present in the AIRs dataset. Second, the residues belonging to the second interface region were present, but the correct structural configuration had a HADDOCK score worse than the wrong sampled configurations. In the present case, the latter reason was the relevant one. In fact, the wrong dimer models in general contained some contacts from both interface regions, thus satisfying a higher number of AIRs than the correct dimer A-D.

To obtain a model of the A-D dimer, we performed a second docking run in which the restraints already satisfied in the best cluster (containing the most favored configuration) of the first run were removed from the input dataset. To this end, we looked at the violation analysis of HADDOCK, and retained all contacts that were not satisfied by the majority of the members of the first cluster by at least 3 Å. This resulted in 9 residues being used as input to a second monomer-monomer docking run. As in the previous calculation, the first cluster was the largest and contained the models with the best HADDOCK score and desolvation energy (Fig. 6A and S2). Superimposing the lowest HADDOCK score water-refined model with the crystal structure resulted in an RMSD of 0.9 Å from the dimer A-D (Fig. 6B).

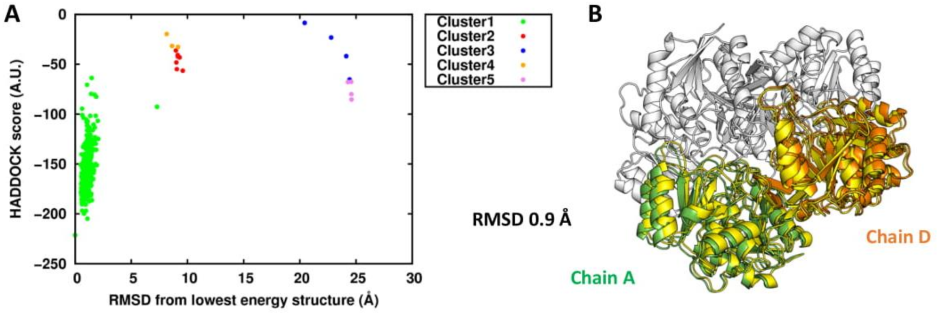


Figure 6. L-asparaginase II monomer-monomer docking using AIRs violated in the A-C dimeric model. **A**) Plot of the HADDOCK score vs RMSD clusters distribution with respect to the lowest HADDOCK score model **B**) Structural alignment between the lowest HADDOCK score model (in yellow) of the first cluster and the crystal structure.

In summary, the two correct dimeric conformations A-C and A-D were obtained performing two distinct docking runs, the first one with the whole AIRs dataset and the second one with the subset resulting from the removal of the AIRs satisfied in the best cluster of the first run. Crucially, this procedure provided us with two compatible non-overlapping dimeric models that, for symmetry, can be used to reconstruct the tetrameric model (Fig. 7). This step strictly depended by the correct identification of the structural model on which the distance violation analysis was carried out. In fact, selecting the third cluster of Fig. 3 to perform the violation analysis instead of the best one resulted in a second docking run that sampled again the dimer A-C in the two best clusters and not-compatible structural configurations in the others (Fig. S3).

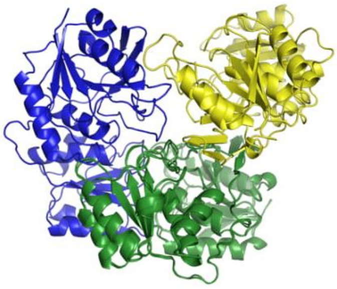


Figure 7. Superimposition on the chain A (in green) of the best structural configurations in the second run (in yellow) and in the first run (in blue).

Extracting the monomer from the PDB of the complex results in a protein model with the side chains oriented in a contact-ready state that favors the correct assembly, in terms of both docking score and RMSD from the experimental structure, as compared to incorrect docking poses.

Thus, to test our protocol in a more realistic condition we generated 15 homology models of L-asparaginase II using the structure of the homolog from *Wolinella succinogenes*(Lubkowski et al., 1996) as the structural template (PDB ID 1WSA, chain A). The homology models had a backbone RMSD lower than 1 Å from the crystal structure of the *E. coli* protein, but widely differing in the orientation of the surface side chains. Each model was used in protein-protein docking with the same input AIRs of the "crystal P 0.25" runs, for both the A-C and A-D dimers. The results of Table 3 show the significant influence of the orientation of side chains on the ability of the docking calculations to sample the correct dimer in the best cluster. Based on the HADDOCK score of the best cluster for each model, the AC runs pointed out that the five runs with the best score also had the lowest RMSD from the crystal A-C dimer, (green gradient in the table). However, for these five models the second calculation with the AIRs providing the A-D dimer resulted in wrong dimeric conformations. Nevertheless, by inspecting the results for all models (Table 3), it turned out that the runs with the best HADDOCK scores (for their first clusters) indeed provided results conformations close to the crystallographic A-D dimer (in particular models 6 and 15). For further comparison, we performed a docking run of the crystallographic monomer with the 34 residues (25% of the whole protein surface) output by the protocol run at a P cutoff of 0.20. Changing the AIRs dataset with a larger one having the same PPV did not significantly affect the results.

Overall, the results described above pointed out the importance of generating a sufficiently large number of homology models to sample many different side chain orientations, thus increasing the probability to capture the orientation permitting residue-residue contacts across the monomeric interface. The best clusters of the two crystal runs showed that ideal side chain orientations provided the top HADDOCK score values. In line with this, the models that had the best HADDOCK scores resulted in the configurations closest to the crystal structure, with a backbone RMSD between 1 and 3 Å from it. For these models, the HADDOCK scores themselves were similar to the values observed for the runs starting from the crystal monomer. Indeed, superimposing on the chain A the AC dimer of model 13 and the AD dimer of model 15 or model6 showed two compatible dimeric models that, taken together, can be used to reconstruct the tetrameric structure (Figure S4)

Table 3. Docking results for homology models of L-asparaginase II. The two "Crystal" runs were performed using the chain A of the crystal structure. Each model mainly differs in the orientation of side chains. For each run the HADDOCK score of the best cluster (calculated as the average value of the 4 best structures of the cluster) and the RMSD of its best structure from the experimental dimer are reported.

	A-C dimer		A-D dimer	
	HADDOCK score	**RMSD**	**HADDOCK score**	**RMSD**
Crystal P 0.25	-218	0.7	-206	0.9
Crystal P 0.20	-170	0.7	-187	1.3
model1	-204	1.4	-121	16.1
model2	-93	17.9	-116	9
model3	-101	21.9	-104	14.5
model4	-72	16.6	-109	3.2
model5	-141	16.1	-91	4.3
model6	-95	14.4	-159	1
model7	-166	1.3	-113	14.5
model8	-72	16.6	-109	3.2
model9	-106	18.9	-120	8.3
model10	-184	2.2	-123	8.8
model11	-187	1.5	-132	11.6

model12	-117	18.9	-124	7.4
model13	-204	1.4	-121	16.1
model14	-101	21.9	-104	11
model15	-134	18.5	-169	2.6

HADDOCK calculations for Sod1

The predicted inter-monomeric ECs at P=0.30 were matched with 7611 ambiguous assignments from solution-state 3D ^{1}H ^{15}N NOESY-HSQC spectrum. The protocol yielded 18 putative interface residues, corresponding to 23% of the whole monomer surface. By comparing the prediction to the of the crystal structure, it appeared that 7 out of 18 residues effectively formed inter-monomeric contacts (Fig. S5).

From the docking calculation starting with the crystal monomer we obtained 7 clusters with comparable HADDOCK score values (Fig. 8A). However, the distribution of the desolvation energies discriminated the second cluster as the most favored (Fig. 8B). Indeed, the structural alignment of the best model of this cluster with the experimental dimer revealed an impressive RMSD of 0.6 Å (Fig. S6A). Instead, the same superimposition on the crystal structure of the first cluster resulted in a dimer in which one of the two monomeric units was rotated by 180° with respect to the corresponding experimental monomer, while preserving the same interface region (Fig. S6B).

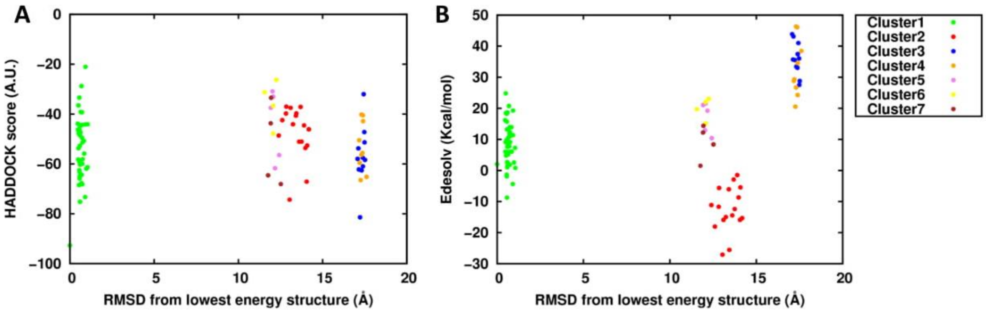


Figure 8. Sod1 clusters distribution with respect to the lowest HADDOCK score model. **A**) HADDOCK score distribution. **B**) Desolvation energy distribution.

DISCUSSION

Solid State NMR is an attractive technique to study large protein assemblies as even systems with high molecular weight can provide very good spectra. However, the determination of their 3D structure involves two very time-consuming steps: the assignment of the side chains in contact at the interface between the subunits and, for homo-oligomeric complexes, the discrimination of intra- vs inter-monomer contacts. In particular, the correct identification of inter-monomer contacts usually requires extensive efforts by an experienced user. From the bioinformatics point of view, focusing on homo- rather than hetero-oligomers makes the interpretation of coevolution signals harder. In fact, the difficult step in the coevolution analysis of hetero-oligomers is the proper pairing of orthologs of interacting proteins and the corresponding removal of paralogs. Once this has been achieved, the creation of a *joint* MSA in which each line contains a pair of interacting proteins allows the straightforward use of predicted inter-protein contacts as restraints to drive the modelling of the quaternary structure (Bitbol et al., 2016; Hopf et al., 2014; Ovchinnikov et al., 2014). Instead, the coevolution analysis of homo-oligomers is based on a single protein MSA, which is relatively effortless to build. Unfortunately, the availability of the three-dimensional structure of the monomeric unit is necessary to successfully separate intra-monomeric and inter-monomeric ECs (Uguzzoni et al., 2017). In this work, we developed a protocol to integrate ECs with NMR-derived ambiguous contacts in order to identify interface residues in homo-oligomers. The input lists of ambiguous contacts can be automatically generated from appropriate solution or solid-state NMR spectra. Our protocol was validated by predicting two difficult cases: the tetrameric L-asparaginase II, in which two distinct dimeric conformations must be recognized to reconstruct the functional complex and the dimeric Sod1, in which the interface region is comparatively small.

The correct identification of interface residues was readily verified by comparing the output of the protocol with the known interfaces in the crystal structures of the two systems (Tables 1 and 2). This analysis showed that NMR data can be beneficial by enriching the predictions in true contacts (i.e. higher PPV). This improvement comes at the cost of reducing the absolute number of predicted residues, which however did not limit the subsequent docking calculations. The requisite for the integration of ECs and NMR data to be successful is that the initial list of potential inter-monomeric ECs contains enough information. This is clearly exemplified by the case of Sod1, for which the absolute number of predictions, after removing all contacts that could be satisfied within the monomer, was quite low. Consequently, many NMR signals could not be matched and the benefit in PPV was modest. Nevertheless, when the total number of predicted interface residues is in a reasonable range (15%-20% of all surface residues, i.e. 12-16 residues for Sod1) the prediction resulting from the integration of ECs and NMR data is more reliable than that based only on ECs.

To generate a 3D structural model of the oligomer, the output of our protocol can be exploited in docking calculations. As a proof-of-principle, we run these calculations starting from the monomer conformation observed in the crystal structure. This is an ideal case, where all the side chains at the protein-protein interface are already in the correct rotameric state to engage in the formation of the complex. All the same, it is important to perform this step to ensure that the output contains enough information to successfully drive the docking. This was indeed the case for the main dimer of L-asparaginase II (A-C) as well as for Sod1. The calculation with the complete AIR dataset could not identify the A-D dimer even though the dataset contained contacts belonging to both interfaces. The A-D interface is somewhat smaller than the A-C interface; as HADDOCK aims to satisfy the highest number of AIRs, the situation where the second chain of the dimer is positioned in between the two interfaces, thus partly satisfying both subsets of AIRs, is favored over the situation in which all of the A-D and none of the A-C AIRs are satisfied. To circumvent this bottleneck, it is necessary to separate the residues belonging to each interface. This was done by removing the contacts already satisfied in the first docking calculation to run a second calculation only with the unsatisfied restraints. The best cluster of the second run indeed matched closely the A-D dimer of the tetramer (Fig 6). Intriguingly, the AIRs derived from ECs only at a P cutoff of 0.8 (Table 1), whose number was similar to the number of AIRs used in the “ECs + NMR” calculations, did not contain information on the A-D dimer interface (not shown). Thus, the information provided by ECs at high levels of confidence is not balanced over the two interfaces, presumably due to the evolutionary history of the system. This makes it necessary to use data at lower P cutoffs, which is efficiently filtered by the ambiguous contacts provided by solid state NMR. The experimental data in fact contain information on both interfaces and thus is useful to extract both sets of true contacts from the list of ECs.

In a more realistic scenario one would use a homology model of the monomer as the input structure to docking calculations. We tested this scenario by generating 15 different models of L-asparaginase II (Table 3) and using the same input AIRs used in the docking of the crystal monomer for all calculations, so that the structure was the only source of variability. For the A-C dimer, we observed that in four cases the best model of the adduct was within 2 Å from the crystal structure, while an additional calculation provided a model with a RMSD of 2.2 Å. The A-D dimer resulted in a similar situation, with two structures within 3 Å and another two at 3.2 Å. Remarkably, there was a very good correlation between the HADDOCK score and the RMSD, allowing the more accurate models to be identified quite straightforwardly. It is also noteworthy that the best results obtained with the homology models had scores close to those obtained with the crystal monomer, which can be reasonably assumed to represent the best possible score. It thus appears that sampling a relatively extensive ensemble of different conformations is an important factor to obtain accurate models of the oligomer in a real-life setting.

In summary, our protocol allowed us to predict homo-oligomeric structure in multimers and in presence of a small homodimerization interface. Notably, this goal was achieved with a minimal user effort, making the determination of the 3D structure of the complex faster than using experimental data alone. The only parameter that must be decided by the user is the probability cutoff P below which the ECs are removed. In our hands selecting a P cutoff such that the number of predicted interface residues was 15%-20% of the number of surface residues in the monomer worked well. The results of our protocol clearly depend upon the quality of the ECs obtained from the online servers. Their integration with NMR data serves two different purposes, namely enriching the input AIRs in true contacts when working at low P cutoffs and removing biases among different regions of the protein. From the point of view of NMR spectroscopists, the present work provides a methodology to analyze homo-oligomers with reduced manual effort.

METHODS

Computational aspects

The protocol described in the "results" section can be carried out running the python script provided (*SI Appendix*). The script needs four inputs: the ECs files, the PDB structure of the monomeric protein, the experimental ambiguous NMR contacts list and the Naccess file (rsa format) with the relative solvent accessibility of the residues. Details about inputs preparation, script steps, and docking protocol adopted for the L-asparaginase II and Sod 1 are described below.

The ECs for both proteins were collected using 3 servers available online: Gremlin (Ovchinnikov et al., 2014) (http://gremlin.bakerlab.org), RaptorX (Wang et al., 2017; Xu et al., 2016) (http://raptorx.uchicago.edu/) and ResTriplet (Yang Li, Chengxin Zhang, Dongjun Yu, 2018) (https://zhanglab.ccmb.med.umich.edu/ResTriplet/). The last two methods are supervised but the PDBs used in this work were not present in the training sets. The MSA in the Gremlin server was generated with the Jackhmmer method and default options (Eddy, 1998). Using different servers adopting different methods in the ECs generation can result in multiple copies of the same EC with different computed likelihood probability. If this is the case, the EC with the highest probability is kept.

The reference protein structures were retrieved from the Protein Data Bank: *E. coli* L-asparaginase II corresponds to PDB ID 6EOK, whereas human apo-Sod1 has the PDB ID 3ECU. Inter-monomeric ECs were identified by removing from the full EC lists all residue pairs with a corresponding Cα-Cα distance < 12 Å in chain A of the structures. This distance was already proved as an excellent threshold in the selection of true contacts across the interface (Uguzzoni et al., 2017).

The experimental procedure for the generation of the ambiguous NMR contacts list is described in the next section.

The per-residue relative solvent accessible area for both main chain and side chain was calculated with Naccess (Hubbard, S. J. and Thornton, 1993). Our python script requires the Naccess file in the rsa format to automatically remove all the residues with a relative solvent accessible area below 40% for both the side chain and the main chain.

The monomer-monomer docking calculations were carried out with the HADDOCK software (Dominguez et al., 2003). The residues chosen to drive the docking run were given as active residues (directly involved in the interaction) to generate ambiguous interaction restraints (AIRs) with the default upper distance limit of 2 Å. The water-refined models were clustered based on the fraction of common contacts (Rodrigues et al., 2012), FCC = 0.75, and the minimum number of elements in a cluster of 4. For the docking run starting from crystal structures, chain A was used as the input monomer. The number of models generated for each step of the HADDOCK docking procedure were set as follow: 10000 for rigid-body energy minimization, 400 for semi-flexible simulated annealing and 400 for refinement in explicit solvent. The distance violation analysis was performed on the best cluster and the corresponding output written in the ana_dist_viol_all.lis file. In this file we selected all the residues with a violation larger than 3 Å to generate a subset of AIRs to drive a second docking run. Thus, the second docking run was performed using exactly the same conditions as the first one.

We generated 15 models of monomeric *E. coli* L-asparaginase II using the structure of *Wolinella succinogenes* L-asparaginase (Lubkowski et al., 1996) as a template (PDB ID 1WSA, chain A) using Modeller (Eswar et al., 2007). The two proteins have 55% sequence identity. The resulting template-based models featured a very similar backbone conformation, lower than 1 Å from the *E. coli* crystal, but different side chain orientations. Each model was assessed in protein-protein docking using the same AIRs used in the “crystal P 0.25” runs, with all the AIRs (A-C dimer calculation) and after the removal of the ones already satisfied by the A-C dimer (A-D dimer calculation), respectively. The number of models generated for each step were reduced as follow: 1000 for rigid-body energy minimization, 200 for semi-flexible simulated annealing and 200 for refinement in explicit solvent.

All the RMSD values reported in this work were measured on the Cα atoms.

Solid- and solution-state NMR data

The L-asparaginase II protein [U- ^{13}C, ^{15}N] was expressed and purified as previously reported (Cerofolini et al., 2019; Giuntini et al., 2017b, 2017a; Ravera et al., 2016), freeze-dried and packed (ca. 20 mg) into a Bruker 3.2 mm zirconia rotor. The material was rehydrated with a solution of 9 mg/mL NaCl in MilliQ H_2O; the hydration process was monitored through 1D {^{1}H}-^{13}C cross-polarization SSNMR spectrum and stopped when the resolution of the spectrum did not change any further for successive additions of the solution (Giuntini et al., 2017b, 2017a; Ravera et al., 2016). Silicon plug, (courtesy of Bruker Biospin) placed below the turbine cap, was used to close the rotor and preserve hydration.

SSNMR experiments were recorded on a Bruker AvanceIII spectrometer operating at 800 MHz (19 T, 201.2 MHz ^{13}C Larmor frequency) equipped with Bruker 3.2 mm Efree NCH probe-head. All spectra were recorded at 14 kHz MAS frequency and the sample temperature was kept at ≈ 290 K.

Standard ^{13}C-^{13}C correlation spectra (Dipolar Assisted Rotational Resonance, DARR) with different mixing times (50, 200 and 400 ms) were acquired using the pulse sequences reported in the literature(Takegoshi et al., 2001). Pulses were 2.6 µs for ^{1}H, 4 µs for ^{13}C; the spectral width was set to 282 ppm; 2048 and 1024 points were acquired in the direct and indirect dimensions, respectively; 128 scans were acquired; the inter-scan delay was set to 1.5 s in all the experiments.

All the spectra were processed with the Bruker TopSpin 3.2 software package and analyzed with the program CARA (Keller, 2007).

The assignment of the carbon resonances of the 2D ^{13}C-^{13}C DARR spectra of rehydrated freeze-dried ANSII was easily obtained by comparison with the 2D ^{13}C-^{13}C DARR spectrum collected on the crystalline and PEGylated preparations of L-asparaginase II (Cerofolini et al., 2019; Ravera et al., 2016).

The experimental data used for the Sod1 protein were taken from deposited solution-state 3D ^{1}H-^{15}N NOESY-HSQC spectrum (Bertini et al., 2009).

Ambiguous assignment lists of the 2D ^{13}C-^{13}C DARR and 3D ^{1}H-^{15}N NOESY-HSQC peaks were generated with the program ATNOS/CANDID (Andreas et al., 2016; Guerry and Herrmann, 2012) by setting the chemical-shift–based assignment tolerances to 0.25 ppm and 0.025 ppm, respectively.

ACKNOWLEDGEMENTS

Financial support was provided by the European Commission (project no. 777536).

We thank Prof. Gaetano Montelione for many useful discussions.

COMPETING INTERESTS

None.

REFERENCES

Andreas LB, Jaudzems K, Stanek J, Lalli D, Bertarello A, Le Marchand T, Cala-De Paepe D, Kotelovica S, Akopjana I, Knott B, Wegner S, Engelke F, Lesage A, Emsley L, Tars K, Herrmann T, Pintacuda G. 2016. Structure of fully protonated proteins by proton-detected magic-angle spinning NMR. *Proc Natl Acad Sci* **113**:9187–9192. doi:10.1073/pnas.1602248113

Anishchenko I, Ovchinnikov S, Kamisetty H, Baker D. 2017. Origins of coevolution between residues distant in protein 3D structures. *Proc Natl Acad Sci* **114**:9122–9127. doi:10.1073/pnas.1702664114

Bai F, Morcos F, Cheng RR, Jiang H, Onuchic JN. 2016. Elucidating the druggable interface of protein−protein interactions using fragment docking and coevolutionary analysis. *Proc Natl Acad Sci* **113**:E8051–E8058. doi:10.1073/pnas.1615932113

Bertini I, Cantini F, Vieru M, Banci L, Girotto S, Boca M, Calderone V. 2009. Structural and dynamic aspects related to oligomerization of apo SOD1 and its mutants. *Proc Natl Acad Sci* **106**:6980–6985. doi:10.1073/pnas.0809845106

Bitbol A-F, Dwyer RS, Colwell LJ, Wingreen NS. 2016. Inferring interaction partners from protein sequences. *Proc Natl Acad Sci* **113**:12180–12185. doi:10.1073/pnas.1606762113

Burger L, van Nimwegen E. 2010. Disentangling Direct from Indirect Co-Evolution of Residues in Protein Alignments. *PLoS Comput Biol* **6**:e1000633. doi:10.1371/journal.pcbi.1000633

Burger L, van Nimwegen E. 2008. Accurate prediction of protein–protein interactions from sequence alignments using a Bayesian method. *Mol Syst Biol* **4**:165. doi:10.1038/msb4100203

Cerofolini L, Giuntini S, Carlon A, Ravera E, Calderone V, Fragai M, Parigi G, Luchinat C. 2019. Characterization of PEGylated Asparaginase: New Opportunities from NMR Analysis of Large PEGylated Therapeutics. *Chem – A Eur J* **25**:1984–1991. doi:10.1002/chem.201804488

Cheng RR, Morcos F, Levine H, Onuchic JN. 2014. Toward rationally redesigning bacterial two-component signaling systems using coevolutionary information. *Proc Natl Acad Sci* **111**:E563–E571. doi:10.1073/pnas.1323734111

Cocco S, Monasson R, Weigt M. 2013. From Principal Component to Direct Coupling Analysis of Coevolution in Proteins: Low-Eigenvalue Modes are Needed for Structure Prediction. *PLoS Comput Biol* **9**:e1003176. doi:10.1371/journal.pcbi.1003176

Dago AE, Schug A, Procaccini A, Hoch JA, Weigt M, Szurmant H. 2012. Structural basis of histidine kinase autophosphorylation deduced by integrating genomics, molecular dynamics, and mutagenesis. *Proc Natl Acad Sci* **109**:E1733–E1742. doi:10.1073/pnas.1201301109

Demers J-P, Fricke P, Shi C, Chevelkov V, Lange A. 2018. Structure determination of supra-molecular assemblies by solid-state NMR: Practical considerations. *Prog Nucl Magn Reson Spectrosc* **109**:51–78. doi:10.1016/J.PNMRS.2018.06.002

Dominguez C, Boelens R, Bonvin AMJJ. 2003. HADDOCK: A protein-protein docking approach based on biochemical or biophysical information. *J Am Chem Soc* **125**:1731–1737. doi:10.1021/ja026939x

dos Santos RN, Morcos F, Jana B, Andricopulo AD, Onuchic JN. 2015. Dimeric interactions and complex formation using direct coevolutionary couplings. *Sci Rep* **5**:13652. doi:10.1038/srep13652

Eddy SR. 1998. Profile hidden Markov models. *Bioinformatics* **14**:755–63.

Ekeberg M, Lövkvist C, Lan Y, Weigt M, Aurell E. 2013. Improved contact prediction in proteins: Using pseudolikelihoods to infer Potts models. *Phys Rev E* **87**:012707. doi:10.1103/PhysRevE.87.012707

El-Gebali S, Mistry J, Bateman A, Eddy SR, Luciani A, Potter SC, Qureshi M, Richardson LJ, Salazar GA, Smart A, Sonnhammer ELL, Hirsh L, Paladin L, Piovesan D, Tosatto SCE, Finn RD. 2019. The

Pfam protein families database in 2019. *Nucleic Acids Res* **47**:D427–D432. doi:10.1093/nar/gky995

Eswar N, Webb B, Marti-Renom MA, Madhusudhan MS, Eramian D, Shen M, Pieper U, Sali A. 2007. Comparative Protein Structure Modeling Using MODELLERCurrent Protocols in Protein Science. Hoboken, NJ, USA: John Wiley & Sons, Inc. pp. 2.9.1-2.9.31. doi:10.1002/0471140864.ps0209s50

Fernández-Recio J, Totrov M, Abagyan R. 2004. Identification of protein-protein interaction sites from docking energy landscapes. *J Mol Biol* **335**:843–65.

Giuntini S, Balducci E, Cerofolini L, Ravera E, Fragai M, Berti F, Luchinat C. 2017a. Characterization of the Conjugation Pattern in Large Polysaccharide-Protein Conjugates by NMR Spectroscopy. *Angew Chemie Int Ed* **56**:14997–15001. doi:10.1002/anie.201709274

Giuntini S, Cerofolini L, Ravera E, Fragai M, Luchinat C. 2017b. Atomic structural details of a protein grafted onto gold nanoparticles. *Sci Rep* **7**:17934. doi:10.1038/s41598-017-18109-z

Göbl C, Madl T, Simon B, Sattler M. 2014. NMR approaches for structural analysis of multidomain proteins and complexes in solution. *Prog Nucl Magn Reson Spectrosc*. doi:10.1016/j.pnmrs.2014.05.003

Guerry P, Herrmann T. 2012. Comprehensive Automation for NMR Structure Determination of ProteinsMethods in Molecular Biology (Clifton, N.J.). pp. 429–451. doi:10.1007/978-1-61779-480-3_22

Hopf TA, Schärfe CPI, Rodrigues JPGLM, Green AG, Kohlbacher O, Sander C, Bonvin AMJJ, Marks DS. 2014. Sequence co-evolution gives 3D contacts and structures of protein complexes. *Elife* **3**:1–45. doi:10.7554/eLife.03430

Hubbard, S. J. and Thornton JM. 1993. NACCESS.

Jones DT, Buchan DWAA, Cozzetto D, Pontil M. 2012. PSICOV: Precise structural contact prediction using sparse inverse covariance estimation on large multiple sequence alignments. *Bioinformatics* **28**:184–190. doi:10.1093/bioinformatics/btr638

Keller R. 2007. The Computer Aided Resonance Tutorial 81.

Lubkowski J, Palm GJ, Gilliland GL, Derst C, Röhm KH, Wlodawer A. 1996. Crystal structure and amino acid sequence of Wolinella succinogenes L-asparaginase. *Eur J Biochem* **241**:201–207. doi:10.1111/j.1432-1033.1996.0201t.x

Marks DS, Colwell LJ, Sheridan R, Hopf TA, Pagnani A, Zecchina R, Sander C. 2011. Protein 3D Structure Computed from Evolutionary Sequence Variation. *PLoS One* **6**:e28766. doi:10.1371/journal.pone.0028766

Marks DS, Hopf TA, Sander C. 2012. Protein structure prediction from sequence variation. *Nat Biotechnol* **30**:1072–1080. doi:10.1038/nbt.2419

Morcos F, Jana B, Hwa T, Onuchic JN. 2013. Coevolutionary signals across protein lineages help capture multiple protein conformations. *Proc Natl Acad Sci* **110**:20533–20538. doi:10.1073/pnas.1315625110

Morcos F, Pagnani A, Lunt B, Bertolino A, Marks DS, Sander C, Zecchina R, Onuchic JN, Hwa T, Weigt M, Zecchina R, Morcos F, Hwa T, Sander C, Pagnani A, Bertolino A, Lunt B, Weigt M. 2011. Direct-coupling analysis of residue coevolution captures native contacts across many protein families. *Proc Natl Acad Sci* **108**:E1293–E1301. doi:10.1073/pnas.1111471108

Ovchinnikov S, Kamisetty H, Baker D. 2014. Robust and accurate prediction of residue-residue interactions across protein interfaces using evolutionary information. *Elife* **2014**:1–21. doi:10.7554/eLife.02030

Procaccini A, Lunt B, Szurmant H, Hwa T, Weigt M. 2011. Dissecting the Specificity of Protein-Protein Interaction in Bacterial Two-Component Signaling: Orphans and Crosstalks. *PLoS One* **6**:e19729. doi:10.1371/journal.pone.0019729

Qian W, He X, Chan E, Xu H, Zhang J. 2011. Measuring the evolutionary rate of protein-protein interaction. *Proc Natl Acad Sci U S A* **108**:8725–30. doi:10.1073/pnas.1104695108

Ravera E, Ciambellotti S, Cerofolini L, Martelli T, Kozyreva T, Bernacchioni C, Giuntini S, Fragai M, Turano P, Luchinat C. 2016. Solid-State NMR of PEGylated Proteins. *Angew Chemie Int Ed* **55**:2446–2449. doi:10.1002/anie.201510148

Rodrigues JPGLM, Trellet M, Schmitz C, Kastritis P, Karaca E, Melquiond ASJ, Bonvin AMJJ. 2012. Clustering biomolecular complexes by residue contacts similarity. *Proteins Struct Funct Bioinforma* **80**:1810–1817. doi:10.1002/prot.24078

Rodriguez-Rivas J, Marsili S, Juan D, Valencia A. 2016. Conservation of coevolving protein interfaces bridges prokaryote–eukaryote homologies in the twilight zone. *Proc Natl Acad Sci* **113**:15018–15023. doi:10.1073/pnas.1611861114

Rose PW, Prlić A, Bi C, Bluhm WF, Christie CH, Dutta S, Green RK, Goodsell DS, Westbrook JD, Woo J, Young J, Zardecki C, Berman HM, Bourne PE, Burley SK. 2015. The RCSB Protein Data Bank: views of structural biology for basic and applied research and education. *Nucleic Acids Res* **43**:D345-56. doi:10.1093/nar/gku1214

Salinas VH, Ranganathan R. 2018. Coevolution-based inference of amino acid interactions underlying protein function. *Elife* **7**. doi:10.7554/eLife.34300

Schug A, Weigt M, Onuchic JN, Hwa T, Szurmant H. 2009. High-resolution protein complexes from integrating genomic information with molecular simulation. *Proc Natl Acad Sci U S A* **106**:22124–9. doi:10.1073/pnas.0912100106

Skwark MJ, Elofsson A. 2013. PconsD: ultra rapid, accurate model quality assessment for protein structure prediction. *Bioinformatics* **29**:1817–1818. doi:10.1093/bioinformatics/btt272

Sun MGF, Kim PM. 2011. Evolution of biological interaction networks: from models to real data. *Genome Biol* **12**:235. doi:10.1186/gb-2011-12-12-235

Sutto L, Marsili S, Valencia A, Gervasio FL. 2015. From residue coevolution to protein conformational ensembles and functional dynamics. *Proc Natl Acad Sci* **112**:13567–13572. doi:10.1073/pnas.1508584112

Szurmant H, Weigt M. 2018. Inter-residue, inter-protein and inter-family coevolution: bridging the scales. *Curr Opin Struct Biol* **50**:26–32. doi:10.1016/J.SBI.2017.10.014

Takegoshi K, Nakamura S, Terao T, Nakamura S. 2001. 13C–1H dipolar-assisted rotational resonance in magic-angle spinning NMR. *Chem Phys Lett* **344**:631–637. doi:10.1016/S0009-2614(01)00791-6

Tang Y, Huang YJ, Hopf TA, Sander C, Marks DS, Montelione GT. 2015. Protein structure determination by combining sparse NMR data with evolutionary couplings. *Nat Methods* **12**:751–754. doi:10.1038/nmeth.3455

Traaseth NJ, Verardi R, Veglia G. 2008. Asymmetric methyl group labeling as a probe of membrane protein homo-oligomers by NMR spectroscopy. *J Am Chem Soc* **130**:2400–2401. doi:10.1021/ja711499r

Uguzzoni G, John Lovis S, Oteri F, Schug A, Szurmant H, Weigt M. 2017. Large-scale identification of coevolution signals across homo-oligomeric protein interfaces by direct coupling analysis. *Proc Natl Acad Sci* **114**:E2662–E2671. doi:10.1073/pnas.1615068114

Wang S, Sun S, Li Z, Zhang R, Xu J. 2017. Accurate De Novo Prediction of Protein Contact Map by Ultra-Deep Learning Model. *PLoS Comput Biol* **13**:e1005324. doi:10.1371/journal.pcbi.1005324

Weigt M, White RA, Szurmant H, Hoch JA, Hwa T, White RA, Szurmant H, Hoch JA. 2008. Identification of direct residue contacts in protein-protein interaction by message passing. *Proc Natl Acad Sci* **106**:67–72. doi:10.1073/pnas.0805923106

Xu J, Zhang R, Wang S, Li W, Liu S. 2016. CoinFold: a web server for protein contact prediction and contact-assisted protein folding. *Nucleic Acids Res* **44**:W361–W366. doi:10.1093/nar/gkw307

Yang Li, Chengxin Zhang, Dongjun Yu YZ. 2018. Contact Prediction by Stacked Fully Convolutional Residual Neural Network Using Coevolution Features from Deep Multiple Sequence Alignment. *CASP13 Abstr B* 154.

SUPPLEMENTARY INFORMATION

The python script to perform the protocol can be downloaded at the following **LINK**.

Supplementary figures

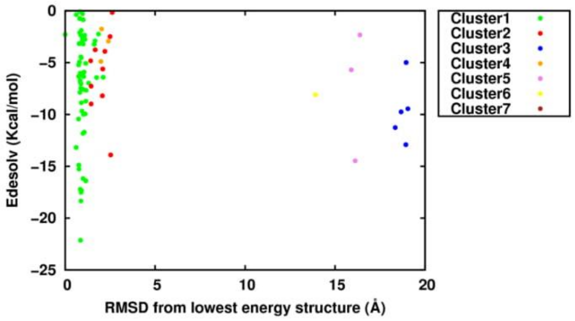


Figure S1. Cluster distribution based on the desolvation energy in the first docking run of L-asparaginase II. The colors of the clusters are the same as in Figure 1.

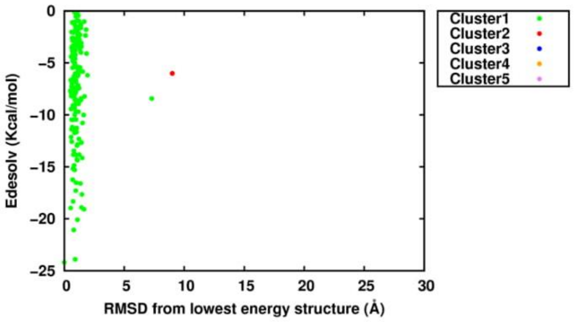


Figure S2. Cluster distribution based on the desolvation energy in the second docking run of L-asparaginase II.

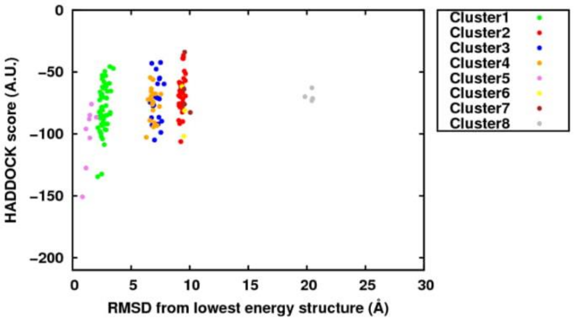


Figure S3 L-asparaginase II clusters distribution obtained from a monomer-monomer docking run performed using the AIRs violated in the third cluster of the first run

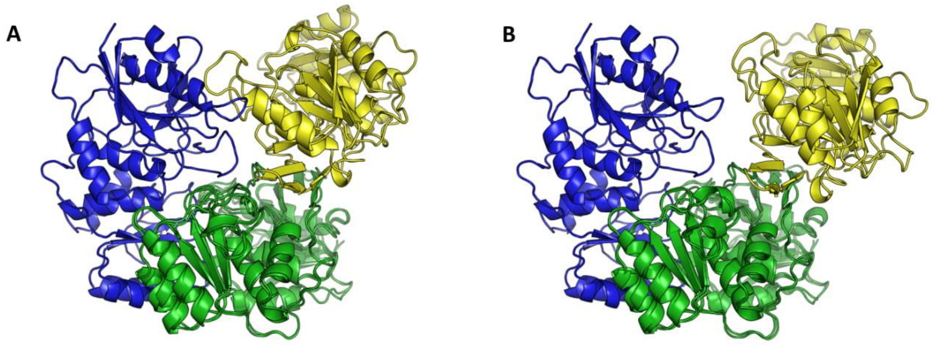


Figure S4. Superimposition on the chain A (in green) of the best L-asparaginase II models. **A**) Model 13 AC dimer is in blue and model 6 AD dimer in yellow. **B**) Mode13 AC dimer is in blue and model 15 AD dimer in yellow.

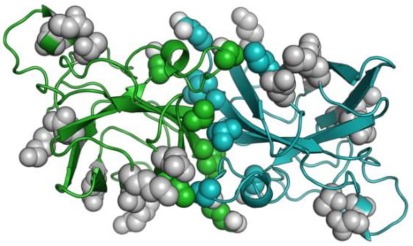


Figure S5. Residues used as AIRS in the docking run of Sod1. Residues forming contacts across the interface are colored as the backbone.

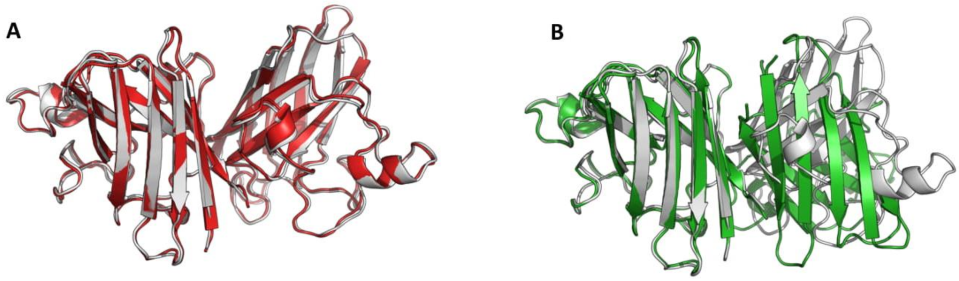


Figure S6. Fitting of the best model of the clusters 1 and 2 on the Sod1 crystal structure . **A**) cluster 2 in red. **B**) cluster 1 in green

2.4 Conclusions

In conclusion, my Ph.D. project contributed in an important aspect of metalloproteins: development of methods to the prediction/determination of (metallo)proteins structure. All of these studies were performed with the contribution of computational methods. In particular, experimental NMR-based data were integrated in theoretical calculations to develop new protocols for the prediction of proteins structure or complex and the refinement of NMR bundles.

In this chapter, we started implementing a rMD protocol that allows NMR structural biologists to simultaneously use PCS restraints derived from the substitution of multiple lanthanide ions in the same site of diamagnetic metals to perform the energetic refinement of structures. This refinement yielded an improvement of structural quality parameters comparable to that of the standard AMPS-NMR protocol for diamagnetic proteins. Importantly, this new protocol allowed structures generated without inclusion of PCS data among the restraints to be successfully refined including the PCS restraints only at the MD stage. This resulted in a satisfactory agreement with all experimental restraints and high stereochemical quality. The refined bundles were of the typical quality for NMR structures deposited in the PDB. This tool exploits the AMBER package for molecular dynamics simulations and is incorporated in the AMPS-NMR portal for NMR structure refinement provided by the WeNMR electronic infrastructure. It complements the available portfolio of software tools for the use of paramagnetic restraints in protein structure determination.

In the CASP13 project, the assessment of the models generated by competitors revealed general higher values of local RMSD on missing data regions. Thus, the presence of data covering most of the target regions positively affected the structural quality of the generated models. Modeling methods that used NMR data generally improved accuracy of prediction over modeling methods used by the same groups without NMR data. However, for most of the targets the best model generated with and without experimental data were similar in term of overall structural quality, i.e. equally distant from the crystal structure. Therefore, the constant raising of computational power and accuracy of prediction methods are close to cover the gap with data-driven approaches, at least for small proteins structure determination.

Solid State NMR can provide ambiguous contacts for large protein assemblies. However, the determination of their 3D structure involves two very time-consuming steps: the assignment of the side chains in contact at the interface between the subunits and, for homo-oligomeric complexes, the discrimination of intra- vs inter-monomeric contacts. In particular, the correct identification of inter-monomeric contacts usually requires extensive efforts by an experienced user. From the bioinformatics point of view, focusing on homo- rather than hetero-oligomers makes the interpretation of coevolution signals harder. In fact, the difficult step in the coevolution analysis of hetero-oligomers is the proper pairing of orthologs of interacting proteins and the corresponding removal of paralogs, that lead to the creation of joint MSA in which each line contains a pair of interacting proteins. Instead, the coevolution analysis of homo-oligomers is based on a single protein MSA, which is relatively effortless to build. At the moment, the availability of the three-dimensional structure of the monomeric unit is necessary to successfully separate intra-monomeric and inter-monomeric ECs. In this work, we developed a protocol to automatically identify residues belonging to the homo-oligomeric interface by matching predicted inter-monomeric ECs with NMR ambiguous contacts list automatically generated from appropriate solution or solid-state NMR spectra.

Chapter 3

ROLE OF METAL IONS IN METALLOPROTEIN DYNAMICS

3.1 Introduction

3.1.1 Metalloproteins Folding

The constant growing in computational power has paved the way to the investigation of long timescale biological processes using *in silico* methods. In particular, modern cost-effective GPGPUs allow exploration of atomic motions through molecular dynamics (MD) methods. MD simulations are particularly useful in the investigation of processes that are difficult to capture with experimental methods, such as protein folding, protein transient conformations, metal ions mobility and others[80]. Among them, folding mechanism occurs when an unfolded protein chain starts to evolve into its stable and functional three-dimensional structure. This can happen in different timescales, from microseconds to seconds and higher, depending on protein features. Despite advances in computational power, in force fields accuracy and MD methods, the exhaustive sampling of protein folding mechanism by MD simulations remains a challenging task[81]. In this regard, the simulation length to observe at least a single folding event or to achieve enough folding/unfolding transitions to define the folding pathway and accurately measure kinetic and thermodynamic quantities is still demanding[82]. In addition, studying the unfolded states is one of the frontiers for all-atom simulations[83]. Alternatively, information on folded and unfolded states can be extracted generating equilibrium trajectory ensembles from independent simulations[84]. Interestingly, the finding of fast-folding proteins has made possible the direct comparison of experimental and computational folding timescales on these systems[85]. As a result, the integration of experimental and computational studies contributed significantly to our knowledge of the pathway(s), thermodynamics and kinetics of folding[86].

Although metalloproteins are involved in many crucial physiological processes, metal-induced folding has been poorly investigated from theoretical point of view. The main reason is the complexity of the metal treatment in model metal-protein interactions[87]. Some of the crucial features of metal binding interactions that are much difficult to reproduce in classical MD simulations are metal induced protonation/deprotonation, multiscale coupling, charge transfer and induced polarization. Some force fields able to accurately model part of the mentioned metal-induced interactions have been developed[22]. However, their applicability on long timescale processes is difficult because of the extensive computational cost.

One of the key question in metalloprotein folding is how the presence of metal affects protein folding in terms of structural and dynamics properties[88]. In this regard, MD simulations of small metalloproteins is a cost-effective way that can be exploited to partially answer this question. However, in most of the cases classical MD simulations are not suitable to study the folding process of proteins due to the long simulation time required that usually must be even larger than the experimental timescale. In fact, a well-known problem in applying classical MD for folding simulations relies on high-energy barriers of proteins that separate minima of the potential energy surface (PES). Consequently, the protein is often trapped in local minima for long time, preventing exploration of the PES and reducing the possibility to sample the native folded state starting from the unfolded protein. This problem can be overcome by applying enhanced sampling methods that greatly boost conformational sampling. Among them, the accelerated MD (aMD) method exploits a bias potential to reduce the height of local barriers and proved to be effective in enhancing the PES exploration[89].

3.1.2 Ferritin

Ferritin is a superfamily of protein-caged Fe_2O_3•H_2O biominerals that plays a key role in the storage and transport of iron ions[90]. Ferritin function is performed in two main steps, 1. Fe^{2+} ions entry and oxidation, with the following nucleation and mineral growth that has a protective effect from metal toxicity and 2. mineral reduction/dissolution with Fe^{2+} ions release from the protein cage into iron-containing systems. Ferritin stores and transports iron in both intra- and extra-cellular environment where it can bind different cell types and receptors. The T cell immunoglobulin and mucin domain-2 (TIM-2) receptor binds human ferritin, allowing for its entry into endosomes where the pH is in the 6.0–4.8 range[91]. Ferritin uptake into endosomes is also triggered by the binding to Transferrin Receptor 1 (TfR1) on the cell surface followed by cellular internalization via endocytosis[92].

Most members of the ferritin superfamily consist of 24 subunits arranged to form a hollow structure able to store up to 4300 iron ions (Figure 6). The highly symmetric 24-mer structure spontaneously self-assembles from the four-helix bundle subunits resulting in a protein with an octahedral symmetry and an overall molecular weight of about 480 kDa. In mammalian, ferritin is mainly composed of H (FtH, heavy 21 ~ kDa) and L subunits L (FtL, light, ~ 19kDa), that have a distinct function, plus the mitochondrial one (FtMt, about 21 kDa) exclusively found in mitochondrial ferritin. Despite the high level of sequence similarity, only FtH harbours the oxidoreductase site responsible of ferroxidase activity, converting Fe^{3+} to Fe^{2+}, a step crucial for iron biomineralization. Instead, the FtL chain contains the nucleation site that can bind Fe^{3+} helping the mineralization process[93]. The formation of the mineral protects cell from normal or pathological oxidative damage that may arise from reactions of iron with oxygen. Differently from the homopolymeric mitochondrial ferritin, the ratio of FtH and FtL subunits in cytosolic ferritin is tissue-specific and can also vary under pathological conditions[94]. Since the two subunits are not interchangeable, the correct balancing of the subunit ratio is critical. Ferritin mutations, disfunction and misregulation have been associated with several diseases[95].

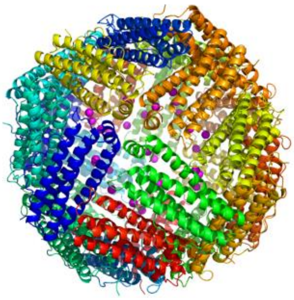


Figure 6. 24-mer ferritin with iron ions in the protein cage.

The 24-mer ferritin structure has two different channels that connect bulk solution to the protein cavity. The two channels have also different chemical properties related to their aminoacidic composition. The hydrophobic C4 channels assemble around the 4-fold symmetry axis and have a role in in the diffusion of dioxygen and hydrogen peroxide[96]. Instead, the hydrophilic and negatively charged C3 channels assemble around the three-fold symmetry axes of the ferritin and are responsible of iron ions mobility in and out from the protein cavity during entry process for mineral formation and during exit after mineral dissolution, respectively[97]. During the entry process, iron ions transit under the form of hexahydrate ions that is attracted by the electrostatic gradient in turn generated by carboxylate groups at the inner end of the C3 channels. Instead, the pH dependence of the iron removal on human H-ferritin has been measured *in vitro* adding a reducing agent in the presence of chelators, a step required to reduce the oxidized Fe^{3+} to the soluble Fe^{2+}, showing that lowering the pH from basic to acidic values increases the amount of iron ions discharged[98]. *In vivo* the process is associated to cage disassembly due to lysosome degradation but also the proteasomal pathway might be involved[95,99]. Interestingly, experimental results have shown the crucial role of the acidic environment for iron extraction in both physiological and pathological conditions[100].

3.1.3 YiiP

Zinc has cytotoxic effects for intracellular concentration higher than few hundred picomolar. Thus, zinc uptake and efflux from cytosol is tightly controlled by specific systems. In mammalians, zinc removal from the cytoplasm into the extracellular space or into the intracellular organelles is mediated by the ZnTs/Slc30 family belonging to the cation diffusion facilitator (CDF) superfamily[101]. At present, all the CDF members known are featured by an antiport mechanism to export divalent cations. These proteins contain an N-terminal domain, a transmembrane domain formed by six helices (TMD) and a long C-terminal domain (CTD). However, an experimental 3D structure for any of the ZnT family proteins is missing so far. This absence is partially compensated by the well-characterized bacterial homolog protein from *Escherichia coli*, the YiiP transporter. YiiP is a cation-proton antiporter that couples a cation ion efflux with a proton influx in a 1:1 exchange stoichiometry. YiiP can transport a broad range of metal cations, but only zinc and cadmium with high efficiency[102]. Its architecture in the crystallographic outward-facing (OF) conformation (PDB 3H90) shows a homodimer with a Y-shaped structure composed by six transmembrane (TM) helices for each monomer[103] (Figure 7A). The six TM helices can be grouped in two separated subdomains, a four-helix bundle (TM1-TM2-TM4-TM5) and a helix pair (TM3-TM6). The TM3-TM6 helix pair protrudes in the cytosolic region in an antiparallel configuration providing a dimeric interface stabilized by four interlocking salt bridges formed between K77 of TM3 and D207 of IL3 close to TM6. Each protein chain has three conserved zinc-binding sites (A-C). The transport site (A) is located in the middle of TMD and is the active site for zinc transport. The tetrahedral coordination geometry adopted in the zinc transport site is made of four highly conserved residues, D45 and D49 from TM2 and H153 and D157 from TM5. These residues are responsible for zinc and cadmium selectivity of the transporter. The binding site B is located in the cytoplasmic loop connecting TM2 and TM3 and is formed by D68, H71 and H75. However, its function is still unknown. Finally, the binding site C is located at the CTD-CTD interface. It harbors two zinc ions for each monomer with very high affinity and has a stabilizing effect on the dimer.

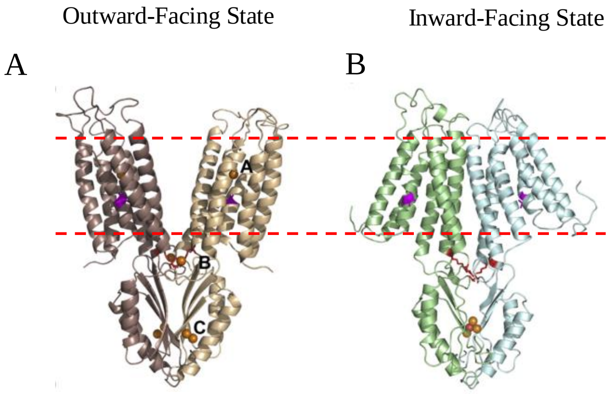


Figure 7. YiiP protein in the outward-facing and inward-facing states[101].

Successively, the cryoelectron microscopy structure of the YiiP homolog from *Shewanella oneidensis* was solved in the inward-facing (IF) conformation and lipid environment at low resolution (PDB 3J1Z)[104] (Figure 7B). Recently, the resolution of the IF conformation structure was improved again with Cryo-EM[105]. It has been suggested that this conformation allows the zinc permeation in the TM cavity from the cytoplasmic side and the following zinc binding to the transport site. In particular, the zinc binding on site C could trigger an allosteric mechanism through which the TMD rearranges to allow the zinc binding on site A. This mechanism connecting CTD to TMD could be mediated by the aforementioned charge interlock and a reorientation of the TM3-TM6 helix pair. The zinc binding to the transport site is followed by the ion release. However, how this process happens is not totally clear. A proposed mechanism relies on the switch from the IF to the OF conformation as a necessary step to discharge the metal ion in the extracellular space or other cellular compartments. A variation of this mechanism has been proposed in which to discharge zinc ions the OF conformation is not strictly necessary but the critical step is a rearrangement of the four-helix bundle against a static TM3-TM6 helix pair[105]. In practice, a persistent dimeric interface acts as a static "rock" versus a mobile "bundle" that modulate the ions permeation and release, in a so-called rocking-bundle mechanism common also in other secondary active transporters[106]. In this regard, the modulation of the zinc accessibility in the TM cavity harboring the zinc transport site was measured through x-ray-mediated hydroxyl radical labelling and mass spectrometry[107]. The measurements revealed the key role played by TM5 of the four-helix bundle in the modulation of the water accessibility from the cytosol to the TMD through a rigid-body motion that puts in contact the L152 and M197 sidechains eventually forming a hydrophobic gate that prevents the access to the transport site.

3.1.4 ZnT8

Levels of zinc in different cellular compartments are controlled by metallothioneins and two zinc transporter families, ZIP and ZnTs, which members are expressed in many tissues. Among them, the mammalian subgroup of four transporters ZnT2, 3, 4 and 8 are responsible of the zinc ions removal from the cytosol, thus raising zinc concentration in vesicles[17]. In particular, ZnT8 is highly expressed in the membrane of insulin secretory granules of pancreatic beta cells[108]. ZnT8 supplies zinc in the secretory granules facilitating insulin storage and maturation. Inside granules zinc is crystallized with insulin to form hexameric crystals. ZnT8 is expressed also in some other tissues, for instance in pancreatic α-cells where is supposed to play a role in glucagon secretion[109]. Interestingly, a single nucleotide polymorphism (SNP) in the ZnT8 gene SLC20A8 encodes two major variants, an arginine or a tryptophan at position 325, located in the CTD, that are associated with high and low risk to develop type-2 diabetes (T2D), respectively[110]. Despite both variants are widespread in population, the higher risk R325 variant is more frequent in humans (>50%). Remarkably, the same variant is also associate with a strong zinc transport activity compared to the loss-of-function W325 variant that is associated with a protective effect[111]. Unfortunately, there is no 3D structure for any of the eukaryotic ZnT transporters. Moreover, homology models based on the bacterial functional homologue YiiP protein revealed a mutation localized in a position not directly involved in the zinc binding on the CTD and very distant to the transport site, thus making the postulation of how the two ZnT8 variants work an arduous task. Recently, biophysical characterization of the ZnT8 CTD suggested a higher zinc affinity and lower thermostability of the W325 variant[112]. Furthermore, both eukaryotic CTD variants exhibited a different zinc-binding stoichiometry from bacterial homologue, highlighting the limitation of the latter as template for model generation of the human proteins.

3.2 Research Projects

Metal-coupled folding is a biological process particularly hard to investigate with MD, not only for the long timescale of the process but also for the limited accuracy of classical force fields in the description of metal-binding effects. However, one of the smallest metalloprotein suitable for a folding study is rubredoxin from *Pyrococcus furiosus* (PfRd). PfRd is a globular protein of 53 amino acids that binds a single iron ion with a tetrahedral coordination. After denaturation, PfRd can refold both in presence and absence of the metal cofactor. The holo-PfRd folding properties were investigated performing a classical MD (F-cMD) simulation and an accelerated (F-aMD) simulation by which an extensive conformational sampling can be reached. A third simulation in the folded apo-form (APO) in addition to experimental data were used to assess the calculated conformational ensembles in term of structural similarity with the native folded state. The crystal structure of the holo-PfRd (PDB 1BRF) was used as starting point for all the MD runs. The RESP charges of the metal-binding site compatible with the Amber14SB force field used in the calculations were taken from the literature. The unfolded conformation needed for the starting point of the refolding simulations was derived from a brief MD at 600K.

As expected, the APO simulation confirmed the stability of the PfRd protein even without the bound metal. In fact, the RMSD from the crystal structure was stably low for the whole simulation. However, the secondary structures showed little rearrangements, especially around the metal-binding site. The gained degrees of freedom due to the iron ion removal was in fact compensated by the formation of up to three non-native β-sheets. The secondary structures detected in the F-aMD conformational ensemble generally mapped with good accuracy the ones present on the crystal structure. The main discrepancies from the folded state were observed in the overestimated occurrence of helical structure elements and in the formation of some non-native β-structures as seen for the APO run. Moreover, the F-aMD simulation was not able to recover the long-range contacts needed to make the triple β-sheet, main responsible of the protein globularity. The NMR chemical shifts of the backbone nuclei of holo-PfRd provided and experimental reference for the comparison of the protein dynamics sampled in our simulations. As expected, the APO simulation showed good agreement with the experimental data, confirming the lack of significant rearrangements. Instead, the other two folding simulations featured modest correlations for the C and N atoms, probably due to a heterogeneous conformational ensemble. However, the extensive sampling of the F-aMD trajectory was able to produce a significant number of conformations with a compact shape relatively similar to the native structure. This statement was supported by the temporal correlation of the lowest RMSD from the crystal with the lowest solvation (SASA analysis) and the highest number of hydrophobic contacts detected among the core residues. In particular, the values sampled were close to the ones observed in the folded APO simulation, with the exception of the RMSD from the crystal structure that never lowered below 6-7 Å. Thus, the sampling achieved in the F-aMD simulation showed at least some transition/intermediate conformations toward the folded state.

To shed light on the molecular mechanism of Fe^{2+} ions release from human H-ferritin nanocage, extensive MD simulations were performed in both acidic and basic conditions. The 24-mer human ferritin model (4134 residues) was built based on the crystal structure of the bullfrog M ferritin (PDB 4DAS). Using H++ server the protonation state of titratable groups was calculated at pH 4 and pH 9. Subsequently, 31 Fe^{2+} ions were added inside the protein's cavity. The two systems were prepared independently using the customized forcefield AMBER99SB-ILDN* on the AMBER16 MD Package. This forcefield merges the ILDN correction of AMBER99-ILDN ff with the CMAP dihedral parameters improvement of CHARMM22* ff. The same NPT condition was applied in both the systems for a simulation time length of 105 ns each.

MD simulations provided a detailed atomic-level view of the mechanism of Fe^{2+} release at pH 4, which occurred through the C3 channels. Within each individual channel, two nearby rings formed by symmetry-related Asp and Glu sidechains define the binding site to which iron ions were rapidly attracted from the internal cavity by the electrostatic gradient. The negatively charged cluster of carboxylates in C3 channels was surrounded by a number of histidine residues with a different protonation state depending on the pH of the simulation. In this regard, the Fe^{2+} ions diffused in the bulk solution only at pH 4, whereas at pH 9 they were immobilized inside the C3 channels by the negative charges. Therefore, a key role in the release process was played by the positive charges of histidine side chains that partially compensated the negative electrostatic charge of the carboxylates. The free energy of the processes sampled by the two unbiased simulations were then measured with the umbrella sampling (US) method. In accord to what seen in the unrestrained simulations, at pH 4 the free-energy profile showed a significantly lower energy barrier than at pH 9. Furthermore, the positions of the energetic minima corresponding to the Fe^{2+} binding site within the channel at pH 4 was significantly closer to the end of the channel, denoting a higher permeability of the channel in acidic than in basic conditions. This hypothesis was tested building a chimeric ferritin in which the His residues inside the C3 channels were doubly protonated as at pH4. As a result of the US run, the mutations caused a reduction of the energetic barrier of Fe^{2+} release but did not change the position of the energetic minimum along the C3 channel. Therefore, the protonation state of the His residues inside the channel is a necessary but not sufficient factor for Fe^{2+} release. In fact, it is only at pH 4, where all the His of ferritin are doubly protonated, that the dynamic opening of the binding site could take place, resulting in a significant shift of the Fe^{2+} coordination site towards the exit of the channel with respect to pH 9.

In a following unpublished work we investigated arachidonic acid's (ARA) contribution to the coordination of iron ions in the mineral nucleation site, situated in the L-chain of horse spleen ferritin (HoSF, PDB 4DE6). The influence of the negatively charged ARA on the mobility of iron ions inside the cage of HoSF was analyzed building the protein model both with and without bound ARA. For each model, four replicas were performed in which the number of ions in the protein cage was increased as follow: 12, 22, 31 and 62. The iron ions were treated in their Fe^{2+} form for concordance to the data acquired by *in vivo* experiments.

The first relevant result regards the stability of the protein in presence of 62 Fe^{2+} ions in the cage, measured as RMSD from the crystal structure. With this significant number of Fe^{2+} ions, ferritin reached stability, i.e. RMSD plateau, only with the bound ARA. On the contrary, the protein without ARA never achieved convergence. Thus, ARA positively contributed to the protein stability in presence of high amount of Fe^{2+} ions. We then used the radial function to evaluate how many Fe^{2+} ions were present in every mineral nucleation site for each simulation. The origin of the function was the centre of the four oxygens of the two glutamate side chains in the nucleation sites. The presence of ARA led to a preference for the peaks closer to the origin in the simulations with 12, 22 and 31 Fe^{2+} ions, but not in the simulation with 62 Fe^{2+} ions (Figure 8).

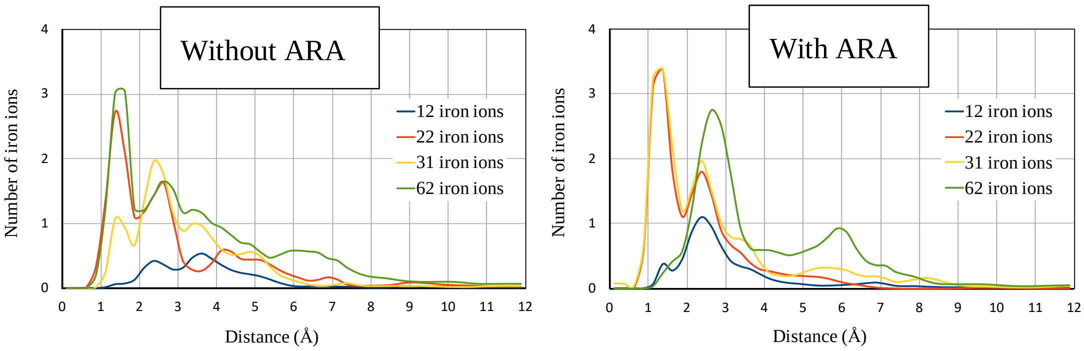


Figure 8. Radial function of iron ions in the nucleation sites.

This might be due to an overpopulation of Fe^{2+} ions in the nucleation site that enabled the possibility of both glutamates to interact with the same metal ion. In this regard, the steric effect and the repulsion of the negatively charged ARA prevented the formation of a “bridge” coordination, instead present in the simulations without ARA, in which two glutamates (Glu56 and Glu60) of two adjacent nucleation sites interacted with the same metal ion (Figure 9). Eventually, ARA reduced the number of possible conformations that the glutamates of mineral nucleation sites can otherwise take to interact with irons.

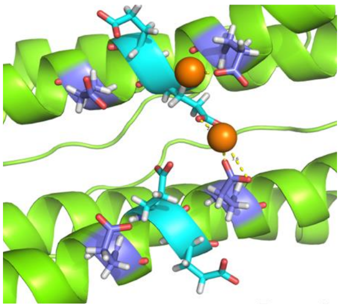


Figure 9. "Bridge" coordination in the nucleation site of MD simulations without ARA.

The zinc-induced YiiP dynamics was investigated performing MD simulations of the IF conformation in presence of Zn^{2+} ions in solution. First, the model of the YiiP inward-facing state was built combining the CTD of the crystal structure (PDB 3H90; 2.9 Å) with the TMD of the Cryo-EM structure (PDB 3J1Z; 13 Å). Then, the transporter was embedded in a DPPC bilayer. Two NVT MD simulations in 0.150 M of $ZnCl_2$ and one in the apo-form were performed for a length of 1 µs each. Finally, two umbrella sampling simulations were performed to measure the free-energy profiles of Zn^{2+} and Na^+ ions going through the TM channel.

The holo-simulations carried out on the YiiP transporter elucidated the Zn^{2+} interaction pathway from the cytosol to the transport site. The Zn^{2+} permeation was preceded by a quite fast disjunction of the four-helix bundle (in particular TM5 and TM4) from the TM3-TM6 helix pair that opened the TM cavity from the IC space allowing one Zn^{2+} ion to reach D49 and H153 of the transport site. Subsequently, a reorientation of TM5 putted in direct contact the side chains of L152 and M197, forming a hydrophobic gate that dropped the water presence in the TM cavity, thus making the channel inaccessible from the cytosol. Moreover, the CTD motion could affect the TMD region tilting and rotating. In fact, the two protein channels had always an alternating accessible/inaccessible state and the CTD was always tilted toward the inaccessible channel. Finally, the free-energy profiles calculated performing two umbrella simulations indicated that both the Zn^{2+} and Na^{+} ions can reach the transport site without encountering high energetic barriers. Conversely, from the transport site to the EC space only Na^{+} ion had a low-energy profile.

Despite the relevance of this transporter, the structure of ZnT8 is still unknow as well as that of eukaryotic homologs. Taking advantage of coevolution analysis and the template structure of a prokaryotic functional homolog YiiP, a model of the dimeric ZnT8 transporter was built. The interatomic distances of the monomeric subunit were calculated using the Gremlin webserver. The multiple sequence alignment was generated with Jackhmmer. The resulting structure was then superimposed to the dimeric YiiP IF structure (PDB 5VRF). The following structure was embedded in a 2:1:1 DOPC:DOPE:DOPG bilayer already proven as the lipid composition able to maximize the zinc transport in vitro[113]. Then, by performing all-atom and coarse-grained MD simulations putative conformational dynamics discrepancies between the active and the loss-of-function ZnT8 variants are under investigation. To date, an all-atom simulation of 2 μs in presence of zinc ions has been performed for each variant (R325 and W325). The recent ff14SB and lipid17 force fields were used for protein and lipidic membrane, respectively. The preliminary analyses performed on the two ZnT8 all-atom simulations pointed out different dynamics of the two transporter variants. In particular, the protective variant W325 showed a cytosolic domain (CTD) tilted toward one of the two channels, blocking its ion permeability (Figure 10). On the contrary, the active R325 variant showed a CTD rotating along its y-axis that allows zinc ions entrance in both the channels by an alternating mechanism.

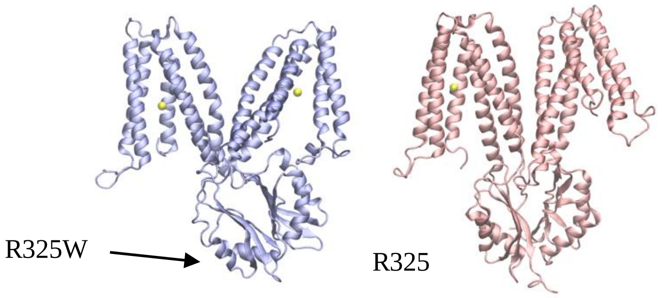


Figure 10. Snapshots of the ZnT8 conformation in the simulations of the R325 (active) and R325W (loss-of-function) variants.

3.3 Results

3.3.1 Molecular dynamics simulations of metalloproteins: a folding study of rubredoxin from *Pyrococcus furiosus*

*Davide Sala[1], Andrea Giachetti[3] and Antonio Rosato[1, 2, *]*

[1]Magnetic Resonance Center (CERM) – University of Florence, Via L. Sacconi 6, 50019 Sesto Fiorentino, Italy

[2]Department of Chemistry – University of Florence, Via della Lastruccia 3, 50019 Sesto Fiorentino, Italy

[3]Consorzio Interuniversitario di Risonanze Magnetiche di Metallo Proteine, Via Luigi Sacconi 6, 50019 Sesto Fiorentino, Italy.

AIMS Biophysics, 2018, 5(1): 77-96.

Biophysics

http://www.aimspress.com/journal/biophysics

AIMS Biophysics, 5(1): 77–96.
DOI: 10.3934/biophy.2018.1.77
Received: 19 January 2017
Accepted: 12 March 2018
Published: 22 March 2018

Research article

Molecular dynamics simulations of metalloproteins: A folding study of rubredoxin from *Pyrococcus furiosus*

Davide Sala[1], Andrea Giachetti[2] and Antonio Rosato[1,3,]*

[1] Magnetic Resonance Center (CERM), University of Florence, Via Luigi Sacconi 6, 50019 Sesto Fiorentino, Italy
[2] Interuniversity Consortium of Magnetic Resonance of Metallo Proteins (CIRMMP), Via Luigi Sacconi 6, 50019 Sesto Fiorentino, Italy
[3] Department of Chemistry, University of Florence, Via della Lastruccia 3, 50019 Sesto Fiorentino, Italy

* **Correspondence:** Email: rosato@cerm.unifi.it; Tel: +390554574267.

Abstract: The constant increase of computational power has made feasible to investigate the folding mechanism of small proteins using molecular dynamics (MD). Metal-binding proteins (metalloproteins) are usually complicated to model, largely due to the presence of the metal cofactor. Thus, the study of metal-coupled folding is still challenging. In this work, we addressed the folding process of *Pyrococcus furiosus* rubredoxin (PfRd), a 53-residue protein binding a single iron ion, using different MD methods. Starting from an extended conformation of the polypeptide chain where we preserved the coordination of the metal ion, a classical MD simulation and an extensive accelerated MD run were performed to reconstruct the folding process of the metal-bound protein. For comparison, we simulated also the dynamics of folded PfRd devoid of the metal cofactor (apo-form), starting from the folded structure. For these MD trajectories, we computed various structural and biochemical properties. In addition, we took advantage of available experimental data to quantify the degree to which our simulations sampled conformations close to the native folded state. We observed that the compaction of the hydrophobic core is the main feature driving the folding of the structure. However, we could not reach a fully folded conformation within our trajectories, because of the incomplete removal of the solvent from the core. Altogether, the various

MD simulations, including that of the folded apo-form of the protein, suggest that an improvement in the accuracy of the protein force-field is still needed.

Keywords: folding; rubredoxin; molecular dynamics; metal; iron; metalloproteins; simulation; forcefield; modelling

Abbreviations: MD: Molecular Dynamics; aMD: accelerated MD; cMD: classical MD; IRED: Isotropic Reorientational Eigenmode Dynamics; PDB: Protein Data Bank; PfRd: *Pyrococcus furiosus* rubredoxin; RMSD: Root Mean Square Deviation; RMSF: Root Mean Square Fluctuation; SASA: Solvent Accessible Surface Area

1. Introduction

Constant advances in computational power and methods have opened the possibility to study many biological processes using *in silico* methods. In particular, the massive computing efficiency of modern GPGPUs can be exploited in molecular dynamics (MD) simulations to explore at the atomistic level timescales of motion that are relevant to functional properties [1,2]. One of the processes that in principle can be studied in a cost-effective way using MD is the folding mechanism [3]. Protein folding occurs when an unstructured polypeptide chain reaches its stable and functional three-dimensional structure. For different proteins, this can happen in a broad range of timescales from microseconds to seconds and higher. Thus, to obtain sufficient sampling to meaningfully comment on folding mechanisms by MD simulations remains a challenging task, notwithstanding the recent advances in computing power, MD methods and accuracy of force-fields [4]. In fact, the simulation length must be at least on the microsecond timescale to stand a good chance of observing a single folding event [5,6]. Achieving enough folding/unfolding transitions to define the folding pathway and accurately measure kinetic and thermodynamic quantities is even more demanding [7]. Alternatively, MD simulations for the folded and unfolded states can be carried out independently [8]. Despite the common difficulties mentioned above, the discovery of fast-folding proteins has provided systems for which the achievable MD timescales match the experimental folding times [9,10]. As a result, the combination of experimental and theoretical studies contributed significantly to our knowledge of the thermodynamics and kinetics of folding [11].

Despite the importance of metalloproteins in many important biological functions [12], little attention has been devoted to the theoretical investigation of metal-induced folding. This is mainly due to the difficulties to correctly model metal-protein interactions [13]. One of the unsolved questions regards how the presence of the metal affects the protein folding in terms of structural and dynamics properties [14]. In this work, we analyzed the folding mechanism of a highly stable metalloprotein: rubredoxin from *Pyrococcus furiosus* (PfRd). PfRd is a globular protein of 53 amino acids that binds a single iron ion in the +2 or +3 oxidation state [15]. The iron atom is coordinated to the protein through cysteinyl sulfurs of two consensus cysteine motifs, CXXC and CPXC. Various studies have addressed the structural basis of PfRd hyperthermostability as well as its ability to refold both in presence and absence of iron [16–19]. The secondary and tertiary structure of PfRd devoid of its iron cofactor (apo-PfRd) are very similar to iron-loaded PfRd (holo-PfRd) [20]. However, experimental studies of Holo- and apo-PfRd showed significant different in their unfolding and refolding processes. Apo-PfRd reaches about 50% unfolding at about 343 K, with the folded 50% of the molecules still displaying structural features consistent with the native structure of the holo-protein. The unfolding process is reversible [20]. On the other hand, the unfolding of holo-PfRd is essentially irreversible [21]. This has been ascribed to presence of the iron binding site; a designed variant of the protein that retains the tertiary structure and thermostability of wild-type PfRd but cannot bind iron features reversible folding [22]. Addition of iron to apo-PfRd in the presence of denaturing agents such as either 6 M urea or 6 M guanidine hydrochloride triggers protein refolding eventually yielding correctly folded holo-PfRd [19]. However, this does not happen if the denaturing agents are removed prior to addition of the metal. More recently, it has been suggested that these subtleties in the refolding process of apo- and holo-PfRd depend on the extent of structure loss upon removal of the iron ion from the holo-protein, which in turn depends on the amount of denaturant added. This is crucial to get the correct packing of the hydrophobic core residues during protein refolding [16].

In this work, we aimed at evaluating the folding properties of holo-PfRd. We applied two different MD methods: a classical MD (cMD) simulation in explicit solvent and an accelerated MD (aMD) simulation where a biased potential enhances the conformational sampling [23,24]. A well-known problem in applying cMD for folding simulations is that proteins usually have high barriers that separate the minima of the potential energy surface. Consequently, the system is often trapped in a local minimum for long periods of simulation time, preventing extensive exploration of the potential energy surface. Eventually, this may prevent reaching the native folded state of the protein starting far away from the global minimum. Even achieving a timescale comparable with the experimental data is not guarantee of success. To overcome these limitations, in this work we applied the aMD method, by which the height of local barriers is reduced, thus allowing the calculation to evolve much faster [25].

2. Materials and methods

This work is based on MD simulations performed using the Amber package of molecular simulation programs [26]. All the MD runs were prepared from the crystal structure of holo-PfRd (PBD 1BRF). The apo-form of the protein was obtained by removing the iron ion from the metal binding site. The holo-form has an iron ion covalently bound to 4 cysteine side chains. Thus, the RESP charges compatible with the AMBER force field for the metal center were used [27]. The systems were solvated with TIP3P water model molecules and the overall charge balanced. The forcefield used was Amber14SB. The starting structure of the folding simulations is derived from a brief MD at 600 K. More precisely, the most elongated structure was chosen as representative of the unfolded state. The MD runs were prepared following the same basic protocol consisting in 4 steps: water minimization, system heating to 300 K in NVT and density equilibration in NPT conditions. All the production runs were performed on a Nvidia Tesla K20m GPGPU applying a PME cutoff of

10 Å. This computational infrastructure is available to users via the AMPS-NMR portal within the West-Life (www.west-life.eu) project [28,29].

The first production run was performed on the folded state of the apo-form for 1 μs (APO). Then, a classical MD run was carried out for 1 μs to fold the protein (F-cMD). In addition, to increase the probability of sampling a folding event the accelerated molecular dynamics method was applied (F-aMD). Accelerated molecular dynamics was carried out starting from the same unfolded structure used for the cMD run. Differently form cMD, during the aMD run the whole potential is boosted as follows:

$$\Delta V_P = \frac{(E_P-(\Delta V_D))^2}{\alpha_P+(E_P-(V+\Delta V_D))} \tag{1}$$

where the torsion potential ΔV_D is given by

$$\Delta V_D = \frac{(E_D-V)^2}{\alpha_D+(E_D-V)} \tag{2}$$

The terms E_P and E_D define the average potential and dihedral energies. The terms α_P and α_D are the inverse strength boost factors for the total and dihedral potential energy, respectively. Prior to the production run a classical MD was carried out for 50 ns to define reasonable constants. As a result, the following values were applied: $E_P = -47.64$ kcal/mol, $E_D = 814$ kcal/mol, $\alpha_P = 2.542$ and $\alpha_D = 42$. The production run was performed for 11.6 μs.

The Root Mean Square Fluctuation (RMSF) is an index to measure the structural flexibility. It is defined by

$$RMSF = \sqrt{\frac{\sum_{t=1}^{\tau}\|r(t)-\langle r\rangle\|^2}{\tau}} \tag{3}$$

where τ is the number of frames, *r(t)* is the atomic position at time *t* and $\langle r\rangle$ is the average structure. Basically, this parameter corresponds to the standard deviation of the atomic positions from the average structure over a trajectory.

The order parameter (S^2) measures the magnitude of the angular fluctuation of a chemical bond vector such as the NH bond, thus reflecting the flexibility of a protein specific site. We used two different methods to calculate the order parameter of NH vectors. The isotropic reorientational eigenmode dynamics (iRED) method relies on a principal component analysis of the isotropically averaged covariance matrix [28]. In the second method, the autocorrelation function (ACF) of the vectors is fitted on a monoexponential curve [29].

The content of secondary structure along the trajectories was measured using the DSSP program [30,31]. Its dictionary consists of 8 classes: random-coil, 3-turn helix, α helix, 5-turn helix, turn, beta-sheet, beta-bridge and bend. In this work, the three helical topologies are indicated with "helix" and the beta-structures as "β-sheet".

The Root Mean Square Deviation (RMSD) is a conformational distance index. It is defined by

$$RMSD = \sqrt{\frac{\sum_{i=1}^{N} m_i\left\|r_i^{(a)}-r_i^{(b)}\right\|^2}{\sum_{i=1}^{N} m_i}} \tag{4}$$

where *N* are the atoms selected, r_i is the position vector and *a* and *b* are two different conformations. In this study, all RMSD values were referred to the crystal structure.

Chemical shifts were predicted using the PPM chemical shifts prediction web server (http://spin.ccic.ohio-state.edu) that was parametrized specifically for MD trajectories [32].

The Solvent Accessibility Surface Area (SASA) is a measure of the exposure to the solvent. For the common water, a probe of 1.4 Å is used to scan the molecular surface. In this study, the SASA is calculated on selected hydrophobic residues belonging to the protein core (residues Trp3, Tyr10, Tyr12, Phe29, Trp36, Phe48, Ile23 and Leu32).

For the present work, we defined "native contact" any contact shorter than a defined cut-off that is present in the reference structure. Only the contacts among core hydrophobic residues closer than 7 Å were considered.

The Principal Component Analysis was performed on the Cα atoms of the F-aMD and APO trajectories. The conformations were fitted on the crystal structure and separated on the average structure. The first two eigenvectors include 44% of the motions.

3. Results

PfRd has a globular shape, with the iron ion covalently bound to 4 cysteine side chains (residues 5, 8, 38 and 41) and is known to be quite rigid in its folded holo form, as shown by the low values of the B factor [33]. Its conserved secondary structure consists of three 3_{10}-helices (residues 19–21, 29–31 and 45–47) and one antiparallel triple stranded β-sheet (residues 2–5, 11–13 and 48–50). The hydrophobic core consists of 8 residues (residues Trp3, Tyr10, Tyr12, Phe29, Trp36, Phe48, Ile23 and Leu32). In this work, we attempted to simulate the folding process of holo-PfRd, starting from an extended structure. To this end, two different simulation schemes were adopted (Table 1). In addition, we simulated the dynamics of apo-PfRd; the starting model for the latter was the folded structure (PDB 1BRF) after removal of the metal ion.

Table 1. Main features of the MD simulations performed.

Type of simulation	Simulation length (μs)	Starting structure
Classical-Apo (APO)	1	Crystal structure without metal ion
Folding Accelerated (F-aMD)	11.6	Unfolded state with the metal bound
Folding Classical (F-cMD)	1	Unfolded state with the metal bound

The acronyms used in the text to identify each simulation are given in brackets in the first column. Both folding runs started from the same elongated conformation with the iron ion bound.

3.1. Simulation of folded apo-PfRd (APO)

The Root Mean Square Fluctuation (RMSF) analysis of the backbone atoms provides a measure of the flexibility of residues with respect to the average structure (Figure 1A). The higher the RMSF of one residue, the higher its mobility. In addition, protein motions can be summarized by calculating the order parameter (S^2) of N-H vectors [34,35]. Other bond vectors can also be used [36,37]. In the APO trajectory, we estimated the S^2 values of N-H vectors by two different methods: the IRED analysis [28] and a more traditional approach where the autocorrelation function (ACF) is fitted to a monoexponential curve [29] (Figure 1B). Despite the absence of the metal coordination by Cys5 and Cys8, the first 8 residues of the N-terminus are very rigid whereas Gly9 has higher flexibility than nearby residues in the RMSF profile. Residues 11 to 13 are quite rigid, and then the RMSF increases and reaches the maximum value in the region 20–21, whereas the order parameters values are still close to the overall protein average. From 22 to 29 the RMSF profile decreases progressively, but then increases again from residue 30 until the peak of the region 34–35. Then, the flexibility drops rapidly for the subsequent two residues. Differently from the RMSF, the order parameter profiles in this central region of the protein have more variability suggesting that the local environment affects the dynamics of the N-H vectors more than larger-scale motions. The absence of the covalently bound iron ion affects in a similar way both the RMSF and the order parameters in the region limited by the metal binding Cys38 and Cys41. In fact, all the analyses revealed high mobility in this region. The RMSF and S^2 profiles agree also in the last protein segment showing a modest flexibility from residue 42 to 51 followed by enhanced mobility for the last two residues of the C-terminus.

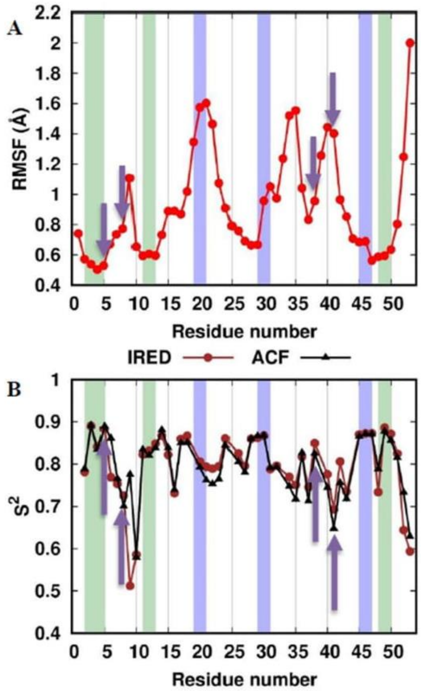


Figure 1. RMSF and order parameter profiles of the APO trajectory. The residues corresponding to β-sheets and 3_{10}-helices in the crystal structure are colored with a green and blue background, respectively. The metal-binding cysteines are indicated with yellow arrows. (A) RMSF of the backbone atoms. (B) The NH vector S^2 profiles computed with the IRED method (IRED, brown) and by calculating the autocorrelation function (ACF, black).

The apo-PfRd structure maintained its native (i.e. present in the crystal) secondary structure elements for most of the simulation (Figure 2). The Root Mean Square Deviation (RMSD) from the crystal structure of the backbone atoms shows a great stability along the trajectory, with an average value lower than 2.5 Å and only a modest increase in the second half of the simulation (Figure 3). The three-stranded β-sheet is present in almost the totality of the frames. On the contrary, the three helices are not as conserved as the triple strand. In particular, the first helix is present in just 23% of the snapshots against 58% and 67% of the second and third helix, respectively. A more detailed analysis revealed that most of the helix-missing frames sampled a turn structure featuring hydrogen bonds typical of helices (Table S1). Hence, just a very small percentage of the conformations were actually unstructured in these regions. Notably, six residues sampled non-native structures with a certain persistency. All of them formed β-strands involving the following couple of residues: 18–23, 38–42, and 37–44 present in the 98%, 42% and 91% of the conformers, respectively.

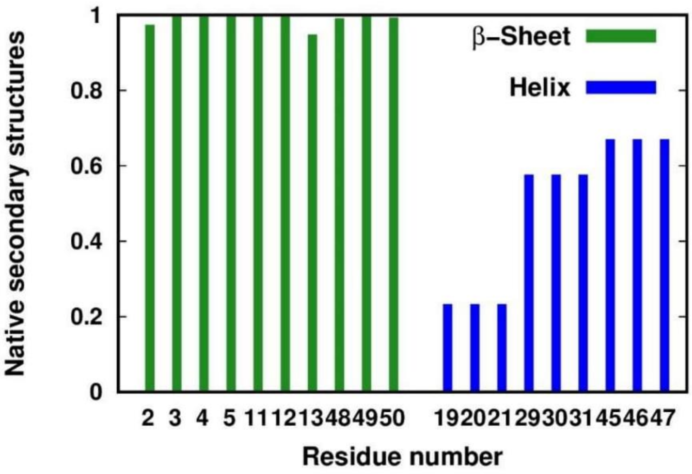


Figure 2. Fraction of native secondary structures sampled in the APO simulation. For each residue in β-sheet (green) or helix (blue) structure in the crystal structure, we report the fraction of conformations with the same secondary structure.

To evaluate the deviation of the APO conformational ensemble from the behavior of the holo-form, we used the NMR data of the latter [40–42]. We compared the chemical shifts of the Cα, C and N atoms computed from the trajectory with the available experimental data [38]. Figure 4 reports the correlation of the predicted values and the experimental values. The Cα atoms display a very high correlation, with a value of the Pearson coefficient as high as 0.98. The N and C atoms have Pearson coefficients of 0.87 and 0.73, respectively. In general, the correlation is good for all the atoms assessed, confirming that the apo-form samples conformations that are consistent with the holo-form.

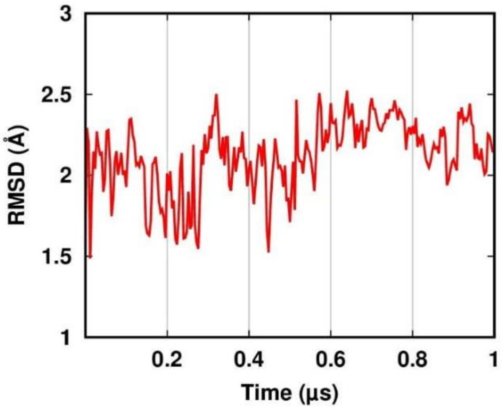


Figure 3. Root Mean Square Deviation from the crystal structure in the APO simulation.

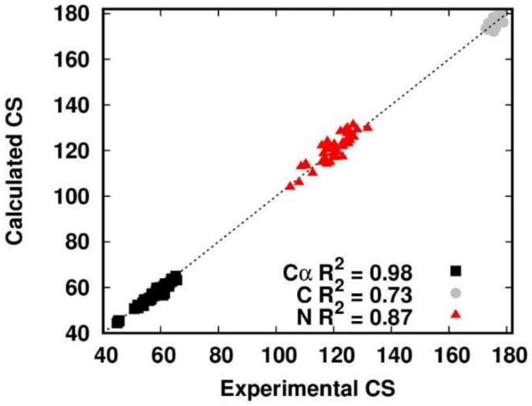


Figure 4. Correlation between experimental and calculated chemical shifts (CS) in the APO simulation. The Cα atoms are shown as black squares. The C atoms are shown as gray circles. The N atoms are shown as red triangles. The dashed line is y = x and is shown only to guide the eye.

In summary, apo-PfRd features a great structural stability. Nevertheless, the lack of metal binding affects the dynamics of local structural elements. For instance, the anomalous flexibility of Gly9 compared with its nearby residues could be the result of the destabilization of the aromatic sidechains in the protein core (Figure 5). Furthermore, some regions sampled non-native secondary structures that can result in a discrepancy between the S^2 and RMSF profiles. This is the case of residues 19–22 and 29–31 (i.e. in correspondence of the first and second native helix) where the RMSF plot shows a relatively high flexibility not detected by the S^2 analysis. The difference is related to the native helices being replaced by H-bonded turns (Table S1). Furthermore, the big loop involving residues 32 to 44, which in the crystal structure is stabilized by the metal binding of Cys38 and Cys41, here is broken in two smaller flexible parts (residues 34–35 and 39–41) by the formation of non-native β-strands involving residues 37–38 with residues 44–43. Thus, the non-native β-sheet partially compensates the stabilizing effect of the metal ion.

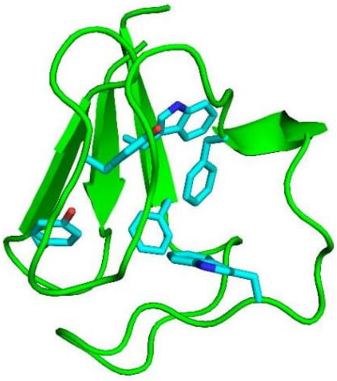


Figure 5. Snapshot at 1 μs showing the 3-D structure of the apo-PfRd. The aromatic side chains in the protein core are shown in cyan (Trp3, Tyr10, Tyr12, Phe29, Trp36 and Phe48).

3.2. Simulation of the folding process

The folding process of holo-PfRd was simulated using two different computational strategies: classical MD (F-cMD) and accelerated MD (F-aMD) (Table 1). The purpose of these simulations was to achieve a structure as close as possible to the folded state starting from an unfolded conformation, where we enforced metal coordination.

Figure 6 shows the S^2 profile for the simulations. As expected, the boost applied in the F-aMD simulation makes the average S^2 values for this simulation lower than for F-cMD. Following this trend, the N-terminus in the F-cMD is more rigid than in the F-aMD run. Nevertheless, the N-H vectors of the range 4–8 are quite rigid in both profiles due to the metal-binding Cys5 and Cys8. The order parameter values of F-cMD in the second half of the protein displays more rigidity than in the first half, especially in the metal binding region. Instead, for the F-aMD profile the second half of the protein appears as flexible as the first half, except in the metal binding region that has values similar to the binding region in the first half of the protein. In summary, while the overall dynamics sampled by the F-aMD and F-cMD simulations have some broad global similarities (Figure 6), we could pinpoint differences in the internal motion variations within specific protein regions.

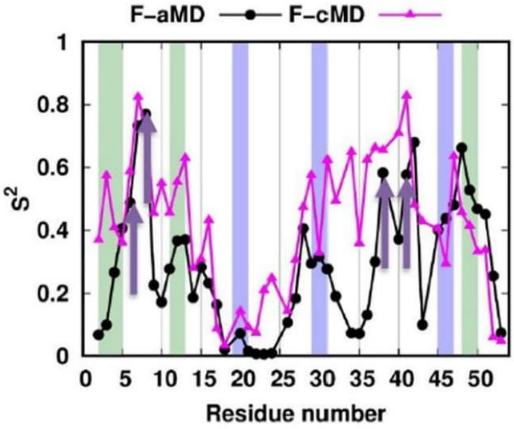


Figure 6. IRED order parameter profiles of the folding simulations. The residues corresponding to β-sheets and 3_{10}-helices in the crystal structure are colored with a green and blue background, respectively. The metal binding cysteines are indicated with yellow arrows. The NH vector S^2 profiles of the F-aMD and F-cMD trajectories are shown as black and magenta linepoints, respectively.

We analyzed the folding simulations in terms of secondary structures sampled (Figure 7). In the F-cMD conformational ensemble, the second and third strands of the β-sheet are consistently missing, whereas the first strand involves only residues 2 and 3 for the 40% and 1% of the frames, respectively (Figure 7A). The F-aMD trajectory has a higher presence of native β-strands except for residue 2. Regarding helical structures, the F-cMD conformational ensemble typically maintains only two residues of the native first 3_{10}-helix and the last residue of the native third helix (Figure 7B). More in detail, residues 20–21 and 47 are in helical conformation in 7% and 37% of the frames, respectively. Also for the helical structures, the F-aMD run shows a better agreement with the crystal. Indeed, all the three helices are sampled at least in 10% of the frames; the third helix in particular is present in more than 60% of the frames. We then conducted a more detailed per-residue analysis of the non-native (i.e. not present in the crystal structure) secondary structures sampled (Tables S2, S3 and Figure 8). In the F-cMD run, a non-native β-sheet structure is observed for two couples of residues: residues 2–28 and 23–51 in 41% and 37% of the frames, respectively. Furthermore, the F-cMD simulation shows an appreciable propensity to form helical structures in correspondence of the second and third native β-strands. The F-aMD simulation sampled more non-native structures than F-cMD. Among the non-native β-structures, the most persistent one involves residues 8 and 43. Notably, the last native β-sheet is replaced by a helix in most of the frames, as also seen in the F-cMD ensemble. As a result, the final helix spans an extended protein region, going from residue 45 to 50. In general, the F-aMD simulation sampled helical structures mostly in regions with defined secondary structures in the crystal or in their adjacent residues (Figure 8).

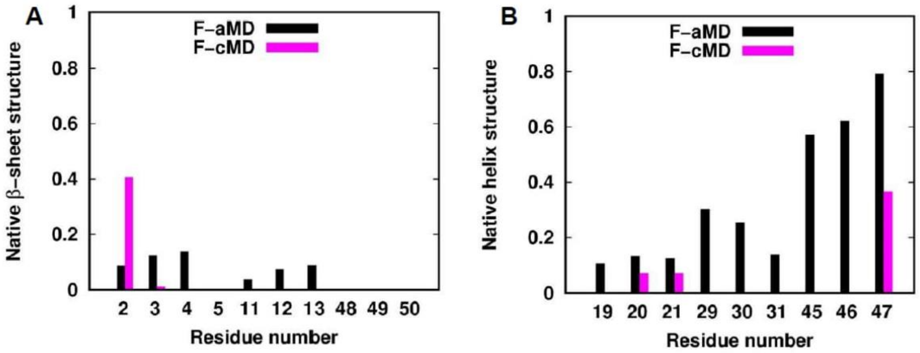


Figure 7. Fraction of native secondary structures sampled in the folding simulations. For each residue being in a secondary structure element in the crystal structure, we report the fraction of conformations that have the same secondary structure in the F-aMD and F-cMD simulations (black and magenta histograms, respectively). (A) Native β-sheets. (B) Native helices.

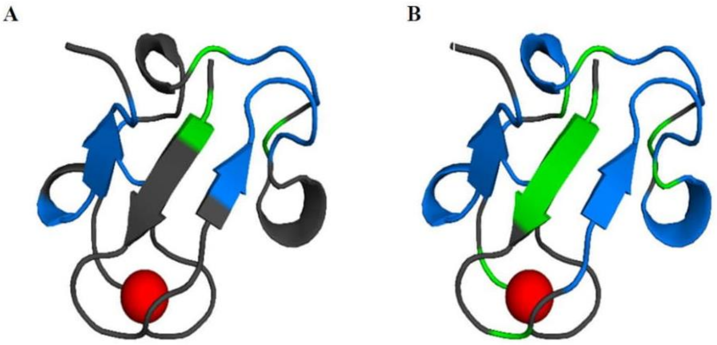


Figure 8. Secondary structures sampled by the F-cMD (A) and F-aMD (B) simulations mapped on the crystal structure of holo-PfRd. The residues where helical or strand structures are sampled in most of the conformations are shown in blue and green, respectively. Unstructured residues are shown in dark gray. The iron ion is shown as a red sphere.

Figure 9 displays the correlation of the chemical shifts predicted from the MD trajectories with experimental NMR data for the F-aMD and F-cMD trajectories. As a reference, a completely linearized protein was built *in silico* and used to compute the same correlation, to be used as a threshold. For the latter system, we observed a high correlation for the Cα atoms, whereas the correlation is poor for both the C and the N atoms. This is not unexpected given the strong dependence of the chemical shifts of the Cα atoms on the aminoacid identity. Both the folding trajectories display a weal correlation for the chemical shifts of the C atoms, with coefficients around 0.41–0.44. The larger variation was observed for the N atoms, with correlation coefficients for the F-aMD, F-cMD and linearized protein of 0.73, 0.70 and 0.68, respectively. Overall, the above data indicate that the F-aMD trajectory has marginally higher correlations with the experimental chemical shifts than the F-cMD trajectory.

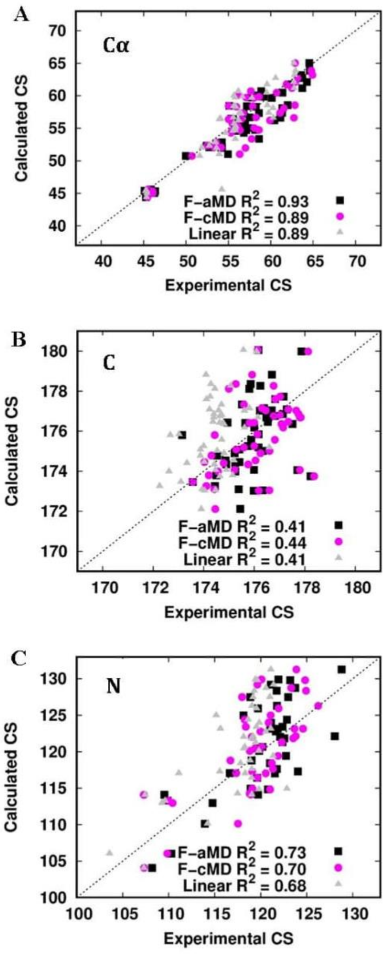


Figure 9. Correlation between experimental and calculated chemical shifts (in ppm) of the F-aMD trajectory, F-cMD trajectory and the linear protein (i.e. a model of the completely linear PfRd). The F-aMD values are shown as black squares. The F-cMD values are shown as magenta circles. The linear protein values are shown as grey triangles. (A) Plot of the Cα atoms chemical shifts. (B) Plot of the C atoms chemical shifts. (C) Plot of the N atoms chemical shifts.

The F-aMD simulation has greater flexibility and a better propensity to make native or native-like secondary structures than F-cMD (Figures 7 and 8). In addition, its predicted chemical shifts are slightly closer to the experimental data, although all the folding simulations failed to reproduce the chemical shifts of the C atoms (Figure 9). On the other hand, the APO simulation (starting from the folded structure) features a significant stability of the tertiary structure of the apo-form of PfRd and a strong correlation with the experimental NMR data.

3.3. Folding events in the F-aMD simulation

We analyzed the F-aMD trajectory along the simulation time to identify the structural features of possible folding events (Figure 10). The RMSD from the crystal structure of the backbone atoms spans a range from 6.5 to 12.5 Å (Figure 10A). The closest conformation to the crystal structure is reached around 2.2 µs; additional RMSD minima are observed after further 1–1.5 µs. Subsequently, the protein remains almost stable in a plateau at 11 Å for 5 µs. In the second half of the simulation a local minimum at about 7 Å from the crystal is reached three times: around 8, 10.1 and 10.7 µs. The relative partially folded conformations are shown.

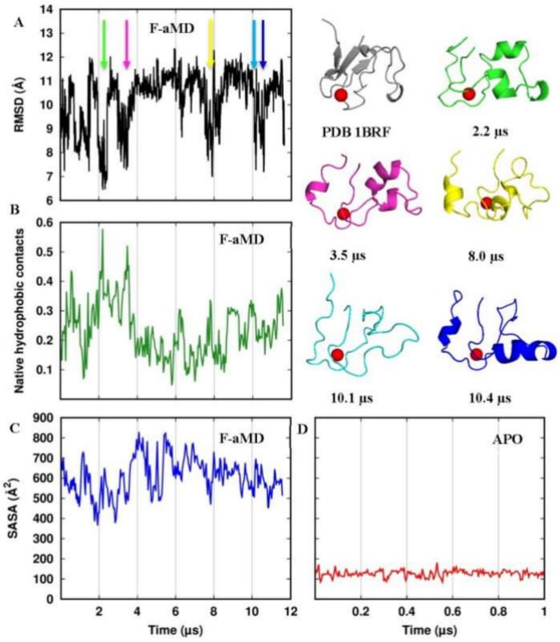


Figure 10. Structural properties along the F-aMD trajectory. (A) backbone RMSD from the crystal structure. The colored arrows indicated minima; the corresponding conformations are shown using the same color. (B) Fraction of native contacts occurring among the side chain residues in the hydrophobic core (residues Trp3, Tyr10, Tyr12, Phe29, Trp36, Phe48, Ile23 and Leu32). (C) SASA of the hydrophobic core. (D) SASA of the hydrophobic core along the APO simulation.

On average, hydrophobic residues tend to be in the core of a protein, where solvent accessibility is low, whereas polar residues tend to reside on the surface, where solvent accessibility is high. Thus, to clarify how the packing of the hydrophobic core affects the folding process, we measured the fraction of native hydrophobic contacts among the core residues (i.e. any contact between two hydrophobic residues of the protein core closer than 7 Å and present in the crystal structure). In the F-aMD simulation, the highest fraction of native contacts is reached at 2.2 µs and 3.5 µs, i.e. concurrently with two RMSD minima (Figure 10B). However, even the highest picks on the plot do not exceed the 60% of all the native contacts in the crystal structure. We additionally monitored the Solvent Accessibility Surface Area (SASA) of the PfRd hydrophobic residues as a function of simulation time (Figure 11C). The profile minimum around 400 Å^2 is achieved a few ns before 2 µs of simulation. Around 3 µs there is a second minimum. Both these minima occur just before the RMSD minima and native contacts maxima at 2.2 µs and 3.5 µs. These correlations suggest that shielding the hydrophobic core from the solvent is crucial to achieve the folding of the protein. After the first 3 µs, the protein was not able to lower the SASA around 400 Å^2 anymore, even though at 8 µs there is another local minimum below 500 Å^2, again just some ns before a RMSD minimum. In all the profiles of Figure 10, the best conformation appears at around 2 µs, when the protein first removes the water molecules from the hydrophobic core, and then achieves the highest number of hydrophobic contacts and the most native-like structure (Figure 10, structure in green). This correlation seems to be related to a unique folding event followed by a second similar attempt occurring about 1 µs later.

In general, the correlations among the parameters assessed confirm the strong relationship between the SASA and the number of native hydrophobic contacts profiles (Table 2). In fact, a negative value of −0.51 means that the protein has the necessity of isolating the protein core from the solvent to connect the hydrophobic residues correctly. The correlation of −0.27 between the RMSD from the crystal structure and the fraction of hydrophobic contacts, albeit weak, could indicate the crucial role of the hydrophobic core in pushing the simulation toward correctly folded conformations. The APO simulation constitutes a useful reference to understand the extent to which the folding events in the simulation produced conformations close to the native structure. Thus, considering that the APO trajectory has an average fraction of native hydrophobic contacts in the protein core of 0.74 (data not shown), the F-aMD top value of 0.58 suggests a good attempt but not a perfect packing. In addition, the SASA values sampled in the F-aMD simulation remain significantly larger than the values of the APO simulation, even in correspondence of the two minima (Figure 10C, D). Thus, even if the folding simulation generates a reasonable number of native contacts in the protein core, the hydrophobic residues remain accessible to the water in the conformations explored.

The projection of the F-aMD and APO conformational ensembles on the first two eigenvectors in the Cα space shows the extent of sampling reached by applying the boost potential in contrast to the stability of the APO simulation (Figure S1). This PCA analysis shows that the two simulations are separated by a relevant free energy barrier; note that the F-aMD free energy has not been reweighted to restore the canonical ensemble. However, the projection of the F-aMD conformation closest to crystal, i.e. the one at 2.2 µs, is located in proximity to the APO trajectory projection.

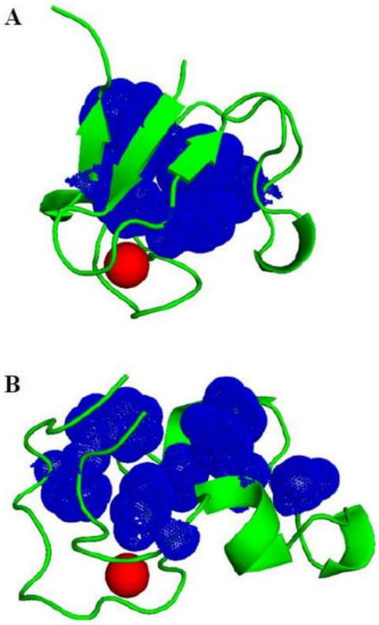


Figure 11. Hydrophobic cluster compactness. The residue side chains of the hydrophobic core are shown as blue dots. The iron ion is shown as red sphere. (A) The highly compact hydrophobic core in the crystal structure PDB 1BRF. (B) The less compact hydrophobic core in the closest to crystal conformation sampled in the F-aMD run.

Table 2. Correlations among the analyses of the F-aMD trajectory.

Structural parameters compared		R
RMSD	SASA	0.24
RMSD	Native hydrophobic contacts	−0.27
SASA	Native hydrophobic contacts	**−0.51**

The strongest correlation is highlighted in bold. R is the Pearson's coefficient. See Figure 10 to observe how each parameter evolves during the simulation.

4. Conclusions

Our analyses indicate that the APO simulation has sampled conformations close to the folded state, as expected. Nevertheless, there are some visible effects due to the absence of the iron ion on the structural properties of apo-PfRd. In particular, up to three non-native β-sheets between residues in spatial proximity, which partly compensates the degrees of freedom gained with the removal of the iron ion. In addition, we observed that the 3_{10} helical regions tend to assume an H-bonded turn configuration; this is more prominent for the first helix. Similar trends are observed also in the

F-aMD simulation, where some non-native β-structures also occur. The occurrence of secondary structures in the F-aMD simulation generally maps with good accuracy to the same residues that are in helical or β-structures also in the crystal structure. This simulation overestimates the occurrence of the elements of helical structure. A significant example is that of the last 3_{10}-helix, which extends to involve also residues of the last β-strand of the native sheet (Figure 7). In fact, the full native triple-stranded β-sheet is never sampled, because the simulation is not able to recover all the long-range contacts needed to make the correct β-structures. The NMR chemical shifts of the backbone nuclei of holo-PfRd provide an experimental reference for the understanding of the average dynamic and structural features of the various simulations. For the APO simulation, the agreement of the back-calculated shifts with the experimental data is very good, confirming the absence of significant rearrangements. The situation is somewhat different for the folding simulations, which feature modest correlations for the C and N atoms.

The extensive sampling of the F-aMD trajectory has produced conformations with a compact shape relatively similar to the native structure. This is supported by the analysis of the RMSD from the crystal, and its correlations with the SASA profile and the number of hydrophobic contacts among the core residues (Figure 10 and Table 2). These data suggest that the F-aMD simulation indeed sampled putative folding pathways. In these events, the SASA reduced to 400–500 Å^2, and shortly after the RMSD values dropped to about 6–7 Å from the crystal structure. At the same time, the number of native contacts in the hydrophobic core increased to the values observed also in the APO simulation. These results indicate that the major obstacle for the complete folding of holo-PfRd is the residual presence of the solvent molecules in the core residues. As a result, one of the potential driving force in the folding process is weakened. This prevents the compaction of N-terminal part of the structure with the rest of the core (Figure 11). Interestingly, the interaction of the first 15 residues of PfRd with the other parts of the protein has been shown to be an important contributor to the thermostability of the system [20]. An additional contribution limiting our simulation is likely the oversampling of helical structure especially for residues 48–50.

The extensive sampling of one of the smallest known metalloproteins achieved here by a combination of accelerated dynamics and the use of GPGPUs shows that metal-coupled folding is still a challenging task for unbiased MD methods. The main difficulties are not limited to the accuracy of the force field describing the metal binding, such as metal induced protonation/deprotonation [40,41], the polarizable effect [42], the charge transfer [43,44] and multiscale coupling [33–35]. Indeed, the standard force field may also induce some bias for the protein part. This is exemplified by the folded apo-form of the protein adopting some non-native secondary structures during the APO simulation. These phenomena affect also the F-aMD and F-cMD simulations. Indeed, there are known limitations in the description of various force fields of interactions that are crucial here, such as those involving phenylalanine side chains [51]. Other authors pointed out that electrostatics and water descriptions could be the weakest force field elements, and proposed that their optimization should consider unfolded proteins [52]. In agreement with this, we suggest that significant methodological work is still needed until unbiased metal-induced folding of metalloproteins can be achieved.

Acknowledgments

This work was supported by the EUROPEAN COMMISSION (project number 675858, West-Life) and by CIRMMP.

Conflict of interest

All authors declare no conflict of interest in this paper.

References

1. Klepeis JL, Lindorff-Larsen K, Dror RO, et al. (2009) Long-timescale molecular dynamics simulations of protein structure and function. *Curr Opin Struct Biol* 19: 120–127.
2. Stone JE, Phillips JC, Freddolino PL, et al. (2007) Accelerating molecular modeling applications with graphics processors. *J Comput Chem* 28: 2618–2640.
3. Perez A, Morrone JA, Simmerling C, et al. (2016) Advances in free-energy-based simulations of protein folding and ligand binding. *Curr Opin Struct Biol* 36: 25–31.
4. Lane TJ, Shukla D, Beauchamp KA, et al. (2013) To milliseconds and beyond: Challenges in the simulation of protein folding. *Curr Opin Struct Biol* 23: 58–65.
5. Freddolino PL, Harrison CB, Liu Y, et al. (2010) Challenges in protein-folding simulations. *Nat Phys* 6: 751–758.
6. Best RB (2012) Atomistic molecular simulations of protein folding. *Curr Opin Struct Biol* 22: 52–61.
7. Piana S, Lindorff-Larsen K, Shaw DE (2012) Protein folding kinetics and thermodynamics from atomistic simulation. *P Natl Acad Sci USA* 109: 17845–17850.
8. Suárez E, Lettieri S, Zwier MC, et al. (2014) Simultaneous computation of dynamical and equilibrium information using a weighted ensemble of trajectories. *J Chem Theory Comput* 10: 2658–2667.
9. Pierce LCT, Salomon-Ferrer R, Augusto F. De Oliveira C, et al. (2012) Routine access to millisecond time scale events with accelerated molecular dynamics. *J Chem Theory Comput* 8: 2997–3002.
10. Kubelka J, Hofrichter J, Eaton WA (2004) The protein folding "speed limit." *Curr Opin Struct Biol* 14: 76–88.
11. Lindorff-Larsen K, Piana S, Dror RO, et al. (2011) How fast-folding proteins fold. *Science* 334: 517–520.
12. Putignano V, Rosato A, Banci L, et al. (2018) MetalPDB in 2018: a database of metal sites in biological macromolecular structures. *Nucleic Acids Res* 41: 459–464.
13. Li W, Wang J, Zhang J, et al. (2015) Molecular simulations of metal-coupled protein folding. *Curr Opin Struct Biol* 30: 25–31.
14. Bentrop D, Bertini I, Iacoviello R, et al. (1999) Structural and dynamical properties of a partially unfolded Fe4S4 protein: Role of the cofactor in protein folding. *Biochemistry* 38: 4669–4680.

15. Blake PR, Summers MF, Park JB, et al. (1991) Determinants of protein hyperthermostability: purification and amino acid sequence of rubredoxin from the hyperthermophilic archaebacterium pyrococcus furiosus and secondary structure of the zinc adduct by NMR. *Biochemistry* 30: 10885–10895.
16. Prakash S, Sundd M, Guptasarma P (2014) The key to the extraordinary thermal stability of P. furiosus holo-rubredoxin: Iron binding-guided packing of a core aromatic cluster responsible for high kinetic stability of the native structure. *PLoS One* 9: e89703.
17. Hernandez G, Jenney FE, Adams MW, et al. (2000) Millisecond time scale conformational flexibility in a hyperthermophile protein at ambient temperature. *Proc Natl Acad Sci USA* 97: 3166–3170.
18. Rader AJ (2010) Thermostability in rubredoxin and its relationship to mechanical rigidity. *Phys Biol* 7: 016002.
19. Bonomi F, Iametti S, Ferranti P, et al. (2008) "Iron priming" guides folding of denatured aporubredoxins. *J Biol Inorg Chem* 13: 981–991.
20. Zartler ER, Jenney FE, Terrell M, et al. (2001) Structural basis for thermostability in aporubredoxins from Pyrococcus furiosus and Clostridium pasteurianum. *Biochemistry* 40: 7279–7290.
21. Cavagnero S, Debe DA, Zhou ZH, et al. (1998) Kinetic role of electrostatic interactions in the unfolding of hyperthermophilic and mesophilic rubredoxins. *Biochemistry* 37: 3369–3376.
22. Strop P, Mayo SL (1999) Rubredoxin variant folds without iron. *J Am Chem Soc* 121: 2341–2345.
23. Hamelberg D, Mongan J, McCammon JA (2004) Accelerated molecular dynamics: A promising and efficient simulation method for biomolecules. *J Chem Phys* 120: 11919–11929.
24. Doshi U, Hamelberg D (2015) Towards fast, rigorous and efficient conformational sampling of biomolecules: Advances in accelerated molecular dynamics. *BBA-Gen Subjects* 1850: 878–888.
25. Miao Y, Feixas F, Eun C (2015) Accelerated molecular dynamics simulations of protein folding. *J Comput Chem* 36: 1536–1549.
26. Case DA, Cerutti DS, Cheatham TE, et al. (2017) Amber 2017, University of California, San Francisco.
27. Carvalho ATP, Teixeira AFS, Ramos MJ (2013) Parameters for molecular dynamics simulations of iron-sulfur proteins. *J Comput Chem* 34: 1540–1548.
28. Bertini I, Case DA, Ferella L, et al. (2011) A grid-enabled web portal for NMR structure refinement with AMBER. *Bioinformatics* 27: 2384–2390.
29. Wassenaar TA, van Dijk M, Loureiro-Ferreira N, et al. (2012) WeNMR: Structural biology on the grid. *J Grid Comput* 10: 743–767.
30. Prompers JJ, Brüschweiler R, Bruschweiler R (2002) General framework for studying the dynamics of folded and nonfolded proteins by NMR relaxation spectroscopy and MD simulation. *J Am Chem Soc* 124: 4522–4534.
31. Korzhnev DM, Billeter M, Arseniev AS, et al. (2001) NMR studies of Brownian tumbling and internal motions in proteins. *Prog Nucl Mag Res Sp* 38: 197–266.
32. Kabsch W, Sander C (1983) Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* 22: 2577–2637.

33. Rost B, Sander C (1993) Prediction of protein secondary structure at better than 70% accuracy. *J Mol Biol* 232: 584–599.
34. Li DW, Brüschweiler R (2012) PPM: A side-chain and backbone chemical shift predictor for the assessment of protein conformational ensembles. *J Biomol NMR* 54: 257–265.
35. Hiller R, Zhou ZH, Adams MW, et al. (1997) Stability and dynamics in a hyperthermophilic protein with melting temperature close to 200 degrees C. *Proc Natl Acad Sci USA* 94: 11329–11332.
36. Ishima R, Torchia DA (2000) Protein dynamics from NMR. *Nat Struct Biol* 7: 740–743.
37. Jarymowycz VA, Stone MJ (2006) Fast time scale dynamics of protein backbones: NMR relaxation methods, applications, and functional consequences. *Chem Rev* 106: 1624–1671.
38. LeMaster DM (1999) NMR relaxation order parameter analysis of the dynamics of protein side chains. *J Am Chem Soc* 121: 1726–1742.
39. Ruschak AM, Kay LE (2010) Methyl groups as probes of supra-molecular structure, dynamics and function. *J Biomol NMR* 46: 75–87.
40. Bougault CM, Eidsness MK, Prestegard JH (2003) Hydrogen bonds in rubredoxins from mesophilic and hyperthermophilic organisms. *Biochemistry* 42: 4357–4372.
41. Prestegard JH, Bougault CM, Kishore AI (2004) Residual dipolar couplings in structure determination of biomolecules. *Chem Rev* 104: 3519–3540.
42. Cho-Chung YS, Pitot HC (1968) Regulatory effects of nicotinamide on tryptophan pyrrolase synthesis in rat liver in vivo. *Eur J Biochem* 3: 401–406.
43. Blasie CA, Berg JM (2002) Structur e-based thermodynamic analysis of a coupled metal binding-protein folding reaction involving a zinc finger peptide. *Biochemistry* 41: 15068–15073.
44. Weinkam P, Romesberg FE, Wolynes PG (2009) Chemical frustration in the protein folding landscape: Grand canonical ensemble simulations of cytochrome c. *Biochemistry* 48: 2394–2402.
45. Devereux M, Gresh N, Piquemal JP, et al. (2014) A supervised fitting approach to force field parametrization with application to the SIBFA polarizable force field. *J Comput Chem* 35: 1577–1591.
46. Wu R, Lu Z, Cao Z, et al. (2011) A transferable nonbonded pairwise force field to model zinc interactions in metalloproteins. *J Chem Theory Comput* 7: 433–443.
47. Sakharov DV, Lim C (2005) Zn protein simulations including charge transfer and local polarization effects. *J Am Chem Soc* 127: 4921–4929.
48. Chakravorty DK, Wang B, Lee CW, et al. (2012) Simulations of allosteric motions in the zinc sensor CzrA. *J Am Chem Soc* 134: 3367–3376.
49. Chakravorty DK, Parker TM, Guerra AJ, et al. (2013) Energetics of zinc-mediated interactions in the allosteric pathways of metal sensor proteins. *J Am Chem Soc* 135: 30–33.
50. Reyes-Caballero H, Campanello GC, Giedroc DP (2011) Metalloregulatory proteins: Metal selectivity and allosteric switching. *Biophys Chem* 156: 103–114.
51. Andrews CT, Elcock AH (2013) Molecular dynamics simulations of highly crowded amino acid solutions: comparisons of eight different force field combinations with experiment and with each other. *J Chem Theory Comput* 9: 4585–4602.

52. Abriata LA, Dal Peraro M (2015) Assessing the potential of atomistic molecular dynamics simulations to probe reversible protein-protein recognition and binding. *Sci Rep* 5: 10549.

AIMS Biophysics, 5(1): 77–96.
DOI: 10.3934/biophy.2018.1.77
Received: 19 January 2017
Accepted: 12 March 2018
Published: 22 March 2018

http://www.aimspress.com/journal/biophysics

Research article

Molecular dynamics simulations of metalloproteins: A folding study of rubredoxin from *Pyrococcus furiosus*

Davide Sala[1], Andrea Giachetti[2] and Antonio Rosato[1,3,*]

[1] Magnetic Resonance Center (CERM), University of Florence, Via Luigi Sacconi 6, 50019 Sesto Fiorentino, Italy
[2] Interuniversity Consortium of Magnetic Resonance of Metallo Proteins (CIRMMP), Via Luigi Sacconi 6, 50019 Sesto Fiorentino, Italy
[3] Department of Chemistry, University of Florence, Via della Lastruccia 3, 50019 Sesto Fiorentino, Italy

* **Correspondence:** Email: rosato@cerm.unifi.it; Tel: +390554574267.

Supplementary

Table S1. Secondary structure persistency per-residue of the APO simulation.

#Residue	Strand	Helix	Turn	#Residue	Strand	Helix	Turn
1	0.00	0.00	0.00	28	0.00	0.00	0.00
2	0.97	0.00	0.00	29	0.00	0.58	0.25
3	1.00	0.00	0.00	30	0.00	0.58	0.41
4	1.00	0.00	0.00	31	0.00	0.58	0.38
5	1.00	0.00	0.00	32	0.00	0.00	0.00
6	0.00	0.00	0.49	33	0.00	0.00	0.00
7	0.00	0.00	0.49	34	0.00	0.00	0.93
8	0.02	0.00	0.32	35	0.00	0.00	0.93
9	0.00	0.00	0.09	36	0.00	0.00	0.00

Continued on next page

#Residue	Strand	Helix	Turn	#Residue	Strand	Helix	Turn
10	0.09	0.00	0.00	37	**0.91**	0.00	0.00
11	1.00	0.00	0.00	38	**0.42**	0.00	0.00
12	1.00	0.00	0.00	39	0.00	0.01	0.98
13	0.95	0.00	0.00	40	0.00	0.01	0.99
14	0.00	0.02	0.97	41	0.00	0.01	0.98
15	0.00	0.02	0.98	42	0.00	0.00	0.42
16	0.00	0.02	0.97	43	**0.42**	0.00	0.00
17	0.10	0.00	0.00	44	**0.91**	0.00	0.00
18	**0.98**	0.00	0.00	45	0.00	0.67	0.32
19	0.00	0.23	0.76	46	0.00	0.67	0.33
20	0.00	0.23	0.76	47	0.00	0.67	0.01
21	0.00	0.23	0.76	48	0.99	0.00	0.00
22	0.00	0.00	0.99	49	1.00	0.00	0.00
23	**0.98**	0.00	0.00	50	0.99	0.00	0.00
24	0.00	0.00	0.94	51	0.05	0.00	0.00
25	0.00	0.00	0.94	52	**0.78**	0.00	0.00
26	0.00	0.00	0.94	53	0.00	0.00	0.00
27	0.00	0.00	0.00				

The residues involved in a native helix and β-strand are colored in blue and green, respectively. Residues with significant (>0.3) persistency of non-native secondary structures are in bold.

Table S2. Secondary structure persistency per-residue of the F-cMD simulation.

#Residue	Strand	Helix	Turn	#Residue	β-sheet	Helix	Turn
1	0.00	0.00	0.00	28	**0.41**	0.00	0.00
2	0.41	0.00	0.00	29	0.01	0.00	0.00
3	0.01	0.00	0.00	30	0.00	0.00	0.00
4	0.00	0.00	0.00	31	0.00	0.00	0.00
5	0.00	0.00	0.00	32	0.00	0.00	0.00
6	0.00	0.00	0.00	33	0.00	0.00	0.00
7	0.00	0.00	0.00	34	0.00	0.15	0.85
8	0.01	0.00	0.00	35	0.00	0.15	0.85
9	0.00	0.00	0.00	36	0.00	0.15	0.79
10	0.00	0.00	0.00	37	0.00	0.00	0.00
11	0.00	0.08	0.40	38	0.00	0.00	0.00
12	0.00	0.12	0.46	39	0.00	0.00	0.00
13	0.00	**0.52**	0.37	40	0.01	0.00	0.00
14	0.00	**0.52**	0.43	41	0.01	0.00	0.00
15	0.00	**0.49**	0.35	42	0.00	0.00	0.00
16	0.00	0.27	0.24	43	0.00	0.00	0.00
17	0.00	0.11	0.26	44	0.00	0.00	0.00
18	0.00	0.00	0.29	45	0.01	0.00	0.00

Continued on next page

#Residue	Strand	Helix	Turn	#Residue	β-sheet	Helix	Turn
19	0.00	0.00	0.29	46	0.00	0.00	0.00
20	0.01	0.07	0.04	47	0.00	0.37	0.57
21	0.01	0.07	0.30	48	0.00	**0.41**	0.57
22	0.01	0.07	0.28	49	0.00	**0.50**	0.48
23	**0.37**	0.00	0.00	50	0.00	**0.45**	0.31
24	0.00	0.00	0.00	51	**0.37**	0.19	0.15
25	0.00	0.14	0.72	52	0.00	0.07	0.10
26	0.00	0.14	0.73	53	0.00	0.00	0.00
27	0.00	0.14	0.26				

The residues involved in native helix and β-strand structures are colored in blue and green, respectively. Residues with significant (>0.3) persistency of non-native strand and helix are in bold and underlined bold, respectively.

Table S3. Secondary structure persistency per-residue of the F-aMD simulation.

#Residue	Strand	Helix	Turn	#Residue	β-sheet	Helix	Turn
1	0.00	0.00	0.00	28	0.17	**0.32**	0.16
2	0.09	0.00	0.06	29	0.12	0.30	0.33
3	0.12	0.00	0.07	30	0.05	0.25	0.39
4	0.14	0.00	0.01	31	0.02	0.13	0.43
5	0.00	0.00	0.01	32	0.11	0.05	0.08
6	0.08	0.00	0.25	33	0.01	0.09	0.12
7	0.00	0.00	0.27	34	0.00	0.10	0.31
8	**0.32**	0.00	0.05	35	0.01	0.09	0.32
9	0.05	0.04	0.11	36	0.00	0.01	0.10
10	0.01	0.13	0.12	37	0.01	0.00	0.00
11	0.04	0.16	0.11	38	0.00	0.00	0.00
12	0.07	0.17	0.11	39	0.01	0.00	0.00
13	0.09	0.19	0.12	40	0.01	0.00	0.00
14	0.01	0.15	0.33	41	0.03	0.00	0.34
15	0.00	0.12	0.35	42	0.00	0.00	0.36
16	0.02	0.08	0.33	43	**0.31**	0.00	0.01
17	0.04	0.05	0.15	44	0.04	0.01	0.03
18	0.16	0.05	0.03	45	0.00	0.57	0.08
19	0.07	0.10	0.33	46	0.00	0.62	0.07
20	0.01	0.13	0.38	47	0.01	0.79	0.11
21	0.02	0.12	0.42	48	0.00	**0.80**	0.10
22	0.08	0.07	0.27	49	0.00	**0.65**	0.22
23	0.06	0.05	0.15	50	0.00	**0.53**	0.26
24	0.07	0.03	0.05	51	0.01	**0.39**	0.24
25	0.06	0.17	0.18	52	0.01	0.08	0.13
26	0.05	0.23	0.22	53	0.00	0.00	0.00
27	0.08	**0.31**	0.17				

The residues involved in a native helix or β-strand structure in the crystal are colored in blue and green, respectively. Residues with significant (>0.3) persistency of non-native strand or helix structure are in bold and underlined bold, respectively.

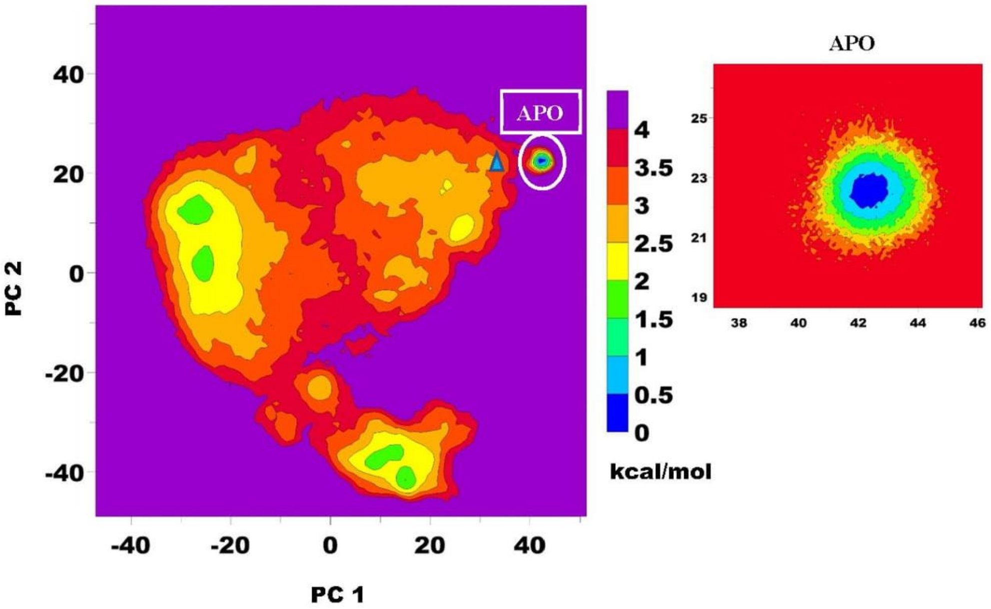


Figure S1. Principal Component Analysis of the F-aMD and APO trajectories. Projection of the two conformational ensembles on the first two eigenvectors. The APO region is zoomed on the right. The F-aMD conformation closest to the crystal structure at 2.2 μs is projected with a cyan triangle.

3.3.2 Investigation of the Iron(II) Release Mechanism of Human H-Ferritin as a Function of pH

Davide Sala[†], *Silvia Ciambellotti*[†], *Andrea Giachetti*[‡], *Paola Turano*[†,§] *and Antonio Rosato*[†,§,*]

[†]Magnetic Resonance Center (CERM), University of Florence, Via Luigi Sacconi 6, 50019 Sesto Fiorentino, Italy.

[‡]Consorzio Interuniversitario di Risonanze Magnetiche di Metallo Proteine, Via Luigi Sacconi 6, 50019 Sesto Fiorentino, Italy.

[§]Department of Chemistry, University of Florence, Via della Lastruccia 3, 50019 Sesto Fiorentino, Italy.

J Chem Inf Model. 2017 Sep 25;57(9):2112-2118. Epub 2017 Sep 12.

JOURNAL OF
CHEMICAL INFORMATION
AND MODELING

Letter

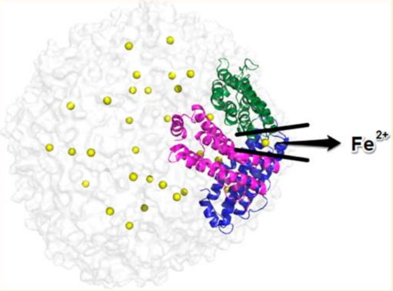
pubs.acs.org/jcim

Investigation of the Iron(II) Release Mechanism of Human H-Ferritin as a Function of pH

Davide Sala,[†] Silvia Ciambellotti,[†] Andrea Giachetti,[‡] Paola Turano,[†,§] and Antonio Rosato*[,†,§]

[†]Magnetic Resonance Center (CERM), University of Florence, Via Luigi Sacconi 6, 50019 Sesto Fiorentino, Italy
[‡]Consorzio Interuniversitario di Risonanze Magnetiche di Metallo Proteine, Via Luigi Sacconi 6, 50019 Sesto Fiorentino, Italy
[§]Department of Chemistry, University of Florence, Via della Lastruccia 3, 50019 Sesto Fiorentino, Italy

S *Supporting Information*



ABSTRACT: We investigated the kinetics of the release of iron(II) ions from the internal cavity of human H-ferritin as a function of pH. Extensive molecular dynamics simulations of the entire 24-mer ferritin provided atomic-level information on the release mechanism. Double protonation of His residues at pH 4 facilitates the removal of the iron ligands within the C3 channel through the formation of salt bridges, resulting in a significantly lower release energy barrier than pH 9.

INTRODUCTION

Maxi ferritins are 24-mer nanocage structures that self-assemble from 4-helix bundle subunits.[1] The resulting structure has octahedral symmetry. In correspondence to the C3 and C4 symmetry axes there are two different channels connecting the bulk solution to the inner cage cavity (Figure 1). The two types of channels differ in their size and chemical properties.[2−5] In vertebrate ferritins, the eight C3 channels constitute the entry points of iron(II) ions,[2,6] which transit under the form of hexahydrated ions under the effect of the electrostatic gradient generated by carboxylate groups at the inner end of the channels. The iron release mechanism in biomineralized ferritin is less characterized.[7] In vitro, release experiments generally involve the addition of a reducing agent in the presence of chelators.[8−11] The process in vivo might be associated with cage disassembly,[12,13] or it could depend on specific acidic environments.[14,15] H-ferritin enters cells via receptor-mediated endocytosis mediated by Transferrin Receptor 1 (TfR1).[16] Binding of H-ferritin to TfR1 results in the uptake of H-ferritin into endosomes and lysosomes in a dynamic process. During their maturation from early to late endosomes, the pH inside these organelles is in the 6.0−4.8 range whereas in lysosomes the pH can drop to values around 4.5.[17,18]

Our kinetic measurements on human H-ferritin in the presence of different reductants showed that iron release has different kinetic profiles but with the same distinct pH dependence (Figure 2A and B). In particular, lowering the pH from 9.0 to 4.0 increased significantly the amount of iron discharged. We used MD simulations to identify the parts of the protein responsible for the interaction with the iron ions during the release process. We focused on the processes occurring at pH 4, where iron release is enhanced, and compared them to pH 9. Furthermore, the umbrella sampling method was applied to estimate the corresponding free-energy barriers. At pH 4, all His residues are doubly protonated, whereas all carboxylates are deprotonated. Consequently, the internal cavity of ferritin provides an environment whose electrostatic potential is close to neutrality due to the balance of the charges of the doubly protonated His side chains (positive) and of the carboxylate side chains (negative). Therefore, the electrostatic surface of the protein's cage does not stabilize significantly metal ions in the absence of mineralization. The iron(II) ions thus respond readily to the gradients of electrostatic potential leading them toward other regions of the structure where negatively charged residues are spatially clustered. Such gradients drive the iron ions toward two main locations: the ferroxidase catalytic sites and the C3 channels (movie S1). The same behavior takes place at pH 9 despite the fact that in each chain nine out of ten His residues are neutral and thus the electrostatic charge within the cavity at this pH is negative. In this work, we obtained evidence that the C3 channels are the pathway for the release of iron(II) ions from the cavity to the bulk solution[19] (movie S2). Our MD simulations provide a detailed atomic-level view of the mechanism of iron release from ferritin, which was still missing.[20]

RESULTS

The eight C3 channels are formed by three symmetry-related motifs,[21] specifically helix α4-loop-α3, from as many chains (A, B, and C). Within each individual channel, two nearby rings formed by symmetry-related Asp and Glu side chains define the binding site to which the iron ions move from the internal cavity (Figure 1). The side chains of the Asp131 residues define

Published: August 30, 2017

ACS Publications © 2017 American Chemical Society 2112 DOI: 10.1021/acs.jcim.7b00306
J. Chem. Inf. Model. 2017, 57, 2112−2118

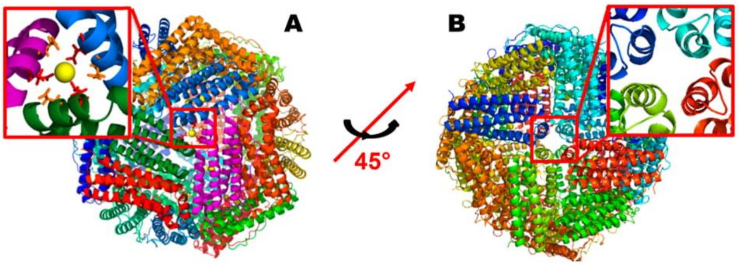


Figure 1. Structure of human H-ferritin viewed down the C3 and C4 axes, related by a 45° rotation around the red axis. Each subunit is colored differently. Iron ions inside the C3 channel are shown as yellow spheres. (A) The C3 channel inside view is illustrated in the red box on the left. Glu134 and Asp131 side chains are shown as red and orange sticks, respectively. (B) The C4 channel is illustrated in the red box on the right.

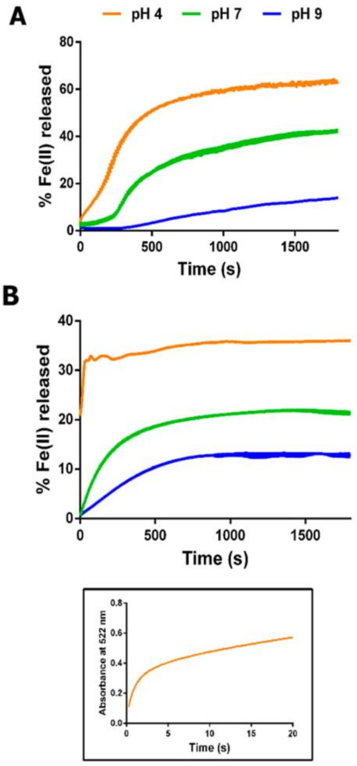


Figure 2. Iron(II) release in human H-ferritin at different pH conditions. The plots show the percentage of iron(II) released over 1800 s at pH 4, 7, and 9 by human holo (480 Fe^{2+} per cage) H-ferritin. The curves represent the mean ± SEM of six independent analyses. The student's *t* test indicated that the curves are significantly different ($P < 0.05$). The graph shows the iron(II) release in the presence of (A) NADH/FMN reductants or (B) sodium dithionite. (B inset) First seconds of the kinetics at pH 4 in the presence of dithionite (represented as absorbance at 522 nm over time) acquired with a stopped-flow spectrophotometer, which allowed us to measure the initial points of the reaction that are lost with manual mixing.

a first ring that is closer to the internal cavity of ferritin, whereas the side chains of the Glu134 residues define a second ring, further from the internal cavity, and thus closer to the protein surface. In our simulation at pH 4, three of the eight C3 sites become populated at different time points by a single iron(II) ion. At pH 9 four sites become populated. Typically, the iron ion reached the site within the first nanoseconds of the simulation. The fact that not all eight channels become populated can be due to the relatively low number of ions introduced in the cavity (31 ions) as well as to the initial position of each ion, e.g., because of the different relative distance between each ion and the closest ferroxidase vs C3 sites. In addition, the local conformations of the eight sites in the initial structural model, built by aligning the human H-chain on the bullfrog crystal structure[22] (PDB entry 4DAS), are not identical.

Before the iron ion actually reaches the C3 site, the Asp131 side chains move apart thereby allowing the metal ion to get into the space between the two rings of carboxylates (Figure S1). Within the site, the iron can be in proximity to the side chains of the Glu134 and Asp131 residues (Figure 1). Once inside, the metal ion moves toward the exit of the channel and, in so doing, it transitions between different coordination configurations. As an example, Figure 3 shows how the coordination of iron in one of the C3 sites at pH 4 is affected by the dynamics of Glu134(C) and its interactions with neighboring histidinium (i.e., the cationic form of His) side chains. When the iron enters the site, it is at a short distance from all three Glu residues. However, after only a few ns Glu134(C) changes conformation and moves away from the metal (Figure 3). This is triggered by the subsequent formation of two salt bridges involving the negatively charged carboxylate of Glu134(C). The first salt bridge is formed with the positively charged side chain of His128(A), which is later (after about 1 ns) displaced by His136(A). The latter event causes the carboxylate of Glu134(C) to rotate away from the iron ion and toward the cavity of ferritin (Figures 3 and S2). The other two Glu134 residues in the site (chains A and B) are released only when the iron ion exits the C3 channel (Figure 4).

The pairwise distances among the three Glu134 residues of the three chains define the process of opening of the channel. For the same C3 site described previously, the three distances are similar in the initial configuration of the ferritin structure (Figure 4C). Upon the arrival of the iron ion in the site, they evolve into one shorter (about 6 Å) and two longer distances (8−8.5 Å). The situation changes further upon the rearrangement of Glu134(C) described in the previous paragraph and Figure 3. After this event, the distance between Glu134(A) and Glu134(B) oscillates between 6 and 8 Å but is typically closer to the lower part of this range, whereas the two distances involving Glu134(C) rapidly increase above 8 Å reaching peaks

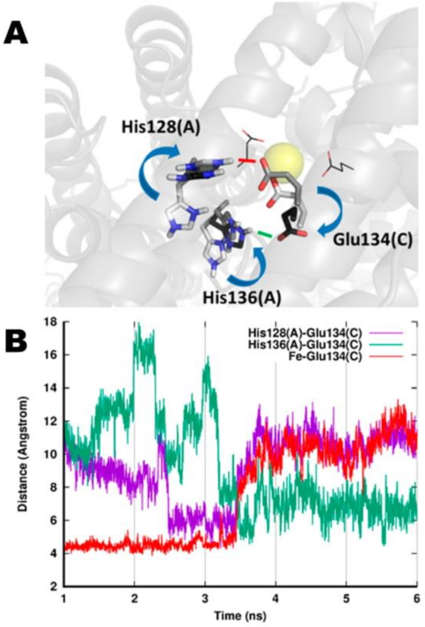


Figure 3. Switch of iron coordination in a C3 site. (A) Overlap of three snapshots at 2.36 ns (white), 3.38 ns (gray), and 3.63 ns (black). Residue side chains involved in the process are shown as sticks. The red dash indicates the salt bridge between His128(A) and Glu134(C) at 3.38 ns that weakens the iron(II)−Glu134(C) interaction. The green dash shows the salt bridge between His136(A) and Glu134(C) that is formed at 3.63 ns, just after iron(II)−Glu134(C) detachment. (B) Iron(II)−Glu134(C), His128(A)−Glu134(C), and His136(A)−Glu134(C) distances for the first channel. His−Glu distances refer to the Hε and Cδ atoms of the His and Glu residues, respectively.

of 12 Å. While the iron ion is in the site, the carboxylate Oε oxygen atoms do not form consistent hydrogen bonds to donor groups in spatial proximity and thus remain close to each other due to electrostatic interaction with the metal ion. At pH 4 this structural configuration evolves during the MD simulation, eventually leading to the release of the iron ion (see below). Instead, the corresponding configuration at pH 9 remains stable throughout the entire simulation (e.g., Figures S3 and S4).

The protein dynamics at pH 4 leads to conformations featuring relatively long-lived salt bridges involving the Glu134 and the doubly protonated His118 residues in the channel. This is apparent between 5 and 20 ns of the simulation (Figure 4A and B). The presence of these salt bridges can be conveniently monitored by looking the hydrogen bond persistence between the side chains (Figure 4D).

The iron ion is released when the A and B chains simultaneously feature long-lived hydrogen bonds, between 15 and 20 ns (Figure 4B and D). This configuration reinforces the correlation between the dynamics of the protein structural environment and the dynamics of the Glu134 side chains. In addition, there is a partial compensation of the negative electrostatic charge of the carboxylates of the Glu134 residues. Eventually, the iron ion escapes the C3 binding site and irreversibly (in our simulations) diffuses into the bulk solution (Figure 4). The channel closes after the iron release and Glu134(C) comes back to the starting orientation (Figure 4C and movie S3). When the His residues are singly protonated, the aforementioned electrostatic mechanisms are no longer possible. Indeed, at pH 9, where all His are singly protonated, the iron ions move quickly toward the nearest C3 channel from their starting position, but the Asp and Glu rings do not open and the ions remain blocked in the site (Figures S3 and S4). Clearly, the double protonation of the His residues is crucial to enable an effective mechanism for the release of the iron ions. Notably, at pH 7 the majority of the His side chains are doubly protonated thereby suggesting that the mechanism described at pH 4 could be relevant also at neutral pH. In the final part of the simulation at pH 4, although the iron ion has already left the protein His128(A) closes the gap with Glu134(C) as previously observed (Figure 3). Thus, the Glu134 distances increase and the channel opens again (Figure 4C and movie S3 at about 100 ns) despite the absence of the metal ion.

As mentioned, during our simulation at pH 4 we observed two iron ions being released through two different channels in the bulk solution. In the second channel, the entry of the iron occurs after a few ns and all three Glu134 side chains are recruited to form the coordination sphere, which remains stable for nearly 40 ns of trajectory (Figure S5). The distance of the three Glu134 side chains from one another are not all identical, with two pairs at about 6 Å and the third (Glu134(A) and Glu134(B)) at 7 Å for nearly 20 ns. Afterward, there is a rearrangement of the site, with the distance between Glu134-(B) and Glu134(C) remaining at 6 Å, whereas the other two distances increase up to 8 Å until 40 ns. Shortly after 40 ns of simulation, Glu134(B) rearranges similarly to what we described for Glu134(C) in the first channel with the exception that His136(A) is closer to Glu134(B) than His128(A) (Figure S6). Nevertheless, the simultaneous proximity of both histidinium side chains allows Glu134(B) to move away from the metal ion. After this event, the distance between the side chain of Glu134(A) and the iron ion oscillates rapidly in the range 4.5−8.0 Å. Instead, the iron−Glu134(C) distance remains within 4.5−5.0 Å, with sporadic transitions at longer distances when the iron−Glu134(A) falls at 5.0 Å or less (e.g., see around 80 ns in Figure S5). In summary, within this second channel the iron ion is mostly bound to Glu134(C). Occasionally the iron ion moves from Glu134(C) toward Glu134(A), then moves back. On top of this, Glu134(A) itself is moving back and forth with respect to the other two glutamates. These extensive structural fluctuations correspond to a situation where the Cδ atoms of the three Glu residues of the C3 site have changed their configuration from a roughly triangular distribution in space to a distribution closer to linearity. The escape from this situation is mediated again by the formation of a persistent salt bridge to positively charged His118 residues (Figure S7). This reinforces the link between the motions of Glu134(C) and Glu134(A), and the rest of the protein chain while simultaneously reducing their electrostatic interaction with the iron ion. In this way, the ion can displace more from its position and eventually escapes from the channel.

The umbrella sampling method[23] was applied to measure the free-energy of the iron(II) release process through the C3 channel. In addition to simulations at pH 4 and pH 9, we prepared a chimeric variant of ferritin in which the protonation state at pH 9 was altered by doubly protonating the side chains of the three histidines that play a crucial role in the release process at pH 4: His118, His128, and His136. At the starting point of the umbrella simulations, the metal ion is located inside the ferritin cage. Then, in all the simulations, the

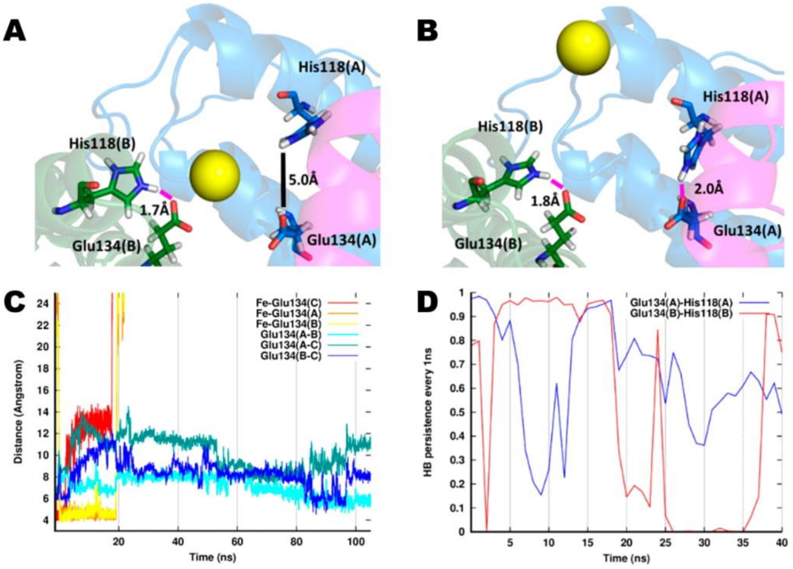


Figure 4. Iron release after the formation of two Glu134−His118 salt bridges in the first channel. Chain A is in blue; chain B is in green. The metal ion is shown as a yellow sphere. (A) Glu134(A) is too far from His118(A) to make a salt bridge, and the iron(II) ion is still coordinated. (B) His118(A) moves in proximity of Glu134(A), and a salt bridge is formed. The electrostatic attraction between the carboxylate Oε oxygen atoms and the metal is interrupted and the iron(II) ion is free to leave the channel. (C) Iron−Glu134 and Glu134−Glu134 distances for one of the three C3 channels populated by iron(II). For Glu residues, all distances refer to the Cδ atoms. Iron(II)−Glu distances are shown using hot colors, Glu−Glu distances are shown using cold colors. The release of iron occurs at about 19 ns (D) The plot shows the persistence of Glu134−His118 hydrogen bonds on a window of 1 ns, normalized by the number of frames in the window. The persistence increases for both hydrogen bonds between 15 and 19 ns of the simulation, i.e. immediately before iron release.

electrostatic gradient pushes rapidly the iron ions inside the C3 channel. Once the metal is inside, the free-energy profile at pH 4 reaches the minimum at 6.2 Å from the innermost part of the channel, shifted by about 4 and 4.5 Å toward the exit with respect to the minima observed for wild-type and chimeric ferritin at pH 9 WT, respectively (Figure 5). In fact, the rings formed by the Asp131 and Glu134 residues (Figures 4C and S1) of the three chains can open dynamically at pH 4, allowing the metal ion to move deeper in the channel (Figure 5A). On the contrary, the iron(II) ion at pH 9 is coordinated closer to the cavity, where negatively charged residues are spatially clustered (Figure 5B). The free-energy minimum for the chimeric ferritin at pH 9 is in a position similar to the WT protein, despite the metal coordination is less compact and some electrostatic interactions with the carboxylates are disrupted by the protonated His residues (Figure 5C). We can conclude that the double protonation of the side chain rings of only the three His residues (His118, His128, and His136) is not sufficient to restore the channel opening observed at acidic pH. Notably, at pH 4 the energy needed to discharge the iron(II) ion or to attract it in the channel are very similar, not exceeding 2 kcal/mol. Instead, for the WT at pH 9 the energetic barrier for the escape from the binding site within the C3 channel is somewhat higher, reaching a plateau of 15.3 kcal/mol around 13.5 Å (Figure S8). Introducing the double protonation for the three key His residues at pH 9 lowers the energy needed to leave the channel to just about 3 kcal/mol.

■ DISCUSSION AND CONCLUSIONS

In summary, at both pH 4 and pH 9 the iron ions are rapidly attracted toward the C3 channels by the electrostatic gradient due to the Glu and Asp side chains that form the C3 site. The Asp residues move apart to allow the metal ion to enter the site, where three Glu side chains are available to coordinate it. This coordination remains stable throughout the simulation at pH 9, whereas at pH 4 we observed for two distinct channels the release of the iron ion into the bulk solution. Overall, the iron release mechanism is very similar for both channels. This process occurs in two main steps (Figure 6): first, the iron ion coordination switches from three to two glutamate side chains (Figure 6A and B). This switch is triggered by the electrostatic attraction between one of the Glu134 side chains and two positively charged histidine residues (His128 and His136). The new orientation of Glu134 is stabilized by His136 through a salt bridge. In the second step, the iron ion is released when its electrostatic interaction with the carboxylate Oε oxygens of the two coordinating Glu134 side chains is interrupted by an increased interaction between these side chains and nearby residues (Figure 6C and D). In this respect, the lifetime of the intrachain Glu134−His118 hydrogen bonds is a useful indicator. Figure 4 suggests that this increase in lifetime is effective for iron release only when it involves at least both Glu residues concurrently. Indeed, the enhanced lifetime of the hydrogen bonds formed by the Glu134 residue from a single chain observed at 5−15 ns in the first channel does not lead to the release of the iron ion. A similar situation occurs in the second channel in the range 65−80 ns. The iron release

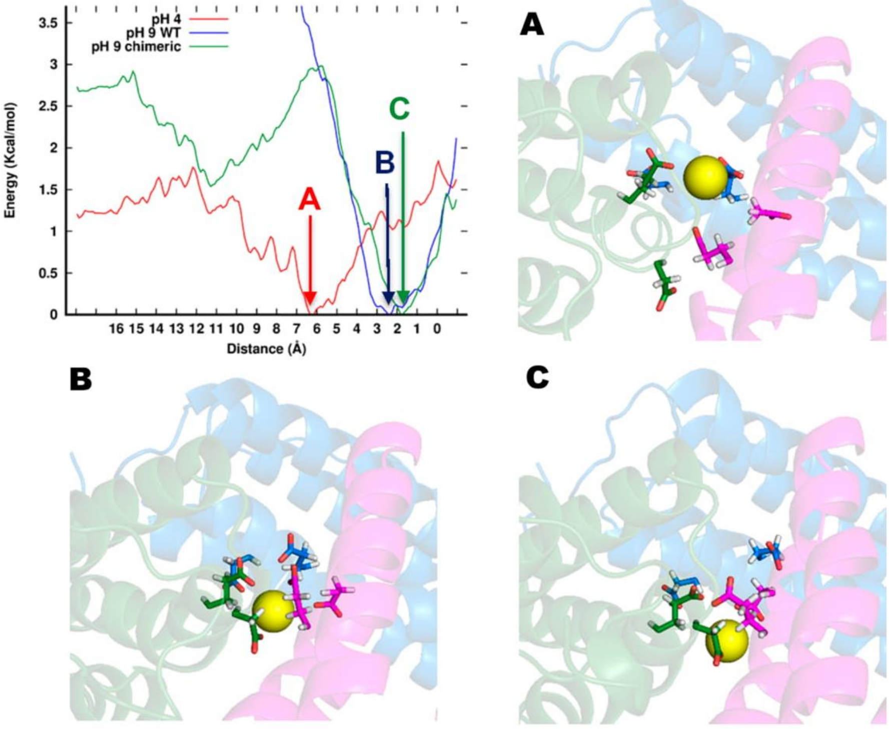


Figure 5. Free-energy profiles along the iron(II) exit coordinate (distances increase from the innermost to the outer part of the channel) through the C3 channel. The free-energy profiles are obtained from the weighted histogram analysis of the umbrella sampling trajectories performed with biased harmonic potential (force constant set at 15 kcal/mol Å^2) for pH 4, pH 9 WT, and pH 9 chimeric (histidine 118, 128, and 136 are doubly protonated as pH 4). The free-energy minima are indicated with arrows, the corresponding metal ion coordinations are shown in pictures A, B, and C. The iron(II) ion is shown as a yellow sphere, Glu134 and Asp131 side chains are shown as sticks, and each chain is colored differently.

mechanism at pH 4 leverages the favorable electrostatics of the doubly protonated His residues enhancing the effectiveness of the stochastic events that bring Glu134 and His118 in proximity for two protein chains simultaneously.

Based on the umbrella sampling simulations we could estimate the free-energy profiles of the initial configurations. Two main things are noteworthy: first, the position of the minima differ by about 4 Å between pH 4 and pH 9, denoting a different accessibility of the channel. Second, the energy barrier is somewhat higher at pH 9 than pH 4. Although all the His residues are doubly protonated at pH 4, three of them are particularly influent in the release process: His118, His128, and His136. We thus leveraged the unique opportunity offered by molecular dynamics simulations to build and study chimeric ferritins to obtain a deeper description of the iron(II) release mechanism. The double protonation of these three His residues introduced in the chimeric form of ferritin at pH 9 could not restore completely the mechanism for the entry of the iron(II) ion in the C3 channel but achieved a nearly full reduction of the energetic barrier for the escape from the binding site into the bulk solution. In conclusion, the positive charges on the side chains of the His118, His128, and His136 residues proved to be crucial to weaken the metal coordination by compensating the negative charges of the Asp131 and Glu134 side chains inside the C3 channel. Instead, only the overall electrostatic contribution of the double protonation of all His at pH 4 could trigger the dynamic opening of the binding site, thus significantly shifting the position of the coordinated iron(II) ion toward the exit of the channel with respect to pH 9. In fact, there is a lower energetic barrier to discharge the iron ion in chimeric than in WT ferritin at pH 9, even though the position of the minimum within the C3 channels remained the same (Figure 5).

At variance with human H-ferritin, the C3 channels are not the entry/exit channels for iron ions in bacterioferritins and bacterial ferritins. These bacterial systems exploit the so-called B-pores instead,[24,25] which are absent in eukaryotic ferritins. For example, in the ferritin of *Pseudomonas aeruginosa* the threefold pores are lined with side chains of alternating charge.[26] Indeed, in the pH 6.0 structure of this system, a negatively charged sulfate ion is found within the C3 channel.

The described pH-dependent mechanism of iron release from ferritin might help defining the role of the acidic environment of endosomes and lysosomes in modulating iron release within these organelles. Encapsulation in endosomes is proposed to be relevant for H-ferritin internalization via the TfR1 receptor[16] and therefore plays a role for the use of ferritin as a drug nanocarrier targeting cancer cells overexpressing TfR1. The delivery of ferritin to lysosomes, instead, has been proposed as a key role in controlling iron cellular homeo-

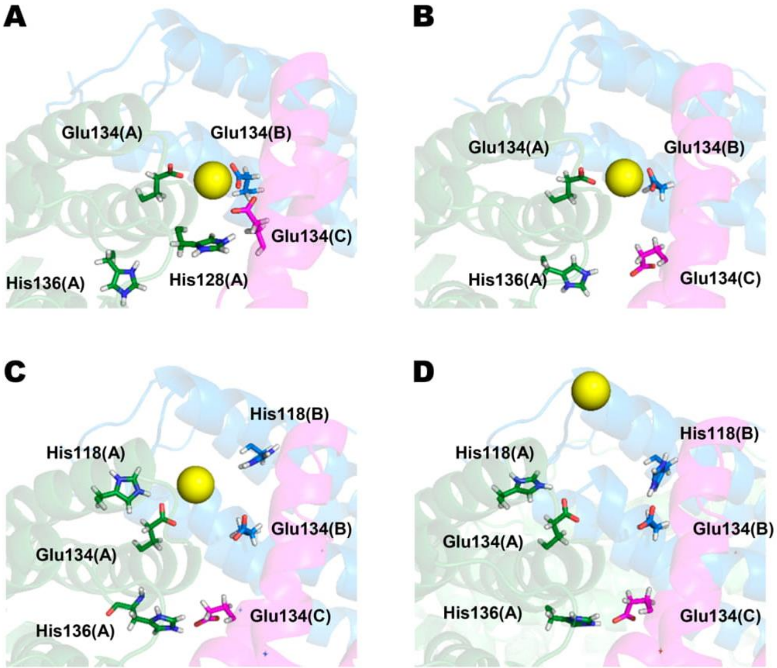


Figure 6. Summary of the crucial steps for iron(II) ion release from C3 channels at pH 4. Each chain is colored differently, and the metal ion is shown as a yellow sphere. Residue side chains involved in the process are shown as sticks. (A) Snapshot at 3.00 ns. The ion is coordinated by three Glu134 residues. His128(A) is approaching the Glu134(C) residue to rip it off from the metal. (B) Snapshot at 3.63 ns. The Glu134(C) is now oriented toward the protein cavity stabilized by His136(A); thus, the ion is now coordinated just by two Glu134 residues. (C) Snapshot at 19.05 ns. During the five nanoseconds before the ion release the interactions between Glu134 and His118 increase, disrupting the metal−carboxylate coordination. (D) Snapshot at 19.24 ns. The simultaneous formation of salt bridges between Glu134 and His118 causes the iron ion release from the channel.

stasis;[15] the interplay between biomineral solubilization upon iron reduction and ferritin degradation at this level is not yet clear, and the observed cage-assisted release at low pH might also come into play.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acs.jcim.7b00306.

Experimental and computational methods and additional figures (PDF)

Movie of a C3 channel, along the whole trajectory, from the protein's cage point of view (MPG)

Movie of the iron(II) ion leaving cavity through the first C3 channel (MPG)

Movie of the first C3 channel, along the whole trajectory, from the protein's cage point of view (MPG)

■ AUTHOR INFORMATION

Corresponding Author

*Tel.: +39 055 4574267. E-mail: rosato@cerm.unifi.it (A.R.).

ORCID

Paola Turano: 0000-0002-7683-8614

Antonio Rosato: 0000-0001-6172-0368

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

This work was supported by Ministero dell'Istruzione, dell'Università e della Ricerca, PRIN 2012, Prot. 2012SK7ASN. S.C. acknowledges a postdoctoral grant by Fondazione Cassa di Risparmio di Firenze (no. 2013.0494) provided by FiorGen.

■ ABBREVIATIONS

MD, molecular dynamics; C3, channel formed by three chains; C4, channel formed by four chains; TfR1, transferrin receptor 1; PDB, protein data bank; WT, wild type

■ REFERENCES

(1) Zhang, Y.; Orner, B. P. Self-Assembly in the Ferritin Nano-Cage Protein Superfamily. *Int. J. Mol. Sci.* **2011**, *12* (12), 5406−5421.

(2) Bernacchioni, C.; Ghini, V.; Theil, E. C.; Turano, P. Modulating the Permeability of Ferritin Channels. *RSC Adv.* **2016**, *6* (25), 21219−21227.

(3) Chandramouli, B.; Bernacchioni, C.; Di Maio, D.; Turano, P.; Brancato, G. Electrostatic and Structural Bases of Fe2+ Translocation through Ferritin Channels. *J. Biol. Chem.* **2016**, *291* (49), 25617−25628.

(4) Theil, E. C.; Behera, R. K.; Tosha, T. Ferritins for Chemistry and for Life. *Coord. Chem. Rev.* **2013**, *257* (2), 579−586.

(5) Crichton, R. R.; Declercq, J.-P. X-Ray Structures of Ferritins and Related Proteins. *Biochim. Biophys. Acta, Gen. Subj.* **2010**, *1800* (8), 706−718.

(6) Pozzi, C.; Di Pisa, F.; Bernacchioni, C.; Ciambellotti, S.; Turano, P.; Mangani, S. Iron Binding to Human Heavy-Chain Ferritin. *Acta Crystallogr., Sect. D: Biol. Crystallogr.* **2015**, *71* (9), 1909−1920.

(7) Linder, M. C. Mobilization of Stored Iron in Mammals: A Review. *Nutrients*; Multidisciplinary Digital Publishing Institute, October 10, 2013; pp 4022−4050.

(8) Theil, E. C. Ferritin: The Protein Nanocage and Iron Biomineral in Health and in Disease. *Inorg. Chem.* **2013**, *52* (21), 12223−12233.

(9) Finazzi, D.; Arosio, P. Biology of Ferritin in Mammals: An Update on Iron Storage, Oxidative Damage and Neurodegeneration. *Arch. Toxicol.* **2014**, *88* (10), 1787−1802.

(10) Melman, G.; Bou-Abdallah, F.; Vane, E.; Maura, P.; Arosio, P.; Melman, A. Iron Release from Ferritin by Flavin Nucleotides. *Biochim. Biophys. Acta, Gen. Subj.* **2013**, *1830* (10), 4669−4674.

(11) Hynes, M. J.; O Coinceanainn, M. M. Investigation of the Release of Iron from Ferritin by Naturally Occurring Antioxidants. *J. Inorg. Biochem.* **2002**, *90* (1), 18−21.

(12) Kidane, T. Z. Release of Iron from Ferritin Requires Lysosomal Activity. *AJP Cell Physiol.* **2006**, *291* (3), C445−C455.

(13) Lewis, H. A.; Buchanan, S. G.; Burley, S. K.; Conners, K.; Dickey, M.; Dorwart, M.; Fowler, R.; Gao, X.; Guggino, W. B.; Hendrickson, W. A.; Hunt, J. F.; Kearins, M. C.; Lorimer, D.; Maloney, P. C.; Post, K. W.; Rajashankar, K. R.; Rutter, M. E.; Sauder, J. M.; Shriver, S.; Thibodeau, P. H.; Thomas, P. J.; Zhang, M.; Zhao, X.; Emtage, S. Structure of Nucleotide-Binding Domain 1 of the Cystic Fibrosis Transmembrane Conductance Regulator. *EMBO J.* **2004**, *23* (2), 282−293.

(14) Mancias, J. D.; Wang, X.; Gygi, S. P.; Harper, J. W.; Kimmelman, A. C. Quantitative Proteomics Identifies NCOA4 as the Cargo Receptor Mediating Ferritinophagy. *Nature* **2014**, *509* (7498), 105−109.

(15) Asano, T.; Komatsu, M.; Yamaguchi-Iwai, Y.; Ishikawa, F.; Mizushima, N.; Iwai, K. Distinct Mechanisms of Ferritin Delivery to Lysosomes in Iron-Depleted and Iron-Replete Cells. *Mol. Cell. Biol.* **2011**, *31* (10), 2040−2052.

(16) Li, L.; Fang, C. J.; Ryan, J. C.; Niemi, E. C.; Lebrón, J. A.; Björkman, P. J.; Arase, H.; Torti, F. M.; Torti, S. V.; Nakamura, M. C.; Seaman, W. E. Binding and Uptake of H-Ferritin Are Mediated by Human Transferrin Receptor-1. *Proc. Natl. Acad. Sci. U. S. A.* **2010**, *107* (8), 3505−3510.

(17) Yamashiro, D. J.; Maxfield, F. R. Acidification of Morphologically Distinct Endosomes in Mutant and Wild-Type Chinese Hamster Ovary Cells. *J. Cell Biol.* **1987**, *105* (6), 2723−2733.

(18) Maxfield, F. R.; Yamashiro, D. J. Endosome Acidification and the Pathways of Receptor-Mediated Endocytosis. *Adv. Exp. Med. Biol.* **1987**, *225*, 189−198.

(19) Tosha, T.; Behera, R. K.; Ng, H. L.; Bhattasali, O.; Alber, T.; Theil, E. C. Ferritin Protein Nanocage Ion Channels: Gating by N-Terminal Extensions. *J. Biol. Chem.* **2012**, *287* (16), 13016−13025.

(20) Bradley, J. M.; Le Brun, N. E.; Moore, G. R. Ferritins: Furnishing Proteins with Iron. *J. Biol. Inorg. Chem.* **2016**, *21*, 13−28.

(21) Haldar, S.; Bevers, L. E.; Tosha, T.; Theil, E. C. Moving Iron through Ferritin Protein Nanocages Depends on Residues throughout Each Four -Helix Bundle Subunit. *J. Biol. Chem.* **2011**, *286* (29), 25620−25627.

(22) Bertini, I.; Lalli, D.; Mangani, S.; Pozzi, C.; Rosa, C.; Theil, E. C.; Turano, P. Structural Insights into the Ferroxidase Site of Ferritins from Higher Eukaryotes. *J. Am. Chem. Soc.* **2012**, *134* (14), 6169−6176.

(23) Roux, B. The Calculation of the Potential of Mean Force Using Computer Simulations. *Comput. Phys. Commun.* **1995**, *91* (1−3), 275−282.

(24) Rui, H.; Rivera, M.; Im, W. Protein Dynamics and Ion Traffic in Bacterioferritin. *Biochemistry* **2012**, *51* (49), 9900−9910.

(25) Yao, H.; Rui, H.; Kumar, R.; Eshelman, K.; Lovell, S.; Battaile, K. P.; Im, W.; Rivera, M. Concerted Motions Networking Pores and Distant Ferroxidase Centers Enable Bacterioferritin Function and Iron Traffic. *Biochemistry* **2015**, *54* (8), 1611−1627.

(26) Yao, H.; Jepkorir, G.; Lovell, S.; Nama, P. V.; Weeratunga, S.; Battaile, K. P.; Rivera, M. Two Distinct Ferritin-like Molecules in *Pseudomonas Aeruginosa*: The Product of the *bfrA* Gene Is a Bacterial Ferritin (FtnA) and Not a Bacterioferritin (Bfr). *Biochemistry* **2011**, *50* (23), 5236−5248.

Supporting Information

Investigation of the iron(II) release mechanism of human H-ferritin as a function of pH

Davide Sala[†], *Silvia Ciambellotti*[†], *Andrea Giachetti*[‡], *Paola Turano*[†,§] *and Antonio Rosato*[†,§,]*

[†]Magnetic Resonance Center (CERM), University of Florence, Via Luigi Sacconi 6, 50019 Sesto Fiorentino, Italy.

[‡]Consorzio Interuniversitario di Risonanze Magnetiche di Metallo Proteine, Via Luigi Sacconi 6, 50019 Sesto Fiorentino, Italy.

[§]Department of Chemistry, University of Florence, Via della Lastruccia 3, 50019 Sesto Fiorentino, Italy.

*rosato@cerm.unifi.it

MATERIALS AND METHODS

1. Kinetic Measurements

Recombinant human H-chain ferritin was made apo (without metal ions) as previously described[1,2] and mineralized (480 Fe^{2+} ions per cage) with a freshly prepared solution of ferrous sulphate in 1 mM HCl in 100 mM MOPS, 100 mM NaCl, pH 7.0. After mixing, the solutions were incubated for 2 h at room temperature and then overnight at 4 °C to complete the iron mineralization reaction. In order to investigate the pH-dependence of the release process, mineralized ferritin was treated with PD-10 desalting columns (GE Healthcare) to be exchanged into 100 mM sodium acetate pH 4.0 and 100 mM MOPS, 100 mM NaCl, pH 9.0 buffers. The sample at pH 7 was also treated with PD-10 column to remove the excess of iron not biomineralized. The release of iron(II) from caged ferritin minerals was initiated by reducing the mineral with added NADH (2.5 mM) and FMN (2.5 mM) or with sodium dithionite (2.5 mM, dissolved with degassed water under anaerobic condition until the addition) and trapping the reduced and dissolved iron(II) with the specific chelator 2,2'-bipyridyl outside the protein cage. All release experiments were performed in aerobic conditions. Iron(II) release from the protein cage was measured as the absorbance of the $[Fe(2,2'\text{-bipyridyl})_3]^{2+}$ complex at the maximum of A_{522} nm and the amount of complexed iron(II) extrapolated using the molar extinction coefficient of the complex (8,430 M^{-1} cm^{-1}). The experiments were performed at iron and protein cage concentrations of 250 µM and 0.52 µM, respectively.

2. Computational Methods

We used the Modeller software[3] (mod9v2) to model the 24-mer human ferritin (4134 residues) based on the crystal structure of the bullfrog M ferritin[4] (PDB 4DAS). Bullfrog and human heavy chains have 68% and 90% of sequence identity and similarity, respectively. Using the H++ server[5] (version 3.2) the protonation state of titratable groups was calculated at pH 4 and pH 9. For this, we used different conformations extracted from a short molecular dynamics simulation at pH 7, which were all submitted to the H++ server. In all cases, the server indicated that all the His side chains are doubly protonated as pH 4. On the contrary, only a few Glu or Asp residues on the external surface of the protein close to the C4 channel are protonated at this pH. These residues were different in the different conformations. For the sake of simplicity and considering that they are far from the C3 channel, we decided to omit the protonation of Asp and Glu residues. The H++ server predicts that the configuration in which all His are doubly protonated and no Asp or Glu is protonated occurs at pH 4.5, i.e. the lowest pH within lysosomes. This configuration is thus physiologically relevant for human H-ferritin.

Subsequently, 31 iron(II) ions were added inside the protein's cavity. The system was solvated with TIP3P water model molecules in a truncated octahedron box with walls distant 9 Å from the solute. The non-bonded parameters for iron(II) were taken in agreement with the use of TIP3P water and of the classical 12-6 LJ non-bonded model[6]. In particular, the Compromise set (CM) of parameters for divalent ions were taken (available from the frcmod.ions234lm_126_tip3p file of the AMBER16 package). The system at pH 4 was balanced adding 110 chloride ions for a total amount of 161,784 atoms. At pH 9 106 sodium ions were added for 161,507 total atoms. The two systems were prepared independently using the customized force-field AMBER99SB-ILDN* on the AMBER16 Molecular Dynamics Package[7]. This forcefield merges the ILDN correction[8] of AMBER99-ILDN force-field with the CMAP dihedral parameters improvement of CHARMM22* force-field[9]. The same simulation protocol was applied for both systems using the pmemd software. At the first stage the water was minimized to remove bad contacts, after a brief MD with the protein restrained, all the system was minimized using Steepest Descent algorithm followed by Conjugate Gradient. The temperature was set to 310 K and controlled by Langevin thermostat with a collision frequency of 20 ps^{-1}. The system was heated linearly for 1 ns in constant volume. Thus, to adjust the system to the correct density we performed a 500 ps MD in NPT condition using Monte-Carlo barostat. The production run was carried out using the following parameters: integration step of 2 fs and constant temperature of 310 K under the control of weak coupling algorithm every 10 ps in NPT conditions controlled by Monte Carlo barostat. Covalently bonded hydrogen atoms were constrained with the SHAKE and SETTLE (for water molecules) algorithms. Finally, the PME method with a cutoff of 8 Å was applied to calculate electrostatic interactions. One snapshot every 1 ps was saved for a total production run of 105 ns on Nvidia Tesla K20m GPGPU. The electrostatic potential was computed using the Adaptive Poisson-Boltzmann Solver software[10].

We carried out umbrella sampling simulations[11] using harmonic biased potential along the iron(II) release process through C3 channels. The force constant was set to 15 kcal/mol/Å^2 with a window spacing of 0.1 Å. At the starting point of the simulations, the metal ion is located inside the ferritin cage, 25 Å away from a water molecule fixed by tight distance restraints to face the exit of the C3 channel. In particular, six distance restraints to the Asn23 and Asn109 α-carbon of each chain forming the C3 channel were applied, using a harmonic potential with a force constant of 20 kcal/mol/Å^2. We chose these residues because they are in a stable position in the middle of two long α-helices. The protocol applied for the production runs was the same as described above, each replica was sampled for 500 ps. The resulting umbrella sampling trajectories were analyzed using the weighted histogram analysis method (WHAM)[12] version 2.0.9 to obtain the free energy profiles of the metal ion release at pH 4 and pH 9.

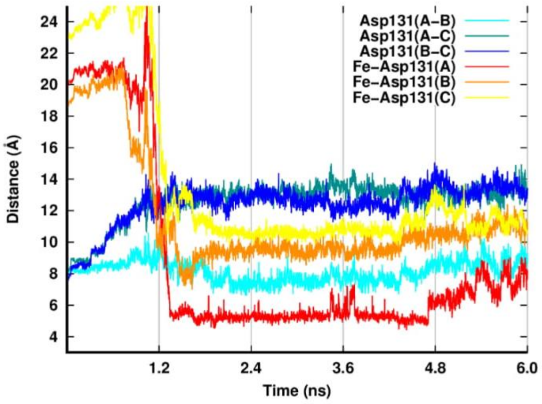


Figure S1 Iron(II)-Asp131 and Asp131-Asp131 distances for the first channel at pH 4. All distances refer to the Cγ atoms.

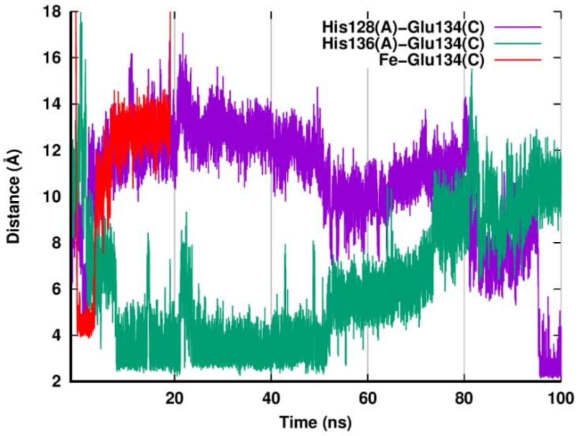


Figure S2 Iron(II)-Glu134(C), His128(A)-Glu134(C), and His136(A)–Glu134(C) distances of the first channel at pH 4 along the whole trajectory.

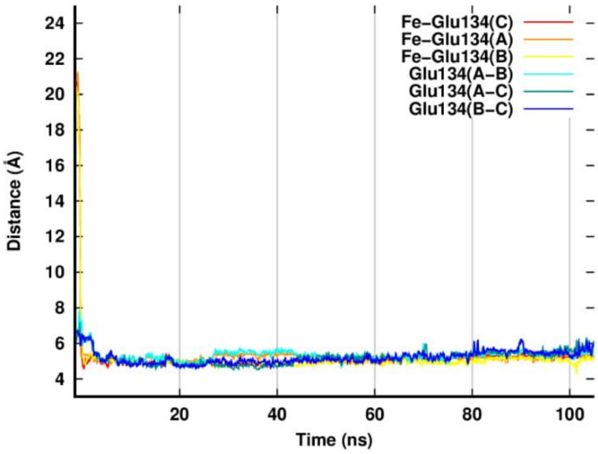


Figure S3 Iron(II)-Glu134 and Glu134-Glu134 distances at pH 9**.** For Glu residues, all distances refer to the Cδ atoms. Iron-Glu distances are shown using hot colors; Glu-Glu distances are shown using cold colors. Essentially, at pH 9 the same profiles are observed for all the four channels populated by iron(II).

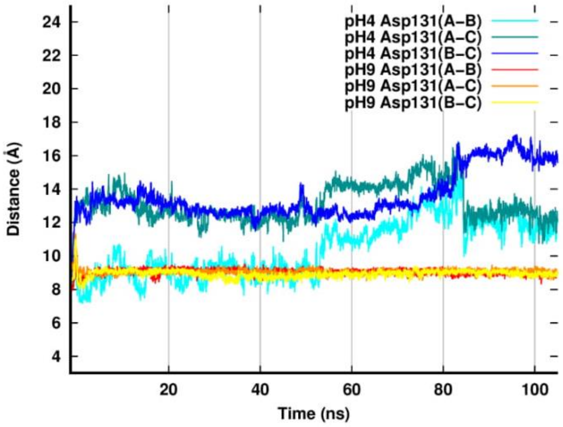


Figure S4 Asp131-Asp131 distances at pH 4 and pH 9 for the same channel.

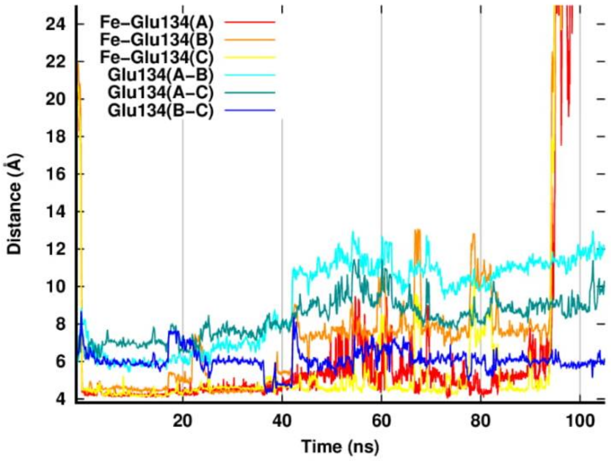


Figure S5 Iron(II)-Glu134 and Glu134-Glu134 distances for the second channel at pH 4. For Glu residues, all distances refer to the Cδ atoms. Iron-Glu distances are shown using hot colors; Glu-Glu distances are shown using cold colors. The release of iron occurs at about 95 ns.

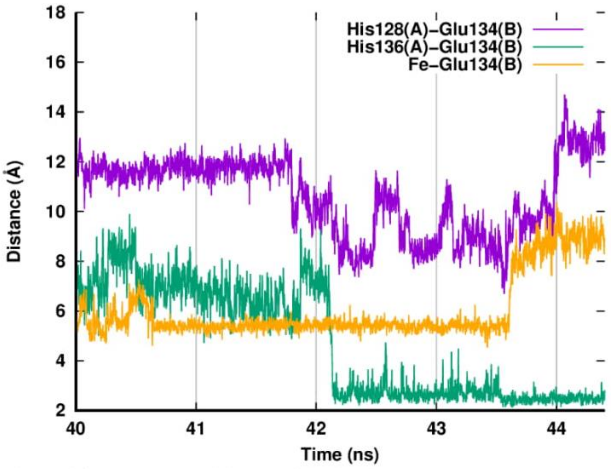


Figure S6 Iron(II)-Glu134, His128(A)-Glu134(B), and His136(A)–Glu134(B) distances of the second channel.

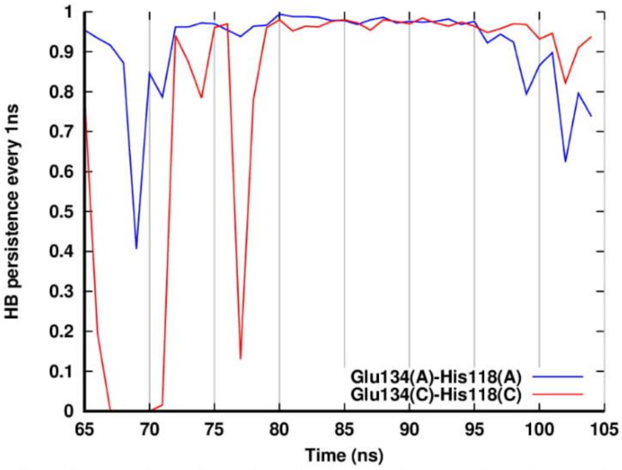


Figure S7 The plot shows the Glu134-His118 hydrogen bonds persistence on a window of 1 ns, the number is normalized over the number of frames in the window. The persistency increases for both hydrogen bonds between 80 and 95 ns of the simulation, i.e. concomitantly with iron release.

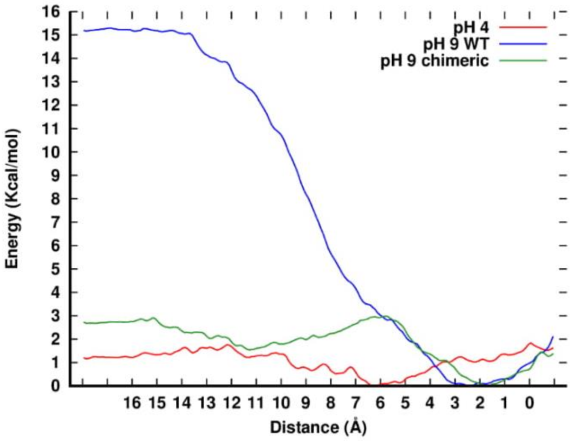


Figure S8 Free-energy profiles along the iron(II) release process (distances are calculated from the innermost to the outer part of the channel) through the C3 channel. The free-energy profiles are obtained from the weighted histogram analysis of the umbrella sampling trajectories performed with biased harmonic potential at pH 4 and pH 9.

MOVIES

Movie S1 Movie of a C3 channel, along the whole trajectory, from the protein cage point of view. Three iron(II) ions are shown. The first one goes inside the channel at the center and leaves the protein. The others two ions are attracted toward two different catalytic sites. The bottom right metal ion is very stable inside the site, the other one on top is transient.

Movie S2 Movie of the iron(II) ion leaving cavity through the first C3 channel. The protein is sectioned to allow the view of the iron inside the channel. Negative electrostatic is shown in red, positive one in blue. Fieldlines represent the electrostatic gradient from positive to negative. The metal ion moves very quickly toward the channel led by a strong electrostatic gradient. When the iron ion reaches the top of the channel, a transient electrostatic attraction, due to Asp residues sidechains, pulls it outside, in the bulk solution.

Movie S3 Movie of the first C3 channel, along the whole trajectory, from the protein cage point of view. Besides the iron ion insertion, it is possible to appreciate the channel closing event followed by a reopening in absence of a metal ion.

REFERENCES

(1) Pozzi, C.; Di Pisa, F.; Bernacchioni, C.; Ciambellotti, S.; Turano, P.; Mangani, S. Iron Binding to Human Heavy-Chain Ferritin. *Acta Crystallogr. Sect. D Biol. Crystallogr.* **2015**, *71*, 1909–1920.

(2) Ravera, E.; Ciambellotti, S.; Cerofolini, L.; Martelli, T.; Kozyreva, T.; Bernacchioni, C.; Giuntini, S.; Fragai, M.; Turano, P.; Luchinat, C. Solid-State NMR of PEGylated Proteins. *Angew. Chemie Int. Ed.* **2016**, *55*, 2446–2449.

(3) Eswar, N.; Webb, B.; Marti-Renom, M. A.; Madhusudhan, M. S.; Eramian, D.; Shen, M.; Pieper, U.; Sali, A. Comparative Protein Structure Modeling Using MODELLER. In *Current Protocols in Protein Science*; John Wiley & Sons, Inc.: Hoboken, NJ, USA, 2007; Vol. Chapter 2, p 2.9.1-2.9.31.

(4) Bertini, I.; Lalli, D.; Mangani, S.; Pozzi, C.; Rosa, C.; Elizabeth, C.; Turano, P. Structural Insight into the Ferroxidase Site of Eukaryotic Ferritins. **2012**, 1–6.

(5) Anandakrishnan, R.; Aguilar, B.; Onufriev, A. V. H++ 3.0: Automating pK Prediction and the Preparation of Biomolecular Structures for Atomistic Molecular Modeling and Simulations. *Nucleic Acids Res.* **2012**, *40*, W537–W541.

(6) Li, P.; Roberts, B. P.; Chakravorty, D. K.; Merz, K. M. Rational Design of Particle Mesh Ewald Compatible Lennard-Jones Parameters for +2 Metal Cations in Explicit Solvent. *J. Chem. Theory Comput.* **2013**, *9*, 2733–2748.

(7) Case, D. A.; Cheatham, T. E.; Darden, T.; Gohlke, H.; Luo, R.; Merz, K. M.; Onufriev, A.; Simmerling, C.; Wang, B.; Woods, R. J. The Amber Biomolecular Simulation Programs. *J. Comput. Chem.* **2005**, *26*, 1668–1688.

(8) Lindorff-Larsen, K.; Piana, S.; Palmo, K.; Maragakis, P.; Klepeis, J. L.; Dror, R. O.; shaw, D. E. Improved Side-Chain Torsion Potentials for the Amber ff99SB Protein Force Field.

PROTEINS-STRUCTURE Funct. Bioinforma. **2010**, *78*, 1950–1958.

(9) Buck, M.; Bouguet-Bonnet, S.; Pastor, R. W.; MacKerell, A. D.; Jr. Importance of the CMAP Correction to the CHARMM22 Protein Force Field: Dynamics of Hen Lysozyme. *Biophys. J.* **2006**, *90*, L36-8.

(10) Baker, N. A.; Sept, D.; Joseph, S.; Holst, M. J.; McCammon, J. A. Electrostatics of Nanosystems: Application to Microtubules and the Ribosome. *Proc. Natl. Acad. Sci. U. S. A.* **2001**, *98*, 10037–10041.

(11) Roux, B. The Calculation of the Potential of Mean Force Using Computer Simulations. *Comput. Phys. Commun.* **1995**, *91*, 275–282.

(12) Grossfield, A. “WHAM: the weighted histogram analysis method”, version 2.0.9 http://membrane.urmc.rochester.edu/content/wham.

3.3.3 An atomistic view of the YiiP structural changes upon zinc(II) binding

Davide Sala[1], Andrea Giachetti[2] and Antonio Rosato[1,3,]*

[1]Magnetic Resonance Center (CERM) – University of Florence, Via L. Sacconi 6, 50019 Sesto Fiorentino, Italy

[2]Consorzio Interuniversitario di Risonanze Magnetiche di Metallo Proteine, Via Luigi Sacconi 6, 50019 Sesto Fiorentino, Italy.

[3]Department of Chemistry – University of Florence, Via della Lastruccia 3, 50019 Sesto Fiorentino, Italy

Biochim Biophys Acta Gen Subj. 2019 Oct;1863(10):1560-1567. Epub 2019 Jun 6

Contents lists available at ScienceDirect

BBA - General Subjects

journal homepage: www.elsevier.com/locate/bbagen

An atomistic view of the YiiP structural changes upon zinc(II) binding

Davide Sala[a], Andrea Giachetti[b], Antonio Rosato[a,b,c,*]

[a] Magnetic Resonance Center (CERM), University of Florence, Tuscany, Sesto Fiorentino, Italy
[b] Consorzio Interuniversitario di Risonanze Magnetiche di Metallo Proteine, Tuscany, Sesto Fiorentino, Italy
[c] Department of Chemistry, University of Florence, Tuscany, Sesto Fiorentino, Italy

ARTICLE INFO

Keywords:
Zinc transporter
Metal transport
Metal homeostasis
CDF
Cation diffusion

ABSTRACT

Background: YiiP is a bacterial zinc-for-proton antiporter belonging to the cation diffusion facilitator family. The zinc(II) ions are transported across the cell membrane, from the cytosol to the extracellular space.

Methods: We performed atomistic molecular dynamics simulations of the YiiP dimer with zinc(II) ions in solution to elucidate how the metal ions interact with the protein while moving from the cytosol to the transport site.

Results: We observed that of the two cavities of the dimer, only one was accessible from the cytosol during transport. Zinc(II) binding to D49 of the transport site triggered a rearrangement of the transmembrane domain that closed the accessible cavity. Finally, we analyzed the free-energy profiles of metal transit in the channel and observed the existence of a high barrier preventing release from the transport site.

Conclusions: The observed dynamics is consistent with the dimer-dimer interface forming a stable scaffold against which the rest of the trans-membrane rearranges.

General significance: Zinc(II) transporters are present in all kingdoms of life. The present study highlights structural features that might be of general relevance.

1. Introduction

Zinc is the second most abundant trace element in all living organisms and plays a fundamental role in many biological processes [1,2]. Despite sufficient intracellular zinc(II) concentration is needed to supply proteins that require zinc(II) ions as cofactor, intracellular concentrations higher than few hundred picomolar produce cytotoxic effects [3,4]. Therefore, the cellular uptake and the efflux of zinc(II) ions are tightly controlled by specific systems, present at all phylogenetic levels, to maintain its concentration in a quite narrow range [5–7]. In mammals, the Zrt-, Irt-related proteins (ZIPs/Slc39s) are responsible of the uptake of zinc(II) ions into the cell [8,9]. On the contrary, the zinc transporters of the Cation Diffusion Facilitator (CDF) superfamily (ZnTs/Slc30) mediate zinc(II) removal from the cytoplasm into the extracellular space or into intracellular compartments such as secretory vesicles [10–14]. Presently, there is no direct experimental information on the three-dimensional structure of any member of the human ZnT family [15,16]. However, this limit is partially compensated by a relatively well-characterized bacterial homolog, the YiiP transporter from *Escherichia coli* [17,18]. Among the human homologs of YiiP, Znt8 has received significant attention owing to the involvement of its W325R common variant in the onset of type-1 and type-2 diabetes [19–22].

YiiP is a cation-proton antiporter catalysing the efflux of zinc(II) against the uptake of a proton with a 1:1 exchange stoichiometry [23]. YiiP can transport several divalent metal cations but only zinc(II) and cadmium(II) with high efficiency [24]. The first X-ray structure of this transporter was solved in the outward-facing conformation [25]; in 2009 a new structure with better resolution became available (PDB ID: 3H90) [26]. These crystal structures featured an architecture shared among all CDF transporters consisting in a transmembrane domain (TMD) connected to a C-terminal, cytoplasmic domain (CTD). YiiP is a Y-shaped homodimer with six TM helices per monomer clustered in a four-helix bundle (TM1–TM2–TM4–TM5) and a helix pair (TM3–TM6). The TM3-TM6 helix pair provides the dimer interface and protrudes in the cytosolic region. Each YiiP protein chain harbours three zinc(II) binding sites (A-C). Site A is located close to the middle of the TMD and constitutes the tetrahedral zinc(II) transport site, involving D45 and D49 of TM2 as well as H153 and D157 of TM5 [24]. The CTD hosts the C sites in which the binding of two zinc(II) ions enhances the stability of

Abbreviations: CDF, Cation Diffusion Facilitator; TMD, transmembrane domain; CTD, cytosolic domain; EC, extracellular; IC, intracellular; TM, transmembrane; PCA, principal component analysis

* Corresponding author at: Magnetic Resonance Center (CERM), University of Florence, Tuscany, Sesto Fiorentino, Italy.
E-mail address: rosato@cerm.unifi.it (A. Rosato).

https://doi.org/10.1016/j.bbagen.2019.06.001
Received 19 February 2019; Received in revised form 28 May 2019; Accepted 3 June 2019
Available online 06 June 2019

the dimer [26]. The function of site B is unclear. The inward-facing state of the YiiP homolog from *Shewanella oneidensis* (PDB ID: 3J1Z) was solved at 13 Å by cryo-electron microscopy [27]. Based on the comparison of all available structures, it has been proposed that the transport of zinc(II) involves a so-called alternating mechanism [28]. This proposed mechanism involves an allosteric connection between sites A and C through which a zinc(II) ion binding in the C site can promote metal binding and transport to the TMD thanks to the reorientation of the TM3-TM6 helix pair [26]. In turn, the binding of the zinc(II) ion in the transport site A triggers the conformational switch from the inward- to the outward-facing state facilitating the release of zinc(II). Recently, a variation of the alternating mechanism has been proposed in which the motion of the four-helix bundle relative to the static TM3-TM6 scaffold is proposed to be the critical step for the zinc (II) transfer across the membrane [29]. This mechanism relies on a persistent dimer interface within the lipid membrane acting as a static scaffold versus a mobile transport domain; a similar strategy is used also by other secondary active transporters [30]. In this regard, the zinc (II) translocation to the transport site depends on the accessibility of the hydrophobic cavity located in the TMD region facing the intracellular space [31].

In this work, we investigated the zinc-induced dynamics of YiiP from *Escherichia coli* by performing atomistic MD simulations of the inward-facing conformation in presence and without zinc(II) ions in solution, respectively. Our results pointed out how the binding of one zinc(II) ion in the transport site triggers the closing of the portion of the channel going from the cytosolic space to the transport site itself, hereafter called the TM cavity. The TM cavity closing was mainly the result of TM4-TM5 of the four-helix bundle moving towards TM3-TM6. In this regard, TM5 played a crucial role through a reorientation that brought the side chains of L152 and M197 in close contact, thus forming a hydrophobic gate as proposed by Gupta et al. Unexpectedly, we observed that zinc(II) binding triggered the selective opening of only one channel out of the two protein chains in the dimer. Our work provides a detailed view of the mechanism of zinc(II) permeation from the cytosol to the TM cavity and gives hints on the subsequent steps of the transport mechanism.

2. Methods

The dimeric model of the inward-facing state of YiiP from *Escherichia coli* was built with the Modeller software [32] (mod9v2) exploiting two different structural templates. Since the inward- and the outward-facing state share the same conformation of the CTDs, we used the structure with the highest resolution available (PDB ID: 3H90) [26] to model the residues in the range 208–290. Instead, the inward-facing TMDs, spanning residues 7-207, were modelled on the cryo-EM structure of the homologous *Shewanella oneidensis* protein (PDB ID: 3J1Z) [27]. The membrane builder module of the CHARMM-GUI was used to embed the protein in a rectangular lipid bilayer composed by 366 DPPC phospholipids [33,34]. The system was solvated with TIP3P water molecules with a hydration ratio of 1:100 lipid to water. Two dummy atoms were bound with harmonic restraints to the two cytosolic binding sites (site C) to mimic the binuclear zinc coordination mediating the interaction between the CTDs [12]. All the histidine residues were kept neutral. The Nδ-protonated tautomer was used for all the histidine residues of the zinc(II) binding sites that coordinate the metal with the Nε atom in the 3H90 crystal structure. Instead, the Nε-protonated tautomer was used for all the remaining histidine side chains.

The holo-simulations were performed in the presence of 0.150 M $ZnCl_2$ corresponding to 94 zinc(II) ions and 174 (188 minus 14 ions to balance the negative charge of the protein) chloride(I) ions in solution. The non-bonded parameters for zinc(II) were taken in agreement with the use of TIP3P water model in combination with the classical 12-6 LJ non-bonded model [35]. In particular, the Compromise set (CM) of parameters for divalent ions was used (available from the frcmod.ions234lm_126_tip3p file of the AMBER18 package).

The No-zinc simulation was prepared by adding 0.150 M of NaCl corresponding to 104 sodium(I) ions and 90 chloride(I) ions in solution. A harmonic potential was applied to prevent the insertion of sodium(I) ions in the TM cavities.

The E79A simulation was prepared starting from the snapshot of the Holo02 run at 50 ns, when the zinc(II) was already bound to D150 and E79. After introducing the E79A mutation, a chloride(I) ion was added to the system to balance the charge.

All the simulations exploited the same force fields for the lipid and the protein portions of the system, the Amber LIPID17 and the Amber ff14SB, respectively [36,37]. All the calculations were performed with the AMBER Molecular Dynamics Package using the pmemd software [38,39], with the same equilibration protocol. An initial minimization step was carried out using the Steepest Descent algorithm followed by Conjugate Gradient. Langevin dynamics with a collision frequency of 1 ps^{-1} was used to linearly heat the system in constant volume for 1 ns, during which the protein and the ions are restrained with a force constant of 10 kcal/mol/Å^2. The equilibrium temperature was set to 323 K, a value commonly used to overcome the phase transition (liquid to gel) temperature of the pure DPPC lipid membrane (about 315 K) [40]. The physiological density of the system was achieved by carrying out an NPT simulation in which the anisotropic pressure scaling is controlled by the Berendsen barostat (pressure relaxation time of 2 ps). Covalently bonded hydrogen atoms were constrained with the SHAKE and SETTLE (for water molecules) algorithms. The PME method with a cutoff of 10 Å was applied to compute electrostatic interactions. The production runs were performed on Nvidia Pascal Xp GPGPU in the same NVT conditions reported above for the heating step [41]. The RMSD of the protein Cα atoms from the starting conformation was calculated over time to check the convergence of the simulations (Fig. S1).

We carried out umbrella sampling simulations using a harmonic biased potential along the zinc(II) diffusion pathway through the TM cavity of the channel [42]. We prepared two runs using a zinc(II) ion as probe and a third run using a sodium(I) ion. The starting structure of the simulations was chosen from the unbiased holo-trajectory when the zinc(II) ion is at the entrance of the TM cavity facing the IC space. In the calculation using the sodium(I) probe, the metal ion was moved some angstroms away from the channel entrance in order to probe also the energetics for its approach to the channel. The metal probes were pulled towards a water molecule fixed by tight distance restraints outside the top exit of the channel. In this way the ion probe is forced to cross the whole TM portion of the transporter. The water molecule was restrained to the Cα of six residues (L22, L58, L95, L131, L152 and L191) located in a stable position in the middle of TM1, TM2, TM3, TM4, TM5 and TM6, respectively. The force constant used to fix the water molecule was 20 kcal/mol/Å^2. Instead, the metal probe was pushed through the channel with a force constant of 15 kcal/mol/Å^2. The total distance was split in windows of 0.2 Å, each one sampled for 1 ns. The resulting trajectories were computed using the weighted histogram analysis method (WHAM) version 2.0.9 to derive the free energy profiles from the potential of mean forces of the metal probes crossing the TM channel [43]. The standard deviations or error bars in the PMF calculations were derived from both bootstrapping and block analysis but are not reported due to their very small values.

All MD trajectories and umbrella sampling simulations are freely available from the Zenodo website at https://zenodo.org/record/2658142 (DOI: https://doi.org/10.5281/zenodo.2658142).

3. Results

Our model of the *E. coli* YiiP dimer in the inward-facing state was built by merging the model of the TMD built on the cryo-EM structure of the *S. oneidensis* homolog with the higher quality CTD obtained from the outward-facing crystal structure of *E. coli* YiiP (Fig. 1A). The CTD is

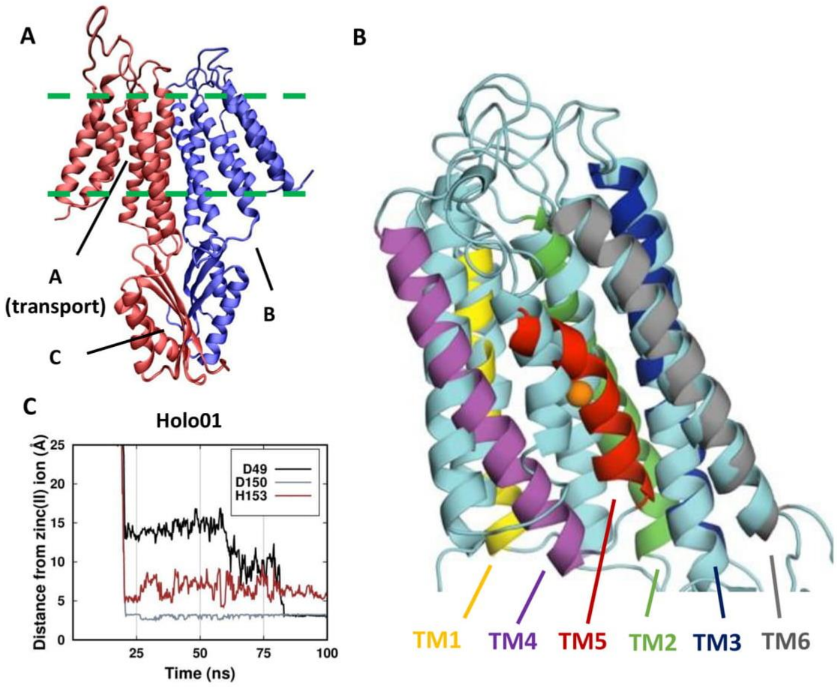


Fig. 1. Model of the *E. coli* YiiP transporter in the inward-facing state. A) The three zinc(II) binding sites are indicated (A-C). The two protein chains are colored differently. B) Fitting of the final snaphot of chain B on the starting conformation in the Holo01 simulation. The starting conformation is in cyan. The zinc(II) ion is shown as an orange sphere. **C)** Zinc(II) interaction pathway in the TM cavity. The distances are measured from the zinc(II) ion to the Cγ atom of D49, D150 and to the Nδ atom of H153. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

structurally identical in both the protein states. The resulting dimeric protein model was embedded in a lipid bilayer and solvated (see Methods). Overall, we carried out four classical MD simulations: three in the presence of 0.150 M of $ZnCl_2$ (holo) and one without zinc(II) ions in solution. The two longest holo-simulations started from the same conformation and sampled a trajectory length of 1 μs each. Hereafter, we will refer to these two runs as "Holo01" and "Holo02". Also the simulation of YiiP in absence of zinc(II) ions ("No-zinc" hereafter) spanned a time length of 1 μs. A further replica of the holo system was performed for 540 ns starting from a snapshot of the "Holo01" run with the zinc(II) ion already bound to the binding site in the channel. We will refer to this replica as "Replica01". Finally, we performed three umbrella simulations using the zinc(II) (2 runs) and sodium(I) ions as probes to assess the metal-dependent free-energy barriers of the channel permeation process.

3.1. Simulations of Holo-YiiP

In the presence of zinc(II) ions in solution, the final structures of the two Holo trajectories at 1 μs featured both sites B on the IL1 loops (Fig. 1A) populated by zinc(II) ions. More importantly, in both runs one zinc(II) ion entered the transmembrane (TM) cavity. The zinc(II) ion entered the channel of chain B in the Holo01 run as opposed to chain A in the Holo02 run. However, only in Holo01 the zinc(II) ion moved from the TM cavity to the transport site (Fig. 1B), whereas in Holo02 the ion remained stably bound at the entrance of the TM cavity (Fig. S2).

The distances from the residues interacting with the zinc(II) ions in the TM cavities were measured along the simulation (Fig. 1C). In the diffusion process from the cytosol towards the transport site in the Holo01 simulation, the zinc(II) ion interacted mainly with three residues: D150, which is part of the region of the TM5 helix facing the intracellular space, D49, which is part of the transport site on the TM2 helix, and H153, which is part of the transport site on the TM5 helix (Fig. 1B and Movie S1). After binding to the transport site, the zinc(II) ion remained coordinated by D49 and D150 until the end of the simulation.

In the Holo02 simulation the metal ion interacted simultaneously with D150 and E79 at the entrance of the TM cavity (Fig. S3). The latter interaction was not observed in the Holo01 trajectory. Although H153 repeatedly approached the zinc(II) ion in the first half of the simulation, the two negative charges of E79 and D150 produced an electrostatic attraction on the metal ion so strong as to prevent further displacement of the ion towards the TM cavity. To analyze whether the additional carboxylate recruited by the zinc(II) ion in Holo02 with respect to Holo01 was indeed the main responsible for preventing the transit of the cation along the channel, we removed the charge of E79 by replacing it with alanine (E79A mutant) and re-started the simulation. Upon mutation, the metal ion was able to move away from the entrance of the channel and bound to D49 similarly to Holo01 (Fig. S4).

The channel accessibility along time can be estimated by measuring a structural parameter, such as the TM cavity opening, and a biochemical parameter, such as the hydration of the channel. In this work, the channel accessibility was assessed by measuring the distances of the stable TM3 (represented by I90) helix from TM1 (A20), TM2 (D49), TM4 (I130), TM5 (L152) and TM6 (M197) (Fig. 2). In all the holo-

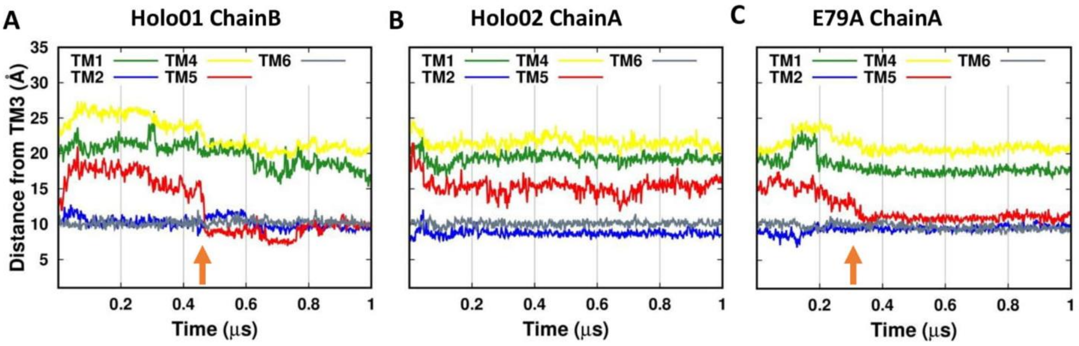


Fig. 2. TM cavity opening and water accessibility as a function of time in Holo01 chain B, Holo02 chain A and mutant E79A chain A, respectively. The graphs show the distances from the Cα atom of I90 in the TM3 helix to the Cα atoms of A20 (TM1), D49 (TM2), I130 (TM4), L152 (TM5) and M197 (TM6). The reorientation of TM5 is marked by an orange arrow.

simulations, TM2 and TM6 showed a very stable behavior preserving a direct contact with TM3. On the contrary, at the beginning of the Holo01 run the chain permeated by the zinc(II) ion displayed a rapid increase of the TM3-TM4 and TM3-TM5 distances followed by a gradual decrease to 15 Å (Fig. 2A). At 470 ns, these distances suddenly dropped due to the reorientation of TM5. In particular, the TM3-TM5 distance reached the same values as TM3-TM2 and TM3-TM6, denoting a direct contact between the TM helices. We can interpret this behavior as the channel being initially in a more exposed (open) configuration, followed by a large-scale rearrangement, induced or stabilized by the presence of the zinc(II) ion in the transport site, that eventually closed the channel. Of note, the Replica01 simulation, where a zinc(II) ion was already bound at the transport site, sampled the same large-scale rearrangement seen in Holo01 (Fig. S5A). Instead, in the chain A of Holo02, where the zinc(II) ion was stably bound at entrance of the TM cavity, all the distance profiles remained fairly constant throughout the whole simulation and did not feature any relevant conformational changes, in what can be regarded as an essentially open configuration (Fig. 2B). Allowing the zinc(II) ion to reach the transport site by restarting the Holo02 run after introducing the E79A mutation resulted in the same reorientation of TM5 observed in Holo01 (Fig. 2C). Furthermore, after this reorientation, the zinc(II) coordination became similar to the tetrahedral geometry of the outward facing configuration for some nanoseconds, with the exception of D150 taking the place of H153 (Fig. S4). This configuration allowed a second zinc(II) ion to approach the site from the extracellular environment, suggesting the existence of an accessible extracellular cavity.

Variations of the channel hydration in response to the conformational motions caused by zinc(II) binding were measured experimentally by Gupta et al. [31]. To obtain a qualitative comparison with these data we measured the time evolution of the number of water molecules (cutoff at 5 Å) surrounding the residues in the TM cavity (Fig. 3). For this analysis, we selected the residues that in experiments displayed the largest response to the presence of zinc(II) ions in solution, namely the average of V48, D49 and I50 (these three residues are grouped together in the experimental work), M151, L152 and M197. The profiles of the V48-D49-I50 group (TM2), L152 (TM5) and M197 (TM6) of chain B displayed a similar trend in the Holo01 and Replica01 simulations (Fig. 3A and S5B). These residues featured the highest water accessibility with TM4 and TM5 far away from TM3. Then, their accessibility dropped when TM5 closed the gap with TM3 closing the channel. Differently from the rest, the water accessibility of M151 (TM5) increased after the channel closed. This effect is more evident in Replica01 and E79A than Holo01. Notably, M151 displayed an opposite trend of experimental hydration change upon zinc(II) exposure with respect to the other residues examined here [31]. For comparison, in the zinc(II)-bound chain A of the Holo02 simulation, V48-D49-I50 and M197 sampled intermediate values between the poorly solvated M151 and the highly solvated L152 (Fig. 3B). Thus, the channel of chain A remained partly accessible from the IC space. The E79A mutant sampled the same TM5 rearrangement as Holo01 and Replica01. However, here only the water accessibility of M197 dropped (Fig. 3C). This is due to the fact that in this simulation a second cavity exposed to the extracellular environment increased the solvation of the top part of the channel.

We then looked at the chains where zinc(II) did not enter the channel. In chain A of Holo01 TM5 sampled short distances from TM3, denoting a TM cavity constantly closed from the IC (Fig. S6A). A similar behavior was observed also for chain B in the Holo02 simulation, with the exception of TM2 being modestly more distant from TM3 (Fig. S6B). Accordingly, the presence of water molecules in the channel was quite low for chain A of Holo01 as well as for chain B of Holo02, with the exception of the V48-D49-I50 triad in the latter simulation (Fig. S6C and S6D).

We monitored the accessibility of the TM cavity also through the distance between the side chains of L152 and M197 (Fig. 4). These residues were always in direct contact in chain A of the Holo01 and Replica01 simulations, corresponding to an inaccessible cavity. On the contrary, for chain B the interaction was suddenly achieved when the channel closed and conserved until the end of the trajectory. The conformational change that closed the channel corresponded to a rotation and tilt of TM5 that brought the L152 side chain in close contact with the side chains of I90 from TM3 and of M197 from TM6 (Movie S2). This interaction created a hydrophobic gate modulating the access of water from the IC space to the TM cavity and in turn to the transport site. The snapshots at 466 ns and 470 ns displaying the top view of the TM cavity on the chain B of Holo01 point out the link between the L152-M197 distance and the presence of water molecules in the cavity (Fig. 5). In particular, the snapshot at 466 ns shows that L152 and M197 were distant enough to allow water/ions to enter the TM cavity from the IC space. In fact, the TM cavity is filled by water molecules. Four ns later, the close contact between the side chains of L152 and M197 prevented the access to the TM cavity from the cytosol. As consequence, the number of water molecules in the cavity dropped. In addition to the above residues, A149 of TM5 approached A83 and I90 of TM3 and to A194 of TM6 making a packed cluster of side chains that reinforced the hydrophobic barrier separating the transport site from the IC space (green sticks in Fig. 5). The time evolution of this hydrophobic cluster closely paralleled the formation of the L152-M197 gate (compare Fig. 4A and Fig. S7).

To obtain further insights into the conformational dynamics of the

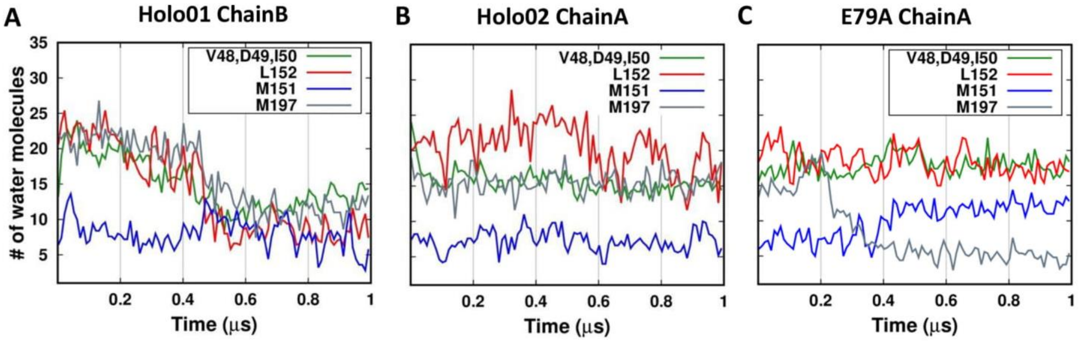


Fig. 3. Number of water molecules within 5 Å from the reported residues in the A) Holo01 chain B, B) Holo02 chain A and C) mutant E79A chain A, respectively.

TM helices in the Holo01 simulation, we calculated the atomic correlation of motions, dividing the whole simulation into two parts: before and after the reorientation of TM5 occurring at 470 ns (Fig. 6). Besides the obvious correlation among residues close in sequence, the loops EL2 (connecting TM3 and TM4) and EL3 (connecting TM5 and TM6) featured the highest correlation during the first 470 ns of simulation (Fig. 6A). This could suggest a sort of communication pathway between the TM3-TM6 helix pair and the TM4-TM5 helices of the four-helix bundle. In the second part of Holo01, the motions of TM5 and the lower part of TM4 were correlated with those of TM3 and TM6 (Fig. 6B). This correlation was absent in the first 470 ns of the simulation. Thus, the reorientation of TM5 linked directly TM4 and TM5 to the TM3-TM6 helix pair.

The principal component analysis (PCA) carried out on the Cα atoms of the Holo01 simulation highlighted the main conformational motions in a few principal components or eigenvectors. In particular, performing the PCA on the TM helices of the Holo01 chain B (excluding the long EC loops) together with the IL1 loop resulted in the first eigenvector containing 62% of the overall motions. The widest motion involved the TM4-TM5 helices (Movie S3), which moved towards the TM3-TM6 helix pair covering a long distance. In this process TM5 not only translated towards the TM3-TM6 helix pair along with TM4, but it also rotated towards the TM cavity as already described in the previous paragraphs. This is accompanied by a smaller-scale reorientation of the TM3-TM6 pair, resulting in a coordinated conformational change that modulates the packing of the four-helix bundle against the TM3-TM6 pair. We then performed a second PCA including also the CTDs (Movie S4). The first eigenvector collected 71% of the overall motions and showed again the main contributions to the mobility given by TM4, TM5 and the IL1 loop. In addition, a rotation of the CTDs was observed, thus suggesting a connection between the CTD dynamics and the conformational changes recorded in the TM region of chain B. Notably, the superposition of the TMDs of the final structures of Holo01 and Holo02 reveals that the two CTDs are tilted in opposite directions (Fig. S8A). In fact, in both structures the CTD tilted towards the chain whose TM cavity was inaccessible. The tilt was greater for the Holo01 run than the Holo02 run, possibly due to the presence of a zinc(II) ion bridging site B (which is located on the IL1 loop) of chain A to the CTD, observed only in Holo01.

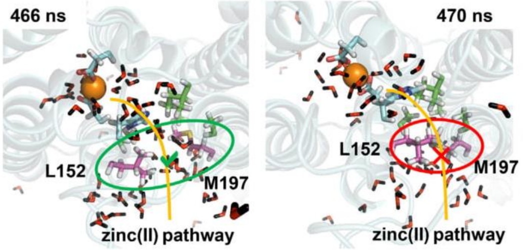


Fig. 5. Top view of the Holo01 chain B channel at 466 and 470 ns. L152 and M197 are depicted as magenta sticks, water molecules are shown as thin sticks. The residues forming the hydrophobic cluster are shown as green sticks. The residues of the transport site coordinating the zinc(II) ion are indicated as cyan sticks. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

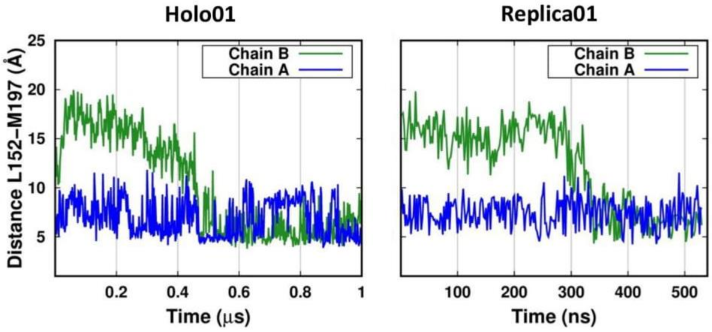


Fig. 4. Hydrophobic gate regulating the access to the TM cavity from the IC space. The distances are measured between the Cγ of L152 and the Cε of M197 in both the protein chains.

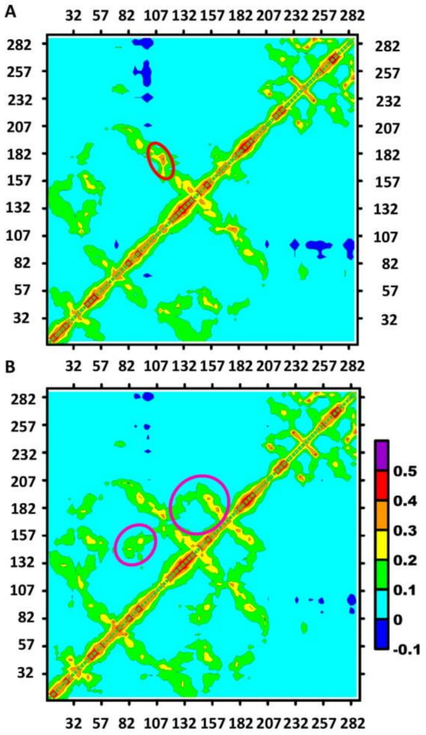


Fig. 6. Residue-residue correlations of the Holo01 Chain B. A) Correlations in the time range 1-470 ns. The red circle indicates the region with the highest correlation. B) Correlations in the time range 471–1000 ns. The magenta circles indicate the regions with correlations arisen in this time range.

3.2. Simulation of YiiP in the absence of zinc(II) ions

The No-zinc simulation was performed without zinc(II) ions in solution for a time sampling of 1 μs. We are aware that using the structure of holo-YiiP as the starting point of the No-zinc run could be far from the correct conformational ensemble sampled by apo-YiiP. Indeed, in this simulation the global structure experienced larger fluctuations than in the Holo-simulations (Fig. S1). However, the main goal of this simulation was to investigate whether the conformational changes featured in the holo simulations are zinc-dependent. To reach this goal we exploited the unique feature of molecular modelling that permits the simulation of non-physiological conditions. Thus, in this simulation a force was applied to prevent the sodium(I) and chloride(I) ions access in the TM cavities, so that the cations could not induce conformational effects similar to zinc(II), while preserving the physiological protein-sodium chloride interactions in the remaining regions of the protein.

To compare with the previous trajectories, we measured the distances of TM3 (I90) from TM1 (A20), TM2 (D49), TM4 (I130), TM5 (L152) and TM6 (M197) as a gauge of the accessibility of the TM cavity (Fig. S9). For chain A, the profile is similar to that observed for chain B of the Holo01 run, consistent with high accessibility of the cavity (Fig. S9A). On the contrary, TM5 in the chain B sampled shorter distances corresponding to the partially accessible or, for short time intervals, inaccessible cavity (Fig. S9B). In accord to this scenario, the L152-M197 distance sampled high values for the chain A and intermediate values for the chain B (Fig. S10). By superimposing the TMDs of the final conformations of the Holo01 and the No-zinc runs, it became apparent that also in this case the CTDs tilted towards the chain whose cavity is inaccessible (chain A of Holo01 and chain B of No-zinc) (Fig. S8B).

3.3. Umbrella simulations

The Holo01 simulation provided us a detailed view of the zinc(II) interactions along the pathway for diffusion towards the transport site. This information can be used to perform a so-called umbrella simulation to derive the free-energy barrier encountered by the ion. The metal dependency of the energetic barriers can be estimated by replacing the zinc(II) ion with other metals. Thus, we performed three umbrella simulations using a zinc(II) or sodium(I) ion as probe on the inward-facing state. The two zinc runs exploited the same starting conformation extracted from the unbiased Holo01 run when the metal ion is located at the entrance of the TM cavity of chain B (corresponding to 0 Å in Fig. 7). Instead, the sodium run started with the ion placed some angstroms away from the channel entrance. The more distant starting point of the sodium(I) than zinc(II) runs is motivated by the fact that there is no evidence in the literature that sodium can enter YiiP. Thus, measuring the energetic barrier of the metal approach to the channel entrance can be relevant. The highest free-energy barrier encountered when moving from the cytosol to the transport site was similar for both metal ions (Fig. 7). In correspondence of the energetic minima, the two metal ions interacted with the same residues (Fig. S11). In particular, the coordination of the zinc(II) ion is the same as in the unbiased Holo01 simulation, with the carboxylates of D150 and D49 interacting with the metal (Fig. S11A). Instead, the energetic profiles when going from the transport site to the EC space differed significantly between the two metals. The sodium(I) ion free-energy increased only slightly before reaching the top of the channel. Instead, in the two zinc(II) runs we observed a dramatic energy increase when detaching from D157 in the transport site to move to the top of the channel. In summary, the inward-facing conformation favored the entrance of the cations in the

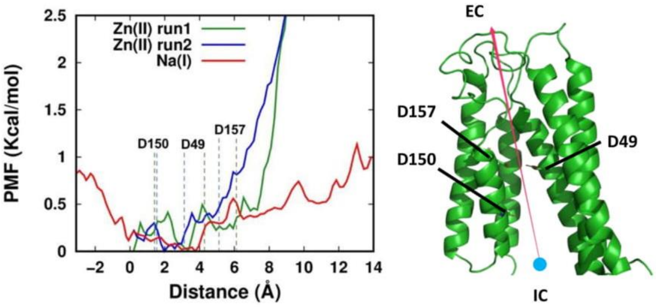


Fig. 7. Free-energy profiles of the zinc(II) and sodium(I) ions crossing the TM cavity. The channel entrance from the IC is aligned at 0 Å. The main binding residues are reported.

TM cavity and their translocation to the transport site. Here, the coordination environment of the zinc(II) ion is well defined; similar interactions with the protein have been observed also for the sodium(I) cation. In the process of release from the transport site, the zinc(II) ion experienced a dramatically high energetic barrier, which effectively prevents its free diffusion from the channel to the bulk solution. This suggests that a significant conformational rearrangement is needed to lower the energetic barrier for the exit of the zinc(II) ion. Interestingly, there is a comparatively much lower barrier for the release of the sodium(I) ion, indicating that the interactions of the latter within the transport site are somewhat weaker, albeit structurally similar, than those formed by the zinc(II) ion. This can be ascribed to the higher charge of the zinc cation.

4. Discussion

The main force that drives the zinc(II) ions towards the YiiP protein channels is electrostatic attraction. Indeed, the YiiP transporter has three negatively charged residues located at the entrance of the TM cavity: E79 on TM3, D150 on TM5 and E200 on TM6. In our holo-simulations the zinc(II) ion permeating the TM cavity bound to D150 or D150 and E79. In particular, the simultaneous interaction with both D150 and E79 occurring in the Holo02 run blocked the metal ion at the TM cavity entrance for the entire simulation. Instead, in the Holo01 run the metal ion interacted initially only with D150, allowing it to subsequently reach the transport site. We thus postulated that the zinc(II) interaction with two negative charges generated a very high energetic barrier preventing the ion from moving deeper in the channel and also inhibiting the rearrangement of TM5. We tested this hypothesis by mutating E79 to alanine. In the E79A mutant, the zinc(II) ion previously blocked at the entrance of the TM cavity reached the transport site as seen in the Holo01 simulation (Fig. S4). Notably, the human homologs of YiiP lack D150, which could result in a similar effect of lowering the energetic barrier for the diffusion of zinc(II) within the channel while preserving the negative electrostatic potential needed for attracting the ions from the IC space [44]. Indeed, the energetic profile of Fig. 7 indicates that once the metal ion has reached the entrance of the TM cavity, its translocation to the transport site encounters modest barriers. Electrostatics is, based on our energetics calculations, a crucial factor also in preventing the free diffusion of the metal ion from the transport site towards the EC space in the inward-facing state of YiiP.

The possibility for the zinc(II) ion to enter the TM cavity from the IC space is related to the accessibility of the TM cavity delimited by the four-helix bundle (TM1, TM2, TM4 and TM5) and the TM3-TM6 helix pair. It has been proposed that the dynamics of TM5 with respect to the TM3-TM6 helix pair is the main factor in modulating the access to the transport site [31]. Our Holo simulations demonstrated a tight correlation between the motion of helices TM5 and, to a lesser extent, TM4 and the presence of water molecules in the TM cavity. The TM helices sampled short distances in the chains not permeated by zinc(II) and had low solvation, corresponding to a closed configuration. Instead, the chains whose TM cavity was permeated by zinc(II) displayed a relevant increase of the distance between the TM4-TM5 pair and the TM3-TM6 pair at the beginning of the simulation (open configuration); this was accompanied by relatively high hydration of the channel (Figs. 2 and 3). After the zinc(II) ion reached the transport site by binding to D49, TM4 and TM5 slowly closed the gap with the TM3-TM6 helix pair. In parallel, the number of water molecules in the cavity gradually decreased. Then, a rapid conformational change took place that put TM5 in close contact with the TM3 making the TM cavity inaccessible from the IC space. The rearrangement of TM5 involved also a motion of TM4, so that the whole structural change can be described as the four-helix bundle achieving a tighter packing against the TM3-TM6 pair. In our Holo01 and Replica01 simulations, this led to the formation of a hydrophobic gate between L152 and M197 (Figs. 4 and 5). This process was highlighted by X-ray–mediated hydroxyl radical labeling measurements, leading to the suggestion that the hydrophobic gate is responsible for the regulation of zinc(II) access to the TM cavity [31]. Our simulations support this view.

In all Holo runs, including the E79A mutant, we found only one of the two available TM cavities was accessible, whereas the other was totally or partially inaccessible. Therefore, the usual symmetric representation of the YiiP transporter does not match the dynamics sampled by our simulations. The open cavity was the one permeated by the zinc(II) ion. In the simulation without zinc(II) ions both chains were in an open or partly open configuration. Thus, it is possible that the binding of zinc(II) initially stabilizes the chain in the open configuration while the other chain rearranges to the closed configuration. Subsequently, when the zinc(II) ion reaches the transport site also the first chain reorganizes to reach the closed configuration.

Based on X-ray–mediated hydroxyl radical labeling measurements, it was suggested that zinc(II) binding at the transport site with tetrahedral coordination geometry is the crucial event to trigger the conformational change of TM5 [31]. In particular, it was hypothesized that this rearrangement would cause the TM cavity to close in the portion facing the IC while opening towards the extracellular side in a sort of alternating mechanism. Here, in the Holo01 and Replica01 simulations YiiP underwent a wide conformational change that caused the closing of the TM cavity upon zinc(II) binding at the transport site. To achieve this, it was sufficient for the zinc(II) ion to reach a stable interaction with only D49 of the transport site, suggesting that tetrahedral coordination is not essential to trigger the reorientation of TM5. Only in the E79A simulation we could observe the coordination sphere of the zinc(II) ion in the transport site achieve a geometry very similar to the tetrahedral geometry of the outward facing configuration. Intriguingly, this event was accompanied by other smaller-scale structural changes, which allowed a second zinc(II) ion to approach the site from the EC side of the protein. This observation is in line with a recent experimental work suggesting that a conformation of the transporter where a persistent dimer interface is combined with an inaccessible intracellular cavity and an accessible extracellular cavity is able to transport zinc(II) across the membrane [29]. Overall, our simulations and the latter experimental data indicate that the four-helix bundle rearranges with respect to a static TM3-TM6 helix pair scaffold, in a rocking-bundle mechanism common also in other active transporters as LeuT [45,46]. This mechanism was apparent in the conformational dynamics sampled in the Holo01 run (Movie S5), even though the simulation length is very short compared to the estimated timescale of the overall zinc(II) transport process [23,47]. Furthermore, our umbrella simulations showed that the IC-facing conformation permitted the translocation of zinc(II) to the transport site but not its release. Thus, we can postulate that a further conformational change involving at least the TM helices harboring the zinc(II)-binding residues (TM2 and TM5) is needed to lower the free-energy barrier of the zinc(II) release process [44].

In conclusion, our simulations combined with the available structural and experimental data on YiiP provide an atomistic view of various steps of the transport mechanism involving the inward-facing conformation. First of all, we observed that during zinc(II) transport the symmetry of the YiiP dimer is lost, with the channels in the two chains consistently featuring a significantly different accessibility. Zinc(II) permeation within the TM cavity of one chain is sufficient to trigger the compaction of the four-helix bundle and the TM3-TM6 helix-pair in the TMD of that chain. After the zinc(II) ion fully enters the transport site further structural changes may put the site in communication with the extracellular space without achieving a full transition of the outward-facing state. Nevertheless, our energy calculations suggest that there is a substantial energetic barrier preventing the release of the ion from the coordination environment in the transport site. Our data do not allow us to evaluate how this barrier is eventually removed, an event which may be combined with the proton antiport performed by YiiP.

Supplementary data to this article can be found online at https://doi.org/10.1016/j.bbagen.2019.06.001.

Acknowledgements

This work was supported by Consorzio Interuniversitario Risonanze Magnetiche di Metallo Proteine (CIRMMP).

References

[1] R.J.P. Williams, J.J.R. Frausto da Silva, The Biological Chemistry of the Elements: The Inorganic Chemistry of Life, Oxford University Press, 2001.

[2] W. Maret, Zinc biochemistry: from a single zinc enzyme to a key element of life, Adv. Nutr. An Int. Rev. J. 4 (2013) 82–91, https://doi.org/10.3945/an.112.003038.

[3] A. Krężel, W. Maret, The biological inorganic chemistry of zinc ions, Arch. Biochem. Biophys. 611 (2016) 3–19, https://doi.org/10.1016/j.abb.2016.04.010.

[4] J.L. Vinkenborg, T.J. Nicolson, E.A. Bellomo, M.S. Koay, G.A. Rutter, M. Merkx, Genetically encoded FRET sensors to monitor intracellular Zn2+homeostasis, Nat. Methods 6 (2009) 737–740, https://doi.org/10.1038/nmeth.1368.

[5] C. Cubillas, P. Vinuesa, M.L. Tabche, A. García-De Los Santos, Phylogenomic analysis of cation diffusion facilitator proteins uncovers $Ni_2{}^+/Co_2{}^+$ transporters, Metallomics. 5 (2013) 1634–1643, https://doi.org/10.1039/c3mt00204g.

[6] B. Montanini, D. Blaudez, S. Jeandroz, D. Sanders, M. Chalot, Phylogenetic and functional analysis of the Cation Diffusion Facilitator (CDF) family: improved signature and prediction of substrate specificity, BMC Genomics 8 (2007) 107, https://doi.org/10.1186/1471-2164-8-107.

[7] S. Barber-Zucker, B. Shaanan, R. Zarivach, Transition metal binding selectivity in proteins and its correlation with the phylogenomic classification of the cation diffusion facilitator protein family, Sci. Rep. 7 (2017) 1–12, https://doi.org/10.1038/s41598-017-16777-5.

[8] H. Zhao, D. Eide, The yeast ZRT1 gene encodes the zinc transporter protein of a high-affinity uptake system induced by zinc limitation, Proc. Natl. Acad. Sci. U. S. A. 93 (1996) 2454–2458, https://doi.org/10.1073/pnas.93.6.2454.

[9] N. Grotz, T. Fox, E. Connolly, W. Park, M.L. Guerinot, D. Eide, Identification of a family of zinc transporter genes from Arabidopsis that respond to zinc deficiency, Proc. Natl. Acad. Sci. 95 (1998) 7220–7224, https://doi.org/10.1073/pnas.95.12.7220.

[10] R.D. Palmiter, L. Huang, Efflux and compartmentalization of zinc by members of the SLC30 family of solute carriers, Pflugers Arch. Eur. J. Physiol. 447 (2004) 744–751, https://doi.org/10.1007/s00424-003-1070-7.

[11] L. Huang, S. Tepaamorndech, The SLC30 family of zinc transporters-a review of current understanding of their biological and pathophysiological roles, Mol. Asp. Med. 34 (2013) 548–560, https://doi.org/10.1016/j.mam.2012.05.008.

[12] O. Kolaj-Robin, D. Russell, K.A. Hayes, J.T. Pembroke, T. Soulimane, Cation diffusion facilitator family: structure and function, FEBS Lett. 589 (2015) 1283–1295, https://doi.org/10.1016/j.febslet.2015.04.007.

[13] I.T. Paulsen, M.H. Saier, A novel family of ubiquitous heavy metal ion transport proteins, J. Membr. Biol. 156 (1997) 99–103, https://doi.org/10.1007/s002329900192.

[14] A. Salusso, D. Raimunda, Defining the roles of the Cation diffusion facilitators in Fe^{2+}/Zn^{2+} homeostasis and establishment of their participation in virulence in *Pseudomonas aeruginosa*, Front. Cell. Infect. Microbiol. 7 (2017) 1–14, https://doi.org/10.3389/fcimb.2017.00084.

[15] E. Bafaro, Y. Liu, Y. Xu, R.E. Dempski, The emerging role of zinc transporters in cellular homeostasis and cancer, Signal Transduct. Target. Ther. 2 (2017) 17029, , https://doi.org/10.1038/sigtrans.2017.29.

[16] T. Kambe, Y. Yamaguchi-Iwai, R. Sasaki, M. Nagao, Overview of mammalian zinc transporters, Cell. Mol. Life Sci. 61 (2004) 49–68, https://doi.org/10.1007/s00018-003-3148-y.

[17] G. Grass, M. Otto, B. Fricke, C.J. Haney, C. Rensing, D.H. Nies, D. Munkelt, FieF (YiiP) from Escherichia coli mediates decreased cellular accumulation of iron and relieves iron stress, Arch. Microbiol. 183 (2005) 9–18, https://doi.org/10.1007/s00203-004-0739-4.

[18] Y. Wei, H. Li, D. Fu, Oligomeric state of the *Escherichia coli* metal transporter YiiP, J. Biol. Chem. 279 (2004) 39251–39259, https://doi.org/10.1074/jbc.M407044200.

[19] D.S. Parsons, C. Hogstrand, W. Maret, The C-terminal cytosolic domain of the human zinc transporter ZnT8 and its diabetes risk variant, FEBS J. 285 (2018) 1237–1250, https://doi.org/10.1111/febs.14402.

[20] R. Sladek, G. Rocheleau, J. Rung, C. Dina, L. Shen, D. Serre, P. Boutin, D. Vincent, A. Belisle, S. Hadjadj, B. Balkau, B. Heude, G. Charpentier, T.J. Hudson, A. Montpetit, A.V. Pshezhetsky, M. Prentki, B.I. Posner, D.J. Balding, D. Meyre, C. Polychronakos, P. Froguel, A genome-wide association study identifies novel risk loci for type 2 diabetes, Nature. 445 (2007) 881–885, https://doi.org/10.1038/nature05616.

[21] G.A. Rutter, F. Chimienti, SLC30A8 mutations in type 2 diabetes, Diabetologia. 58 (2015) 31–36, https://doi.org/10.1007/s00125-014-3405-7.

[22] B. Yi, G. Huang, Z. Zhou, Different role of zinc transporter 8 between type 1 diabetes mellitus and type 2 diabetes mellitus, J. Diabetes Investig. 7 (2016) 459–465, https://doi.org/10.1111/jdi.12441.

[23] Y. Chao, D. Fu, Kinetic study of the Antiport mechanism of an Escherichia coli zinc transporter, ZitB, J. Biol. Chem. 279 (2004) 12043–12050, https://doi.org/10.1074/jbc.M313510200.

[24] Y. Wei, D. Fu, Selective metal binding to a membrane-embedded aspartate in the *Escherichia coli* metal transporter YiiP (FieF), J. Biol. Chem. 280 (2005) 33716–33724, https://doi.org/10.1074/jbc.M506107200.

[25] M. Lu, D. Fu, Structure of the zinc transporter YiiP, Science (80-) 317 (2007) 1746–1748, https://doi.org/10.1126/science.1143748.

[26] M. Lu, J. Chai, D. Fu, Structural basis for autoregulation of the zinc transporter YiiP, Nat. Struct. Mol. Biol. 16 (2009) 1063–1067, https://doi.org/10.1038/nsmb.1662.

[27] N. Coudray, S. Valvo, M. Hu, R. Lasala, C. Kim, M. Vink, M. Zhou, D. Provasi, M. Filizola, J. Tao, J. Fang, P.A. Penczek, I. Ubarretxena-Belandia, D.L. Stokes, Inward-facing conformation of the zinc transporter YiiP revealed by cryoelectron microscopy, Proc. Natl. Acad. Sci. U. S. A. 110 (2013) 2140–2145, https://doi.org/10.1073/pnas.1215455110.

[28] O. Jardetzky, Simple allosteric model for membrane pumps, Nature. 211 (1966) 969–970, https://doi.org/10.1038/211969a0.

[29] M.L. Lopez-Redondo, N. Coudray, Z. Zhang, J. Alexopoulos, D.L. Stokes, Structural basis for the alternating access mechanism of the cation diffusion facilitator YiiP, Proc. Natl. Acad. Sci. (2018) 201715051, , https://doi.org/10.1073/pnas.1715051115.

[30] Y. Shi, Common folds and transport mechanisms of secondary active transporters, Annu. Rev. Biophys. 42 (2013) 51–72, https://doi.org/10.1146/annurev-biophys-083012-130429.

[31] S. Gupta, J. Chai, J. Cheng, R. D'Mello, M.R. Chance, D. Fu, Visualizing the kinetic power stroke that drives proton-coupled zinc(II) transport, Nature. 512 (2014) 101–104, https://doi.org/10.1038/nature13382.

[32] N. Eswar, B. Webb, M.A. Marti-Renom, M.S. Madhusudhan, D. Eramian, M. Shen, U. Pieper, A. Sali, Comparative Protein Structure Modeling Using MODELLER, Curr. Protoc. Protein Sci, John Wiley & Sons, Inc, Hoboken, NJ, USA, 2007, pp. 2.9.1–2.9.31, , https://doi.org/10.1002/0471140864.ps0209s50.

[33] E.L. Wu, X. Cheng, S. Jo, H. Rui, K.C. Song, E.M. Dávila-Contreras, Y. Qi, J. Lee, V. Monje-Galvan, R.M. Venable, J.B. Klauda, W. Im, CHARMM-GUI membrane builder toward realistic biological membrane simulations, J. Comput. Chem. 35 (2014) 1997–2004, https://doi.org/10.1002/jcc.23702.

[34] S. Jo, T. Kim, V.G. Iyer, W. Im, CHARMM-GUI: a web-based graphical user interface for CHARMM, J. Comput. Chem. 29 (2008) 1859–1865, https://doi.org/10.1002/jcc.20945.

[35] P. Li, B.P. Roberts, D.K. Chakravorty, K.M. Merz, Rational design of particle mesh Ewald compatible Lennard-Jones parameters for + 2 metal cations in explicit solvent, J. Chem. Theory Comput. 9 (2013) 2733, https://doi.org/10.1021/ct400146w.

[36] R. Gould, I.R. Skjevik, A.A. Dickson, C.J. Madej, B.D. Walker, Lipid17: A Comprehensive AMBER Force Field for the Simulation of Zwitterionic and Anionic Lipids, (2018) In Prep.

[37] J.A. Maier, C. Martinez, K. Kasavajhala, L. Wickstrom, K.E. Hauser, C. Simmerling, ff14SB: improving the accuracy of protein side chain and backbone parameters from ff99SB, J. Chem. Theory Comput. 11 (2015) 3696–3713, https://doi.org/10.1021/acs.jctc.5b00255.

[38] D.A. Case, T.E. Cheatham, T. Darden, H. Gohlke, R. Luo, K.M. Merz, A. Onufriev, C. Simmerling, B. Wang, R.J. Woods, The Amber biomolecular simulation programs, J. Comput. Chem. 26 (2005) 1668–1688, https://doi.org/10.1002/jcc.20290.

[39] D.A. Case, I.Y. Ben-Shalom, S.R. Brozell, D.S. Cerutti, T.E.I. Cheatham, V.W.D. Cruzeiro, T.A. Darden, R.E. Duke, D. Ghoreishi, M.K. Gilson, H. Gohlke, A.W. Goetz, D. Greene, R. Harris, N. Homeyer, S. Izadi, A. Kovalenko, T. Kurtzman, T.S. Lee, S. LeGrand, P. Li, J. Liu, T. Luchko, R. Luo, D.J. Mermelstein, K.M. Merz, Y. Miao, G. Monard, C. Nguyen, H. Nguyen, I. Omelyan, A. Onufriev, F. Pan, R. Qi, D.R. Roe, A. Roitberg, C. Sagui, S. Schott-Verdugo, J. Shen, C.L. Simmerling, J. Smith, J. Swails, R.C. Walker, J. Wang, H. Wei, R.M. Wolf, X. Wu, L. Xiao, D.M. York, P.A. Kollman, AMBER 2018, (2018).

[40] J.F. Nagle, Area/lipid of bilayers from NMR, Biophys. J. 64 (1993) 1476–1481, https://doi.org/10.1016/S0006-3495(93)81514-5.

[41] R. Salomon-Ferrer, A.W. Götz, D. Poole, S. Le Grand, R.C. Walker, Routine microsecond molecular dynamics simulations with Amber on Gpus. 2. Explicit solvent particle mesh Ewald, J. Chem. Theory Comput. 9 (2013) 3878, https://doi.org/10.1021/ct400314y.

[42] B. Roux, The calculation of the potential of mean force using computer simulations, Comput. Phys. Commun. 91 (1995) 275–282, https://doi.org/10.1016/0010-4655(95)00053-I.

[43] M. Andrec, The Weighted Histogram Analysis Method (WHAM), Spring, 2010, https://doi.org/10.1136/bmj.39349.437442.43.

[44] Y. Golan, R. Alhadeff, F. Glaser, A. Ganoth, A. Warshel, Y.G. Assaraf, Demonstrating aspects of multiscale modeling by studying the permeation pathway of the human ZnT2 zinc transporter, PLoS Comput. Biol. 14 (2018) e1006503, , https://doi.org/10.1371/journal.pcbi.1006503.

[45] G. Jeschke, A comparative study of structures and structural transitions of secondary transporters with the LeuT fold, Eur. Biophys. J. 42 (2013) 181–197, https://doi.org/10.1007/s00249-012-0802-z.

[46] L.R. Forrest, G. Rudnick, The rocking bundle: a mechanism for ion-coupled solute flux by symmetrical transporters, Physiology. 24 (2009) 377–386, https://doi.org/10.1152/physiol.00030.2009.

[47] Y. Chao, D. Fu, Thermodynamic studies of the mechanism of metal binding to the *Escherichia coli* zinc transporter YiiP, J. Biol. Chem. 279 (2004) 17173–17180, https://doi.org/10.1074/jbc.M400208200.

Supporting information

Supplementary figures

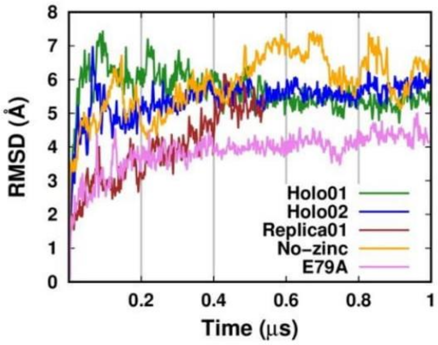


Figure S1. RMSD of the Cα atoms over time.

Holo02

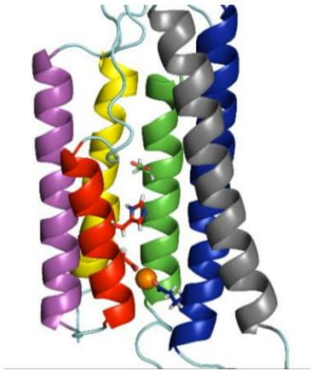


Figure S2. Final snapshot of the holo02 simulation.

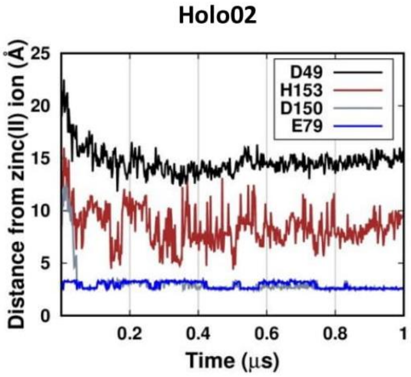


Figure S3. Zinc(II) interaction pathway in the TM cavity of chain A of the Holo02 run. Distances were measured from the zinc(II) ion to the Cγ atoms of D49 and D150, to the Nδ atom of H153 and to the Cδ atom of E79.

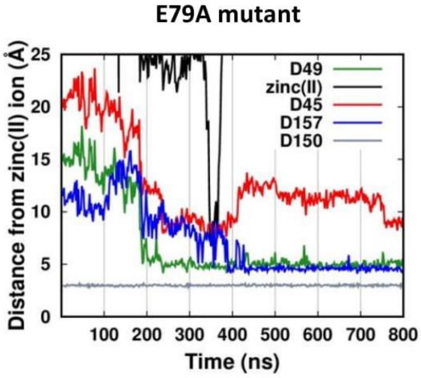


Figure S4. Zinc(II) interaction pathway in the TM cavity of the E79A mutant simulation. All the distances refer to the zinc(II) ion permeating the channel from the IC space. Thus, the black line indicates the zinc(II)-zinc(II) distance between the ion bound in the transport site and another ion coming from the EC space.

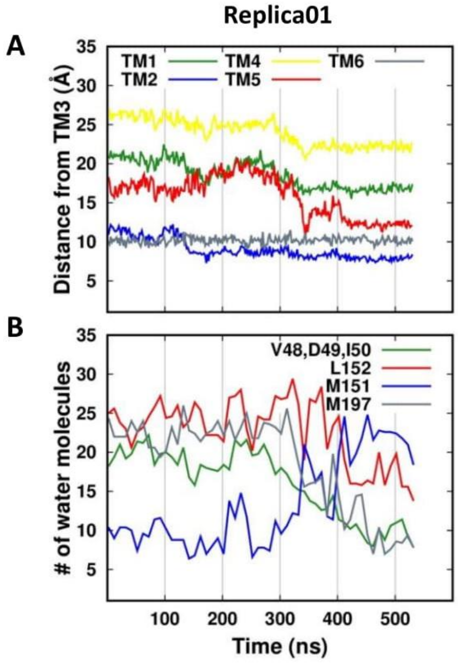


Figure S5. TM cavity opening and water accessibility of chain B of the Replica01 simulation. **A**) Distances from the Cα atom of I90 on the TM3 helix the Cα atom of A20 (TM1), D49 (TM2), I130 (TM4), L152 (TM5) and M197 (TM6). **B**) Number of water molecules within 5 Å from the residues.

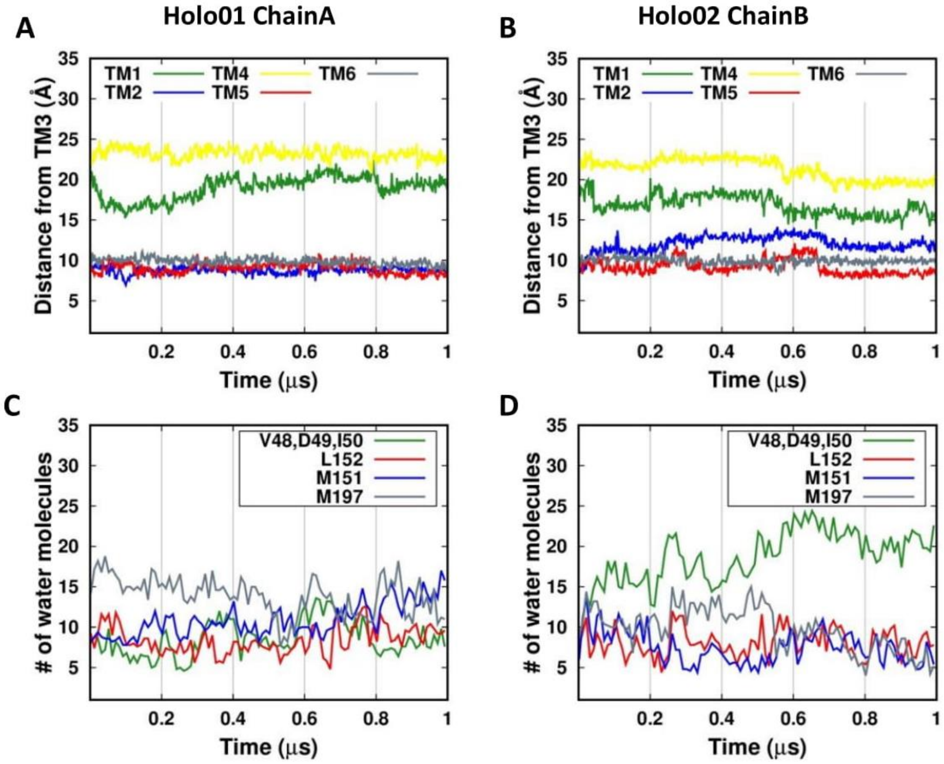


Figure S6. Opening and water accessibility of the TM cavity vs time in the Holo01 chain A and Holo02 chain B, respectively. **A-B**) Distances from the Cα atom of I90 on TM3 helix to the the Cα atoms of A20 (TM1), D49 (TM2), I130 (TM4), L152 (TM5) and M197 (TM6). **C-D**) Number of water molecules within 5 Å from the residues.

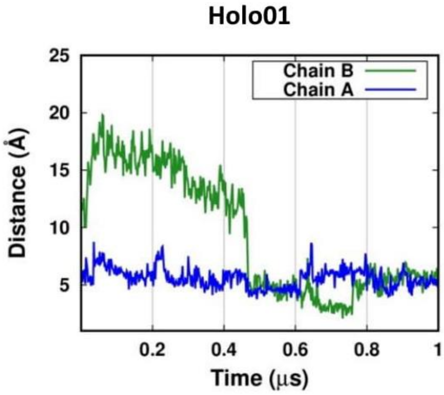


Figure S7. Distance between groups of aliphatic residues in the TM cavities of the Holo01 simulation. The distance is measured between the side chains of A149(TM5)-L152(TM5) and the side chains of A83(TM3)-I90(TM3)-A194(TM6)-M197(TM6).

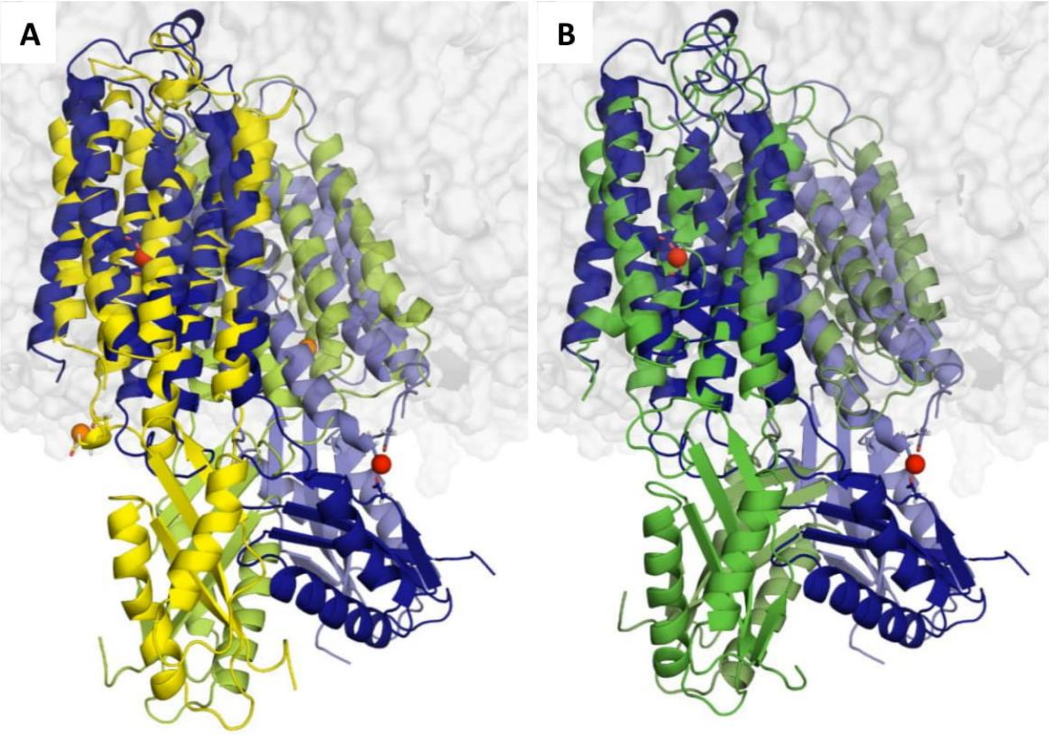


Figure S8. Superimposition of the final structure of the simulations. The conformations are aligned in the TMD region. The two protein chains have different shades of the same color. The Holo01 zinc(II) ion bound to the transport site and the one bridging the Site B on IL1 and the CTD are shown as red spheres. The residues coordinating the zinc(II) ions are shown as sticks. **A**) The Holo01 and Holo02 conformations are in blue and yellow, respectively. The Holo02 zinc(II) ions bound at the TM cavity entrance and the one bound to the SiteB on IL1 are shown as orange spheres. **B**) The Holo01 and No-zinc conformations are in blue and green, respectively.

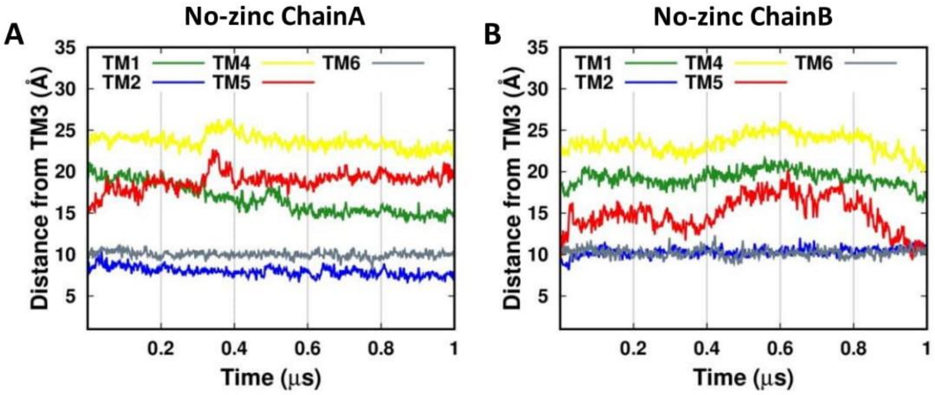


Figure S9. The TM cavity opening along time in the No-zinc simulation. Distances among the Cα atoms of I90 on TM3 helix from A20 (TM1), D49 (TM2), I130 (TM4), L152 (TM5) and M197 (TM6). **A**) Chain A. **B**) Chain B.

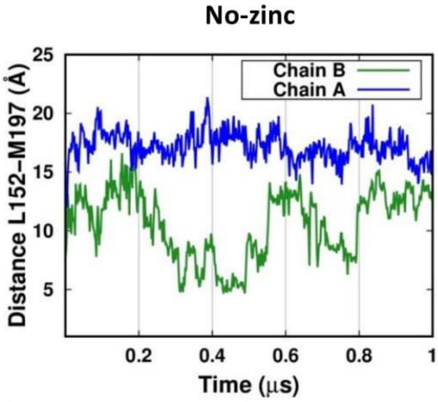


Figure S10. Hydrophobic gate regulating the access to the TM cavity from the IC space in the No-zinc simulation. The distances are measured between the Cγ of L152 and the Cε of M197 in both the protein chains.

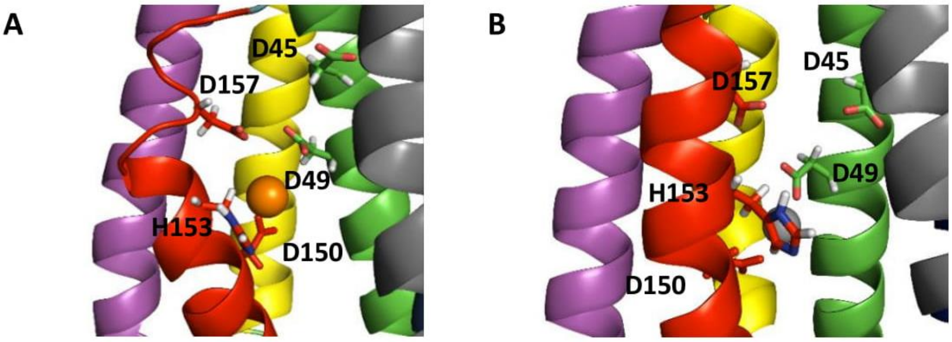


Figure S11. Metal ions coordination at the energetic minimum of the corresponding umbrella simulation. **A**) zinc(II). **B**) sodium(I).

3.4 Conclusions

My Ph.D. project contributed to the investigation on the role of metals in metalloproteins conformational dynamics by performing MD studies on metal transporters and by studying relevant biological processes as metal-coupled folding. In so doing, different molecular dynamics approaches were implemented to get an atomistic view of the molecular mechanisms under exploration.

Fist, our study on the small PfRd metalloprotein showed some difficulties in the description of the metal-coupled folding not linked to the technological gap. The main limitations were due to the partial accuracy of the classical force field describing the metal binding. In fact, they do not take in account important effects such as metal induced protonation/deprotonation, the polarizability, the charge transfer and multiscale coupling. Furthermore, this is only part of the problem, the counterpart regards the bias induced by the force field on the protein portion. In this regard, some limitations affecting the force field quality are well-known, such as the description of phenylalanine side chains. Other authors pointed out the importance of taking in account also the unfolded proteins in the description of electrostatic and water interactions. In this regard, the recent force fields CHARMM36m[114] or a99SB-disp[115] in combination with TIP4P-D showed excellent performances. However, significant methodological work is still needed until unbiased metal-induced folding of metalloproteins can be achieved.

We described the pH-dependent mechanism of iron release from human H-ferritin. Based on the umbrella sampling simulations we could estimate the free-energy profiles of the initial configurations. In conclusion, the positive charges on the side chains of three His residues proved to be crucial to weaken the metal coordination by compensating the negative charges of the Asp and Glu side chains inside the C3 channel. The illustrated pH-dependent process of iron discharge from ferritin might help defining the role of the acidic environment of endosomes and lysosomes in modulating iron release within these organelles. Encapsulation in endosomes was proposed to be relevant for H-ferritin internalization via the TfR1 receptor and therefore plays a role for the use of ferritin as a drug nanocarrier targeting cancer cells overexpressing TfR1. The delivery of ferritin to lysosomes, instead, has been proposed as a key role in controlling iron cellular homeostasis. The interplay between biomineral solubilization upon iron reduction and ferritin degradation at this level are not yet clear and the observed cage-assisted release at low pH might also come into play.

The data collected in a following study showed that in presence of high amount of Fe^{2+} ions, the structure of the ferritin with arachidonic acid (ARA) is much more stable than the protein alone. This might be due to the higher amount of negative charged present in the ARA form of the protein or because ARA can play a proper structural role inside the ferritin cage. Furthermore, it has been observed that ARA can increase the initial rate of biomineralization[116], but is not clear how such effect is produced. Our data suggested that ARA prevent the formation of some conformations but promote the conformational selection of specific iron-bound configurations in the nuclease site. This influence from the ARA could increase the initial rate of biomineralization, decreasing the “range” of possible conformations to explore in order to get to the actual conformation/s responsible for the biomineralization process.

In literature, the dimeric YiiP transporter was always depicted as a symmetric antiporter, with the two channels either accessible or inaccessible. Instead, in our simulations the two channels had always an alternating state, one accessible and the other one inaccessible, with the CTD always tilted toward the channel inaccessible. The Zn^{2+} entrance in the cavity triggered the channel closing, through the TM4-TM5 motion. Then, the Zn^{2+} binding on the transport site caused the reorientation of TM5 in turn forming the L152-M197 hydrophobic gate that prevented the water and ions access from the cytosol. In conclusion, our results suggested a correlation between the CTD tilted position and the conformational dynamics upon Zn^{2+} binding at the transport site. A defined communication pathway connecting the CTD to the TMD could be significant not only in the switch between IF and OF states as previously postulated, but also in the signal transmission between the two protein chains, especially if the two channels have an alternating open/closed state and the dimeric interface does not need to splay apart to transport Zn^{2+} across the membrane, as recently suggested[105].

MICROREVIEW

Application of Molecular Dynamics to the Investigation of Metalloproteins Involved in Metal Homeostasis

Davide Sala[a], Francesco Musiani[b] and Antonio Rosato[a,c]*

[a] Magnetic Resonance Center (CERM), University of Florence, Via Luigi Sacconi 6, 50019 Sesto Fiorentino, Italy

[b] Laboratory of Bioinorganic Chemistry, Department of Pharmacy and Biotechnology, University of Bologna, Viale Giuseppe Fanin 40, I40127 Bologna, Italy

[c] Department of Chemistry, University of Florence, Via della Lastruccia 3, 50019 Sesto Fiorentino, Italy

Eur. J. Inorg. Chem. 2018, 4661–4677

Application of Molecular Dynamics to the Investigation of Metalloproteins Involved in Metal Homeostasis

Davide Sala,[a] Francesco Musiani,[b] and Antonio Rosato*[a,c,d]

Abstract: Available estimates indicate that 30–40 % of all proteins need at least one metal ion to perform their biological function. Therefore, they are called metalloproteins. The correct biosynthesis of metalloproteins requires living organisms to be able to cope with issues such as the limited bioavailability or the potential cytotoxicity of several essential metals. Thus, organisms have developed complex machineries that guarantee the proper intracellular concentration and distribution among compartments of each metal, i.e. metal homeostasis. To understand how the different proteins responsible for metal homeostasis carry out their function, it is necessary to investigate their three-dimensional (3D) structure and mobility at the atomic level. Nuclear magnetic resonance spectroscopy is one of the main experimental techniques providing this information. Computer simulations of molecular dynamics (MD) complement experimental information by showing how the 3D structure fluctuates over time and as a function of environmental conditions, with the possibility of exploring a wider range of timescales and conditions than usually amenable to experiment. Here we review numerous applications of MD for the investigation of the structure and dynamics of metalloproteins, and we also mention some technical aspects related to the parametrization of metals in commonly used force fields.

1. Introduction

30–40 % of the proteins encoded by the genome of most organisms are metalloproteins, i.e. their biochemical role depends on the binding of at least one metal ion in vivo.[1,2] Consequently, several metal ions are essential micronutrients. On the other hand, many essential metals, especially from the d-block, have limited environmental availability and are cytotoxic in high amounts. These unfavorable properties required the development of cellular mechanisms to control the uptake, storage and excretion of metal ions in a selective manner. The mechanisms underlying metal homeostasis maintain the appropriate metal concentrations in the cytoplasm. In parallel, specific pathways for intracellular metal trafficking control the delivery of the appropriate metal ion to metalloproteins and the distribution of metal ions among sub-cellular compartments.[3–5] In particular, each metal ion has unique chemical and physical properties, so that metalloproteins use a specific metal ion to perform their function.[6] By balancing the uptake, excretion, storage, and intracellular distribution of metal ions, the intracellular concentration of each metal is maintained at optimal levels. This subtly regulated ensemble of processes is called metal homeostasis. Defects in metal homeostasis or in its regulation may lead to pathologic conditions.[7] Most cells respond to depletion and repletion of metal ions in a metal-specific manner. This suggests that the mechanisms of metal-dependent regulation, metal transport and storage, and metalloprotein maturation can discriminate the correct cofactor among other metal ions.

Metal-responsive transcriptional regulators, also known as metalloregulatory proteins or metal sensors, are crucial players of metal homeostasis. Each regulator in this family recognizes a specific metal ion or a group of metal ions within the cell. Binding of the cognate metal to the regulator modulates its DNA binding affinity and/or specificity, thereby linking the intracellular concentration of the metal and gene expression levels.[8–10] At the structural level, this happens by propagating the rearrangements that take place within the metal binding site to the other regions of the metalloregulator structure, thus inducing a variation of protein structure and/or dynamics. In turn, the latter variation affects the affinity of the protein for DNA. The effect of this chain of events on gene expression depends on the specific metal sensor. Ultimately, this typically leads to a significant up- or down-regulation of the production of a variety of proteins involved in metal usage or metal homeostasis, from metal-dependent enzymes to metal-transporting and

[a] *Magnetic Resonance Center (CERM),*
University of Florence,
Via Luigi Sacconi 6, 50019 Sesto Fiorentino, Italy
E-mail: rosato@cerm.unifi.it
http://www.cerm.unifi.it/

[b] *Laboratory of Bioinorganic Chemistry,*
Department of Pharmacy and Biotechnology,
University of Bologna,
Viale Giuseppe Fanin 40, I40127 Bologna, Italy

[c] *Consorzio Interuniversitario di Risonanze Magnetiche di Metallo Proteine,*
Via Luigi Sacconi 6, 50019 Sesto Fiorentino, Italy

[d] *Department of Chemistry,*
University of Florence,
Via della Lastruccia 3, 50019 Sesto Fiorentino, Italy

ORCID(s) from the author(s) for this article are available on the WWW under https://doi.org/10.1002/ejic.201800602.

metal-storage systems.[11–13] Moreover, some metalloregulators act on processes beyond metal homeostasis, such as the response to oxidative stress or acid adaptation.[4,14,15]

In this review, we will address explicitly, among other metals, iron and zinc metabolism. Iron is the most abundant trace element in the human body and is essential to all forms of life and participates in fundamental biological processes, such as photosynthesis, respiration and nitrogen fixation.[16] Its most relevant oxidation states in vivo are +2 (ferrous) and +3 (ferric). Higher oxidation states appear transiently during the catalytic cycle of some iron-dependent enzymes. Iron-dependent proteins may bind individual ions but also complex iron-containing cofactors, such as heme or iron–sulfur clusters. The use of these different cofactors is related to the environmental niche occupied by the organism, with iron–sulfur clusters being among the most ancient from the evolutionary point of view.[17] Zinc is the second most abundant trace metal in the human body; approximately 10 % of all human proteins require zinc.[18] Acquisition and efflux of zinc ions are mediated by a number of specific systems that transport zinc across the biological membrane. Human zinc transporters can be divided into two main families: (i) zinc uptake proteins (ZIP/Slc39)[19] and (ii) CDF (Cation Diffusion Facilitator) proteins for zinc export from the cytoplasm (ZnT/Slc30).[13] ZnTs transfer zinc ions from the cytosol to intracellular compartments or outside the cell. On the contrary, ZIPs raise the cytosolic concentration of zinc.

Knowledge of three-dimensional (3D) structure and dynamics are crucial to understand the function of biological systems. The NMR or X-ray structure of a protein describes the conformation corresponding to the global (or sometimes a local) energy minimum. Nevertheless, higher energy conformations can contribute significantly to the protein function, thus introducing a relationship to protein mobility.[20,21] Molecular dynamics (MD) simulations provide a convenient way to obtain an atomic-level view of the mobility of proteins. Indeed, computational approaches based on MD simulations have become a precious tool to acquire detailed insights on biological systems.[22,23] MD simulations solve Newton's equations of motion for a system of *N* interacting atoms: Equation (1).

$$m_i \frac{\partial^2 r_i}{\partial t^2} = F_i \; ; \; i = 1, \cdots, N \tag{1}$$

where m_i, r_i and F_i are the mass, the position and the force acting on the *i*-th atom, respectively. Because of the number of atoms typically involved in these systems, MD simulations are performed with the use of classical additive potentials [force-fields, V (r_1,…r_N)], which are functions of the atomic positions only:

Equation (2) is integrated in small time steps, providing an atomistic view of the evolution of the system over time (the so-called trajectory). By analyzing the trajectory after its initial "equilibration" phase, during which e.g. water molecules and ions redistribute around the macromolecular system, the macroscopic properties of interest can be extracted from the simulation.

$$F_i = -\frac{\partial V(r_1, \cdots, r_N)}{\partial r_i} \; ; \; i = 1, \cdots, N \tag{2}$$

Here we describe selected applications of MD to the investigation of metalloproteins. We also mention some technical difficulties associated with MD simulations of metalloproteins, focusing on the parametrization of the metal site.

Davide Sala received his Master's degree in Industrial Biotechnology (Bioinformatics CV) at the university of Milan-Bicocca in 2014. Currently he is a PhD student in Structural Biology at the University of Florence and his research focuses on the use of molecular dynamics methods to investigate metalloproteins.

Francesco Musiani received a PhD degree in Chemistry from the University of Florence in 2003, with Prof. Bertini. Following postdoctoral studies at the University of Bologna (Italy) with Prof. Ciurli and at the International School for Advanced Studies (SISSA/ISAS) of Trieste (Italy) with Prof. Micheletti, he was also a visiting scientist at the German Research School for Simulation Sciences (Juelich, Germany) with Prof. Carloni. He is currently a Fixed-term Senior Assistant Professor at the University of Bologna. His research is focused on computational biochemistry and biophysics applied to bioinorganic chemistry.

Antonio Rosato received his PhD in Chemistry in 1998. He is Associate Professor at the Department of Chemistry of the University of Florence since 2002. He has received the Prizes "Gastone De Santis" and "Raffaello Nasini", both from the Italian Chemical Society, in 2005 and 2009 respectively. Metalloproteins are the main focus of his research activities. Antonio Rosato has developed innovative methodologies for the study via NMR of the solution structure of paramagnetic metalloproteins. He is actively working on molecular dynamics and on the implementation of bioinformatics research of these systems. He co-authored about 110 articles in international scientific journals and has contributed to the determination of the structure in solution of around thirty metalloproteins.

2. Determination of the 3D Structure of Metalloproteins

2.1. NMR-Based Structure Determination of Metalloproteins

The biological function of metalloproteins is intrinsically tied to the properties of the bound metal ion(s), defined by its coordination chemistry. Therefore, the characterization of the metal coordination environment (number and types of ligands, geometry of the donor atoms) are a crucial part of the structure determination process for metalloproteins. The use of NMR-active metal nuclei to observe directly protein–metal interactions is feasible only for a handful of metals. In this regard, NMR studies of ^{199}Hg and ^{113}Cd have been particularly instructive, thanks to the information provided both by the chemical shifts of these nuclei when bound to proteins with respect to those observed in model compounds and by the measurement of the coupling constants between the NMR active metals and the protein nuclei (^{1}H/^{15}N/^{13}C). Unfortunately, many biologically interesting metals are not suitable for direct study by NMR, e.g. because the nuclear spin is greater than 1/2 (e.g. ^{43}Ca, ^{67}Zn) or because they harbor unpaired electrons in their biologically relevant states (e.g., Mn^{2+}, Cu^{2+}). In fact, ^{113}Cd has been used also as an NMR-active replacement for other metals, mostly Zn.[24]

For proteins binding diamagnetic metal ions, the NMR chemical shifts of protein nuclei can afford information on the metal environment. For instance, the chemical shifts of the ^{13}Cβ nuclei of zinc-bound cysteines (ca. 34 ppm) are significantly higher than all other cysteines (ca. 27 ppm). This permits the identification of zinc-bound cysteines.[25] While the above is mainly a through-bond effect, metal binding always induces variations in the chemical environment of the residues involved in metal-binding as well as of their spatial neighbors, giving rise to changes in the NMR chemical shifts of their nuclei. By mapping these chemical shift perturbations on the protein structure, it is possible to evaluate the affinity and stoichiometry of metal binding. Besides metal binding, chemical shift mapping is routinely exploited to investigate protein-ligand as well as protein–protein and protein–nucleic acid interactions. As an example, this approach enabled the characterization of the binding of diamagnetic Cu^{+} to the soluble domains of intracellular copper shuttle,[26,27] as well as to study metal-mediated formation of heterodimers based on simple titrations followed via ^{1}H-^{15}N HSQC NMR spectra.[28,29] However, it is difficult to use chemical shift mapping to define the 3D structure of a metalloprotein at atomic detail. In practice, NMR structure determination of metalloproteins typically relies on SAXS measurements or alternatively X-ray crystallography to define accurately the ligand geometry around the metal ion(s).[30] With this information available, the structure determination of a metalloprotein containing a diamagnetic metal ion does not differ significantly from the routine procedure used for any protein. It has been shown that quantum mechanical/molecular mechanical (QM/MM) molecular dynamics simulations restrained by NMR-derived data permit the refinement of the zinc coordination in protein structures even in the absence of X-ray or SAXS information.[31] However, the technical complexity of such an approach practically prevents its widespread application. Indeed, for the majority of currently available NMR structures of zinc-proteins, zinc-donor atom distances as well as bond angles involving the metal ion have been taken directly from libraries or based on available force field parameters (see section 5).

For metalloproteins containing paramagnetic metal ions, additional structural information is provided by the interaction between the NMR-active nuclei of the protein and the unpaired electron spin density on the metal.[32] This interaction affects the NMR spectra of the protein in a measurable way; these measurements can be converted into so-called paramagnetic restraints, to be used together with traditional restraints for NMR-based structure determination. A distinct advantage of paramagnetic restraints is that they are long-range in nature (i.e. they involve residues that are distant in protein sequence). Thanks to the higher magnetic moment of the electron, electron–nucleus interactions can be appreciable at much greater distances than nucleus–nucleus interactions. Thus, paramagnetic effects can provide structural restraints at distances up to ca. 40 Å from the metal site.[33] The most exploited paramagnetic restraints are derived from the enhancements of nuclear relaxation rates (typically called paramagnetic relaxation enhancements, PREs) and pseudocontact shifts[32,34–36] (PCSs). In metalloproteins, one can exploit Solomon's law[37,38] to convert the measured PREs into nucleus–metal distance restraints, which are then added to the list of the usual NMR-based internuclear distance restraints.[39] The electronic properties of the paramagnetic metal ion are treated as an effective constant that can be calibrated against an initial structural model. The treatment can be slightly more complex if the protein binds multiple metal ions with different electronic properties, as in the case of some iron-sulfur proteins. In this case, the PRE-based restraints are handled as a weighted average of the inverse sixth power of all the nucleus–metal distances.[40] The use of PCSs as structural restraints requires the determination of the magnetic susceptibility anisotropy tensor ($\Delta\chi$) of the paramagnetic metal ion, by fitting the experimental PCS data to an initial structural model.[41,42] This initial model is then refined against the PCS restraints, together with all other available restraints, and used to re-evaluate $\Delta\chi$, in an iterative fashion. It has been necessary to introduce specific routines in the structure generation programs to allow the use of PCS restraints.[43–45] There are various approaches to introduce lanthanide ions in proteins, such as the replacement of Ca^{2+} ions in calcium-binding proteins[33,46] and the chemical attachment of so-called lanthanide binding tags to proteins.[47–50] Thanks to the diverse number of unpaired electrons in the lanthanide series, different lanthanide ions induce different PCSs. This allows to produce multiple samples each with a different lanthanide ion and then combine the corresponding PCS data. This approach affords structural information within multiple shells at variable distances from the metal ion.[33]

Paramagnetic restraints typically cannot be used to determine the coordinates of the nuclei of the protein residues that coordinate the paramagnetic metal ion. This is caused either by the very strong paramagnetic broadening of their NMR signals, which makes them undetectable, or by their chemical shifts resulting from the sum of PCSs and contact shifts, which can be difficult to separate with sufficient accuracy. However, if the $\Delta\chi$ is negligible, the contact shifts experienced by the nuclei of the coordinating residues can sometimes be exploited to obtain dihedral angle restraints, e.g. for iron-sulfur clusters in ferredoxins.[51]

2.2. Molecular Dynamics for the Refinement of (Metallo)protein Structures

Usually, NOE-derived distance restraints provide the bulk of the information for NMR-based macromolecular structure determination. In addition, dihedral angle restraints, commonly obtained from chemical shifts,[52] residual dipolar couplings (rdc's),[53] and paramagnetism-based restraints[32] can be used. All the available restraints constitute the input for the simulated annealing calculations that generate the 3D structure of the macromolecule of interest. After simulated annealing, a so-called structure refinement step is performed to improve the quality of the final NMR structure, especially from the energetic point of view. Such refinement is often based on MD calculations with state-of-the-art force fields and explicit or, sometimes, implicit water.[54–58] Other possible approaches are based on structure rebuilding[59] or statistical potentials.[60] Structure refinement can improve significantly the structural quality, as measured by stereochemical parameters. In addition, it can improve the accuracy of the structure, as measured by the agreement with the experimental data.[61,62] Unfortunately, sometimes these protocols can make errors in the NMR structures less evident, especially in terms of the analysis of stereochemical parameters.[63,64]

In spite of the latter drawback, structure refinement in explicit solvent is now a routine practice within all projects of NMR-based macromolecular structure determination. General MD programs, such as AMBER[65] or GROMACS,[66] exploit conversion tools to convert the restraints used by the simulated annealing software to their own format. This is crucially needed in order to maintain the structure consistent with all experimental data also during the energetic refinement procedure. Unfortunately, such conversions are often not straightforward, e.g. due to conflicting rules for atom naming. To address this issue, we developed a web interface that greatly facilitates the setup of MD-based structure refinement. Calculations are run over a computational Grid infrastructure with many thousand CPUs. By using a distributed computational infrastructure, users do not have to rely on local PC clusters and enjoy a reduction of the wall time needed to obtain their results thanks to parallelization. In fact, each conformer of the NMR bundle is refined independently. This interface, which we called AMPS-NMR[67] (AMBER-based Portal Server for NMR structures), is available free of charge for academics at http://py-enmr.cerm.unifi.it/access/index. The development of AMPS-NMR took place in the context of the European FP7 project WeNMR.[68] After taking care of the conversion of restraints, AMPS-NMR automatically implements a predefined multi-step protocol for structure refinement, so that the user does not need to know the intricacies of the various parameters to be adjusted in order to perform an MD simulation. By exploiting the AMPS-NMR interface, the user may obtain substantial improvements in the quality of his/her NMR structure.[67] As of today, more than 450 users have registered on the AMPS-NMR portal.

More recently we extended the functionality of AMPS-NMR by enabling calculations also over a cluster of GPUs (graphics processing units). GPUs provide a great speed-up of MD calculations,[69] resulting, according to our benchmarks,[70] of a 100-fold time gain in the refinement of the NMR-based structure of a small protein. Thus, calculations that require 12 hours on a traditional Linux PC with a single CPU core can be performed in less than 10 minutes on a GPU card. In addition, this development enables computing longer unrestrained MD simulations via the simple web interface of AMPS-NMR spectroscopy.

Another significant recent improvement was the extension of AMBER to handle multiple PCS data sets. As described in the previous sections, PCS are information-rich paramagnetic restraints that can be measured when the protein binds a paramagnetic metal ion. For lanthanide-substituted proteins it is possible to measure multiple sets of PCS data by preparing different samples where a different member of the lanthanide series has been incorporated. These data sets can be used simultaneously to generate structural models, thanks to the incorporation of specific routines in different structure generation programs.[45,71] However, it was only 2016 that we extended the PSEUDOAMBER routines,[72] enabling the energetic refinement of such structural models in the presence of all available PCS restraints[73] (also together with diamagnetic restraints, if any). Our new extension implements a refinement protocol that is finely tuned version of the AMPS-NMR standard protocol. Tuning the previously developed protocol was necessary to obtain consistent convergence with multiple PCS datasets.[73] This optimized protocol allowed us to successfully refine the NMR structure of bovine calbindin D_{9k} using PCS data from multiple samples. In each sample, a different member of the entire series of lanthanide trivalent cations (excluding the radioactive Pm^{3+} and the isotropic Gd^{3+}) was substituted into the C-terminal calcium-binding site of calbindin.[24,45] This refinement yields an improvement of structural quality parameters comparable to that of the standard AMPS-NMR protocol for diamagnetic proteins. Interestingly, this new protocol allowed structures generated without inclusion of PCS data among the restraints to be successfully refined including the PCS restraints only at the MD stage; this resulted in a satisfactory agreement with all experimental restraints and high stereochemical quality.[73]

3. Molecular Dynamics Studies on Metal Storage and Transport

A number of metal ions are essential to life. Other metal ions are instead poisonous to living organisms, even when present in the environment at very low concentrations. Sometimes, the same element may be beneficial or noxious depending on speciation. Thus, the cell must control tightly the uptake and export of metals through efficient transporting systems. In addition, specific storage mechanisms allow the cell to calibrate the intracellular concentration of metal according to needs.

3.1. Ferritin and Iron

Ferritin plays a key role in the storage and transport of iron ions.[74] The main functions of ferritin are performed through safe iron oxidation, which protects the cell from metal toxicity, iron mineralization and finally controlled iron release for metal incorporation into iron-containing systems. Ferritin stores iron intracellularly, but it also circulates and binds specifically to different cell types. Binding of ferritin to Transferrin Receptor 1 (TfR1) on the cell surface triggers the uptake of ferritin into endosomes, where the pH is in the 6.0–4.8 range, and in cellular internalization via endocytosis.[75] From the structural point of view, ferritin is a globular protein consisting of 24 subunits arranged to form a nanocage structure able to store up to 4300 iron ions. The 24-*mer* structure has an octahedral symmetry; it self-assembles from the four-helix bundle subunits resulting in an overall molecular weight of about 480 kDa. Three types of subunits can be found in mammalian ferritins: H (FtH, heavy 21 ca. kDa), L (FtL, light, ca. 19 kDa) and the mitochondrial one (FtMt, about 21 kDa) exclusively found in mitochondrial ferritin. The sequence similarity among the subunits is high; nevertheless, only FtMt and FtH harbor the oxidoreductase site responsible of the oxidation of Fe^{2+}.[76] Instead, the FtL chain contains the nucleation site that can bind Fe^{3+} and catalyze the formation of the biomineral.[77] At variance with homopolymeric mitochondrial ferritin, the ratio of the two subunits in the cytosolic protein can vary depending on the tissue and also under pathological conditions.[78] The main functions of ferritin are performed through safe iron oxidation, which protects the cell from metal toxicity, iron mineralization and finally controlled iron release for metal incorporation into iron-containing systems.

Two different channels connect external environment and the ferritin cavity at the C3 and C4 symmetry axes of the 24-*mer* structure. The two channels have different chemical properties related to their amino acidic composition. The hydrophobic C4 channel has a role in the diffusion of dioxygen and hydrogen peroxide,[79] whereas the hydrophilic and negatively charged C3 channel is responsible for the uptake of Fe^{2+} due to the favorable electrostatic field potential.[80,81] The iron transport from the external environment to the catalytic site has been extensively studied in the past years; instead, the opposite process, i.e. the release of biomineralized iron from the cage to the bulk solution, is not totally clear. In vivo, the release process is associated to cage disassembly,[82] mainly due to lysosome degradation[83,84] but also the proteasomal pathway might be involved.[85,86] Importantly, an acidic environment is crucial for iron extraction in both pathological and physiological conditions.[83,87,88] The pH dependence of the iron removal on human H-ferritin has been recently measured in vitro.[89,90] In the presence of reductants, which are needed to reduce biomineralized Fe^{3+} to the more soluble Fe^{2+} species, lowering the pH from basic to acidic value increases the amount of iron ions released in the bulk solution. We performed molecular modelling and MD simulations to study the release mechanism of Fe^{2+} ions from human homopolymeric FtH ferritin (H-ferritin). The work was focused on the structural and energetic differences between the process at pH 4 and pH 9, with the aim of understanding the determinants of the enhanced Fe^{2+} release in acidic conditions. Thus, it was important to compute the protonation state of titratable groups at pH 4 and pH 9; this was done using the H++ server.[91] The AMBER package was used for molecular dynamics (MD) simulations;[65] details on the MD setup are given in Sala et al.[92]

In our simulations, the Fe^{2+} ions rapidly entered the C3 channels due to the electrostatic gradient going from the cavity to the interior of the channels. The gradient is present at both pH 4 and pH 9. It is the result of two groups of residues making a spatial cluster of negative charges inside the channel. The two groups can be described as two rings of three residues each from as many chains: the innermost (i.e. closer to the cavity) ring made by three Asp residues and the most external one made by three Glu residues. The negatively charged cluster of carboxylates in C3 channels is surrounded by a number of histidine residues with a different protonation state depending on the pH of the simulation. At both pH 4 and 9, a single iron ion reached a position between the two rings of carboxylates in less than five nanoseconds of simulation.

From this point, the Fe^{2+} ions diffused in the bulk solution exclusively at pH 4, whereas at pH 9 they were immobilized inside the C3 channels by the negative charges. Therefore, a key role in the release process was played by the different protonation state of the titratable sites. In fact, the positive charges on the His rings inside the C3 channels (His118, His128 and His136) were crucial to weaken the iron coordination environment. This is presumably due to the compensation of the negative charges of the Asp and Glu carboxylate groups (Figure 1). Further insights on the release process were provided by so-called umbrella simulations, which are used to evaluate the energetic barriers derived from the potential of mean force (PMF) along a reaction coordinate.[93] For ferritin, we simulated the diffusion of the Fe^{2+} ion within the C3 channels, along the trajectory that we observed in unrestrained simulations. The umbrella simulations highlighted two main features: first, the energy barrier of the Fe^{2+} release process was higher at pH 9 than at pH 4. Second, the positions of the energetic minima corresponding to the Fe^{2+} binding site within the channel were different between pH 4 and pH 9. At pH 4, the energetic minimum is significantly closer to the end of the channel, denoting a higher permeability of the channel in acidic than in basic conditions. In addition, we performed an umbrella simulation at pH 9 on a chimeric ferritin in which the His residues inside the C3 channel were doubly protonated as at pH 4. This simulation showed that the double protonation of His118, His128 and His136 at pH 9 did not change the position of the energetic minimum along the C3 channel. Instead, it greatly reduced the energetic barrier preventing the Fe^{2+} ion from leaving the binding site in the channel and diffusing into the bulk solution. In summary, we demonstrated that the compensation of the negative charges inside the C3 channel by the doubly protonated His residues was a critical factor. However, it is only by doubly protonating **all** the His of ferritin at pH 4 that the dynamic opening of the binding site could take place, resulting in a significant shift of the Fe^{2+} coordination site towards the exit of the channel with respect to pH 9.

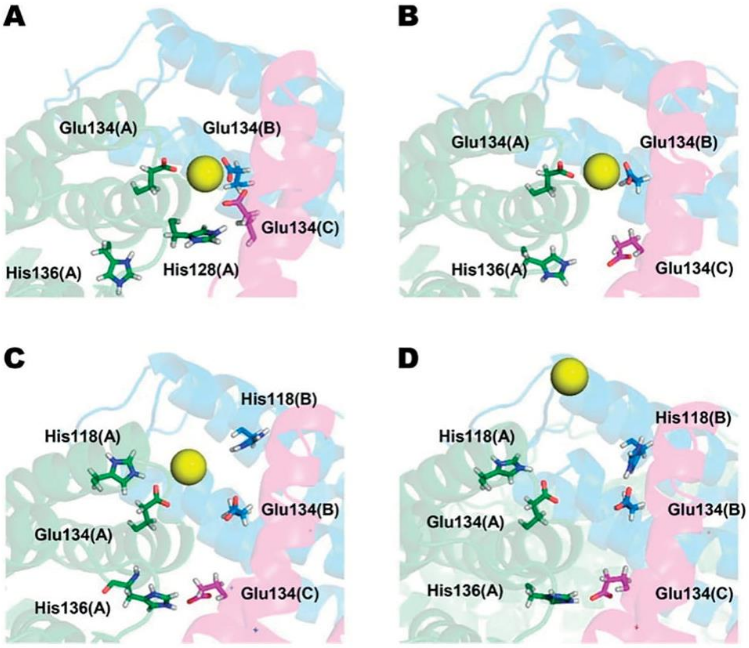


Figure 1. The main steps allowing Fe^{2+} exit from the human H-ferritin via the C3 channels at pH 4. Each chain has a different color; the yellow sphere represents the Fe^{2+}ion. The side chains of the most important residues along the pathway are represented as sticks. [A] Snapshot at 3.00 ns. Three Glu134 residues bind to the Fe^{2+} ion. His128(A) is close to Glu134(C), favoring its subsequent detachment from the ion. (B) Snapshot at 3.63 ns. Glu134(C) points toward the cavity of ferritin: its position allows the interaction with His136(A); the metal is now bound to two rather than three Glu134 residues. (C) Snapshot at 19.05 ns. Prior to the release of the metal there is a reinforcement of the interactions between Glu134 and His118, affecting the coordination of Fe^{2+}. (D) Snapshot at 19.24 ns. Glu134 and His118 from chains A and B engage in two salt bridges; this triggers the exit of the Fe^{2+} ion release from the channel. Reprinted with permission from ref.[86] Copyright 2017 American Chemical Society.

3.2. YiiP and Zn^{2+}

The YiiP transporter from *Escherichia coli* is a relatively well-characterized member of the CDF (Cation Diffusion Facilitator) family. All CDF transporters known so far use an antiport mechanism to export divalent cations.[94] Their architecture comprises a transmembrane domain (TMD) connected to a cytosolic C-terminal domain (CTD). The high-resolution crystal structure[95] (PDB 3H90) showed that YiiP is a homodimer with a Y-shaped structure composed by twelve transmembrane (TM) helices. The six TM helices from each monomer are grouped into two bundles with four (TM1–TM2–TM4–TM5) and two (TM3–TM6) helices, respectively. Each YiiP chain contains three zinc-binding sites (A–C) but only sites A and C are conserved in all CDF members. Site A (the transport site) is positioned near the center of the TMD and constitutes the active site for zinc transport. Sites B and C are in the loop connecting TM2 and TM3 and at the CTD-CTD interface, respectively. The transport site adopts a tetrahedral coordination geometry that involves D45 and D49 of TM2 as well as H153 and D157 of TM5.[96] The structure has four interlocking salt bridges, involving K77 of TM3 and D207 of IL1, connecting the TMD and CTD regions. The charge interlock among the four salt bridges determines the stabilization of the dimer and the positioning of the TM3–TM6 bundle. The low-resolution cryoelectron microscopy structure of the YiiP homolog from *Shewanella oneidensis* (PDB 3J1Z) provided a view of the inward-facing state.[97] It has been proposed that this protein conformation allows the zinc ions to bind the transport sites from the cytosol. The proposed mechanism for this process is mediated by sites A and C. They are connected through an allosteric mechanism by which a perturbation in the CTD region can affect site A in the TMD region. This connection is mediated by the aforementioned charge interlock. In the proposed mechanism, zinc binding in the C site induces a reorientation of the TMD helices, thus allowing access to the transport site. Zinc binding in the transport site triggers the switch from the inward-facing conformation to the zinc-discharging outward-facing conformation. Diffusion of the Zn^{2+} ion to the A site depends on the accessibility to the hydrophobic cavity located in the TMD region facing the intracellular space. Thus, the possibility to reach the transport site A is regulated by the packing of hydrophobic residues belonging to TM3, TM5 and TM6 helices. X-ray-mediated hydroxyl radical labelling and mass spectrometry data suggested that TM5 plays a key role in gating water access to the TMD region through a rigid-body motion that affects the orientation of the Leu152 side chain, in turn modulating the access to the transport site.[98]

We investigated some of the latter aspects by performing MD simulations of the inward-facing conformation in the presence of Zn^{2+} ions in solution (0.150 M of $ZnCl_2$). The transporter was embedded in a DPPC lipid bilayer. In our simulations, one of the zinc ions reached residues Asp49 and His153 in the transport site, elucidating the pathway attracting the metal ion from the cytosol to the TMD region. Zinc binding in the transport site triggered the closing of the side of the TMD that faces the cytosolic space. The motion of the TM1–TM2–TM4–TM5 helices with respect to TM3–TM6 tuned the access to the TMD region; an important contribution came also from the internal dynamics of TM5. In particular, the rotation of TM5 affected the orientation of the hydrophobic Leu152 side chain blocking the water access to the cavity as proposed by Gupta et al.[98] (Figure 2). In addition, umbrella sampling simulations were performed to measure the free-energy profiles of zinc and sodium ions going through the TM channel. In the inward-facing conformation both metals have a favorable energetic profile to reach the transport site. Asp49 and Asp150 provided the coordination site corresponding to the energetic minimum. Only the sodium ion was able to exit the channel with a reasonable energetic barrier, demonstrating that the inward-facing conformation is suitable for zinc access to the transport site but not for its release in the bulk solution.

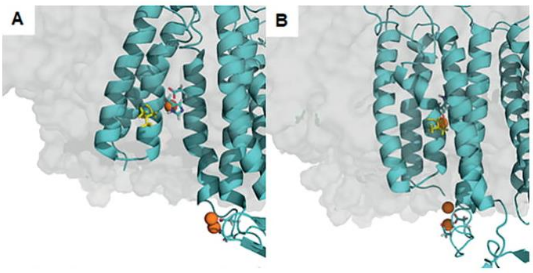


Figure 2. Closing of the TM cavity upon Zn^{2+} ion binding on the transport site of the YiiP antiporter. The orange spheres represent the Zn^{2+} ions. The residues coordinating the metal and Leu152 are represented as sticks. Leu152 is colored in yellow. The DPPC membrane surface is shown in gray. (A) Snapshot at 400 ns. The TM cavity is accessible from the intracellular (IC) space and one zinc ion reached the transport site. (B) Snapshot at 1 μs. The access to the transport site is closed from the IC space.

The case studies reported above define a common strategy for the investigation of the diffusion processes of metal ions using MD. The strategy can be resumed as follow: first, MD unrestrained simulations are performed to collect information about the diffusion pathways, the interactions of the metal ions with the protein frame and the associated conformational changes. Second, when the metal pathway of interest is identified, umbrella simulations provide the energetic barriers along the process. Furthermore, a deeper description of diffusion processes can be achieved by exploiting the unique possibility provided by MD to easily build and investigate the behavior of mutants or chimeras that can probe the role of specific interatomic interactions. In summary, the proposed strategy yields a detailed atomic description of the mechanisms of metal ion transport in turn helping the interpretation of low-resolution structural data and physiological/biochemical data.

4. Molecular Dynamics Studies on Metal-Responsive Transcriptional Regulators

In recent years, the resolution of several crystallographic structures of bacterial metalloregulatory proteins prompted the study of their mode of action by using both experimental and computational methods. In this section, we describe some recent results obtained through the use of molecular modelling techniques.

4.1. The DtxR Family

The DtxR family comprises metallosensors specific for the Fe^{2+} and/or Mn^{2+} ions. They are metal-dependent repressors of transcription. The metal ions are co-repressors that induce the interaction of the regulators with palindromic DNA regions upstream of the set of regulated genes[99,100] (Figure 3A).

In 2007, Bhattacharya et al. studied the diphtheria toxin repressor (DtxR) from *Corynebacterium diphtheriae* by using atomistic MD simulations and ^{15}N-NMR based relaxation measurements.[101] DtxR is the primary regulator of iron homeostasis, oxidative stress resistance, and virulence in Gram+ bacteria. DtxR is composed by two domains linked by a flexible proline-rich (Pr) stretch: a large N-terminal domain responsible for iron binding, dimerization, and DNA binding not considered in the simulations, and a C-terminal domain with structural similarity to eukaryotic SH3 domains. In the iron-bound (*holo*) protein, the Pr stretch associates with and stabilizes the N-terminal domain helices that constitute the dimer interface. In the metal-free (*apo*) protein, the Pr stretch is in contact with the SH3 domain, resulting in an intramolecular complex[102] (Pr–SH3) stabilizing the monomeric form of the protein. In both forms, the dynamics of the SH3 domain featured ordered secondary structures with high generalized order parameters and low effective correlation times. Instead, the amino acids of the loops connecting the β-strands featured reduced generalized order parameters, with extra motional terms needed to correctly model the relaxation rates. The amino acids in the Pr region had low order parameters and long effective correlation times. In addition, the SH3 domain exhibited motions on the ms time scale, whereas the Pr region showed motions on the 0.1 ms time scale. The MD simulations conducted on the Pr-SH3 complex, the SH3 domain and the Pr region highlighted structural changes that could provide a contribution to the observed relaxation rates. This might indicate that the Pr region is involved in a binding-unbinding equilibrium. However, the absence of the metal binding N-terminal domain in the simulations precluded the possibility of a complete picture of the molecular mechanism of action of DtxR.

In 2015, Ghosh et al. published an extensive study of IdeR from *Mycobacterium tuberculosis*,[103] another member of the DtxR family binding up to two Fe^{2+} ions per monomer. The IdeR

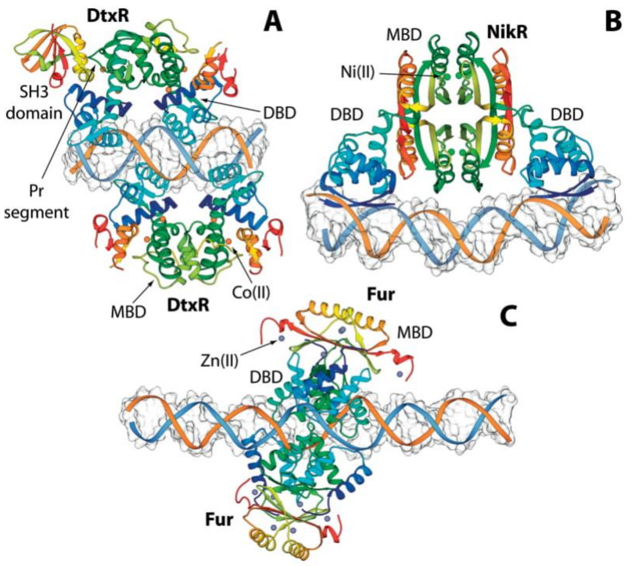


Figure 3. Crystal structures (A, B) and model structure (C) of the metal sensor/DNA complex of DtxR, NikR and Fur, respectively. The metal sensors are reported as ribbons colored from blue to red going from the N- to the C-terminal. Metal ions are depicted as spheres colored accordingly to the atom type. The DNA double strand is reported as light blue and orange ribbons. PDB code: 1C0W and 2HZV for panel A and B respectively; in panel C is reported a model complex from ref.[124]

monomer contains three domains: the DNA binding domain (DBD), the dimerization domain (DD), and the SH3-like domain (SH3). While the function of the DBD and DD domains is relatively clear, the role of the SH3-like domain is still object of debate. Two monomeric subunits form a dimer that binds to DNA, mainly via the DNA-binding helices. Two IdeR dimers bind to opposite sides of a 19 base-pairs-long conserved DNA sequence (the "iron box"[104,105]) resulting in a dimer-of-dimer complex. The authors performed atomistic MD simulations in explicit solvent of several systems in different metallation states and without or with DNA. The parametrization of the Fe^{2+} centers were conducted by using the Metal Center Parameter Builder[106] (MCPB). The simulations were integrated with protein structure networks to characterize the impact of Fe^{2+} binding on the 3D structure and function of the metalloregulator. A significant structural variation between the *apo* and the *holo* systems was observed. In particular, the simulations suggested that Fe^{2+} improves the stability of the monomeric subunit, which in turn enhances dimerization. By simulating the dynamics of the IdeR-DNA complex without Fe^{2+}, the authors were able to observe the unbinding of the protein subunits from the DNA after 100 ns of simulation. Moreover, Fe^{2+} acts as an allosteric regulator of IdeR that enhances the formation of the complex between the metalloregulator and DNA.

4.2. NikR

NikR regulates the expression of genes encoding proteins involved in nickel metabolism.[4,107,108] About 30 species of bacteria and archaea are known to have NikR homologs. The Ni^{2+}-bound form of the protein represses the nikABCDE operon, which encodes an ABC transporter responsible for selective Ni^{2+} uptake.[109,110] Moreover, NikR from *Helicobacter pylori* is a pleiotropic regulator of several genes. Among its nickel-dependent roles, it functions as a repressor of *H. pylori* NikR itself and of the Ni^{2+} permease NixA, and as a promoter of the expression of urease.[111]

NikR[4] has a homodimer-of-homodimer structure, which can be regarded as an ensemble of three domains (Figure 3B). The first one is the central metal-binding domain (MBD), involving the C-terminal portion of the four monomers forming the quaternary structure of the protein. The MBD harbors four regulatory metal-binding sites symmetrically located at the tetramerization interface. The coordination geometry around each Ni^{2+} ion is square planar, with three fully conserved His and one Cys residues coordinating the metal. This geometry is maintained also thanks to the ligand field stabilization of the d^8 Ni^{2+} ion. The network of hydrogen bonds around the metal-binding site is the likely means through which the occupancy of the metal site modulates the overall structural and dynamic behavior of NikR. The MBD is connected to two peripheral DNA-binding domains (DBD) by flexible linkers. Each DBD is formed by the N-terminal portions of two monomers of the metallosensor and contains a ribbon-helix-helix motif, as often encountered in prokaryotic transcription factors.[112] The available crystal structures of NikR showed that this metallosensor can adopt three alternative conformations: open, *trans*, and *cis*, differing because of the relative position of the DBDs with respect to the MBD. The crystal structure of the NikR-DNA complex from *Escherichia coli* demonstrated that the *cis* conformation of the *holo* sensor can bind DNA[113] (Figure 3B), a result confirmed also by a computational model built on the basis of chemical shift perturbation data.[114]

Atomistic MD simulations of *apo* NikR from *E. coli* identified a number of residue interrelationships that constitute an allosteric communication pathway between the metal- and DNA-binding sites, which are as far as 40 Å.[115] Additional atomistic MD calculations in implicit solvent are available for NikR from *Pyrococcus horikoshii* in the *apo*-open, *holo-trans*, and *apo-trans* forms.[116] The analysis of these three 100 ns-long simulations highlighted a slow conformational kinetics and independent motion of the DBDs. A noticeable relationship was observed between structurally and biological important regions. Finally, also the effect of Ni^{2+} on DNA binding by *H. pylori* NikR has been studied using several atomistic MD simulations in explicit solvent together with experimental data from solution NMR spectroscopy. The protein was simulated in the *apo*, $2Ni^{2+}$-bound and $4Ni^{2+}$-bound forms and starting from all the possible NikR conformations.[117] From these simulations it appears that both *apo*- and *holo*-NikR exist in solution as inter-converting conformations sampling the whole space from the *cis* to the open to the *trans* conformations observed in the crystal. Metal binding does not result in the stabilization of the *cis* conformation that has high affinity for DNA. Rather, it enhances the relative dynamics of the DBDs with respect to the MDB. These results support the view that the binding of the metallosensor to its operator DNA is mediated by conformational selection together with an induced fit mechanism assisted by the coordination of the metal.[117] In the latter two studies, the Ni^{2+} binding site was modelled through a bonded approach.[116,117]

4.3. The MerR Family

Metallosensors belonging to the MerR family are activators in their *holo*-forms. They induce the expression of genes coding for proteins involved in metal export and detoxification.[13] The MerR family includes metal sensors able to bind monovalent or divalent cations.[4] Members of the MerR family bind their DNA operator sequences regardless of the presence of the metal and activate gene expression at low metal ion concentrations.

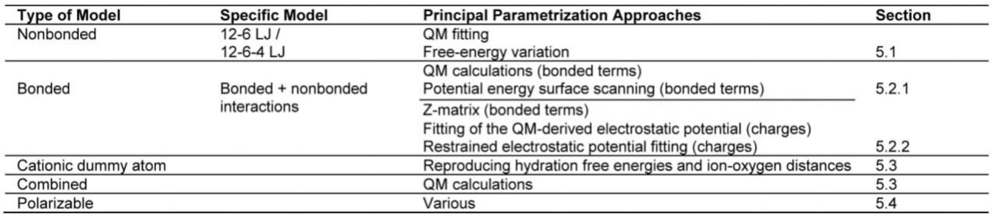
In 2010, Guo and co-workers studied the archetypal member of this family, MerR from *Shigella flexneri*, upon Hg^{2+} binding.[118] In particular, the authors performed 50 ns-long atomistic MD simulations in explicit solvent on the model structure of the *apo* and *holo* forms of the protein. The Hg^{2+}-binding site was modelled by the Automatic Frequency Matching Method[119] and potential energy scanning to parameterize the potential function from quantum chemical calculations for the metal in a trigonal-planar environment. The structure was predicted to be an all-alpha homodimer formed by six α-helices. Helices $\alpha1$ and $\alpha2$ constitute the DBD, and helices $\alpha3$ and $\alpha4$ comprise the "coupling domain" (CD). The CD mediates the communication between the Hg^{2+}-binding site and the DBD. The dimerization interface is constituted by an *anti*-parallel coiled coil dimer involving helix $\alpha5$; two Hg^{2+}-binding sites are located at the two ends of the dimer. The simulations, corroborated by SAXS data, revealed that the average torsional angle between the DBDs is about 65°. The interdomain motions occur on a timescale of ca. 10 ns and involve domain opening-and-closing movement, together with changes of the torsional angle between the DBDs. This correlated motion may propagate from the Hg^{2+}- to the DNA-binding site.[118]

4.4. The ArsR/SmtB Family

The ArsR/SmtB family is the most widespread group of known metalloregulators. It includes over 500 transcriptional regulators able to bind a plethora of metal ions.[4] In general, the members of the ArsR/SmtB family are repressors in their *apo* form, recognizing and binding an imperfect 12–2-12 inverted repeat on DNA within the promoter region. DNA-binding prevents transcription by the RNA polymerase. The binding of the cognate metal ions lowers the affinity of the regulator for DNA, resulting in the dissociation of the protein:DNA complex.[120]

In 2012, Lee *at* al. solved the solution structure of homodimeric *apo*-NmtR from *Mycobacterium tuberculosis*.[121] Binding of Ni^{2+} to the regulatory sites induces a very significant paramagnetic broadening of the NMR signals from the first ten residues as well as from amino acids in helix $\alpha5$. Atomistic MD simulations were performed on the apo, Zn^{2+}- and Ni^{2+}-bound form of the protein in explicit solvent. The Zn^{2+} and Ni^{2+} binding site were parametrized through quantum chemical (QM) computations. The calculations revealed that NmtR accommodates Zn^{2+} ions in a pseudo-tetrahedral geometry and Ni^{2+} ions in a pseudo-octahedral geometry involving the α-amino group, His3, His104 and His107 from the helix $\alpha5$ of the first monomer, together with residues Asp91, His93 from the helix $\alpha5$ of the second monomer. In the same year, Chakravorty and co-work-

Table 1. Main methods used for the parametrization of d-block metal ions. Only the approaches described in this review are listed. For each group of models (type of model), we indicated the specific model discussed (if relevant) and then the corresponding parametrization approaches. The last column indicates the section of this review where the method is discussed.



Type of Model	Specific Model	Principal Parametrization Approaches	Section
Nonbonded	12-6 LJ / 12-6-4 LJ	QM fitting Free-energy variation	5.1
Bonded	Bonded + nonbonded interactions	QM calculations (bonded terms) Potential energy surface scanning (bonded terms)	5.2.1
		Z-matrix (bonded terms) Fitting of the QM-derived electrostatic potential (charges) Restrained electrostatic potential fitting (charges)	5.2.2
Cationic dummy atom		Reproducing hydration free energies and ion-oxygen distances	5.3
Combined		QM calculations	5.3
Polarizable		Various	5.4

ers studied the Zn^{2+} sensor CzrA from *Staphylococcus aureus*.[122] The study was conducted through classical and quantum mechanical/molecular mechanical (QM/MM) MD simulations. The Zn^{2+} site was parametrized using the MTK++ program.[123] Zn^{2+} binding to apo CzrA induces the transition from a "closed" conformation to an "open" conformation that reduces the DNA binding affinity of the protein. The simulations could reproduce the closed to open transition for DNA-bound *holo* CzrA and provided a view of the residues involved in correlated and anticorrelated motions underlying the communication from the zinc- to the DNA-binding sites. The authors suggested that zinc binding restricted the conformational space available to the regulator, resulting in a suboptimal configuration of the DNA binding interface.

4.5. The Fur Family

Proteins belonging to the Fur family of dimeric metalloregulators usually are repressors in their *holo* form. When coordinating the cognate metal ions, they prevent transcriptional initiation by binding their target DNA within the promoter region. The metal ions are co-repressors. The progenitor of this family (the ferric uptake regulator, Fur) binds Fe^{2+} ions, while other members of the family can bind Zn^{2+} (Zur), Ni^{2+} (Nur) or Mn^{2+} (Mur) ions.[4]

Fur regulates iron metabolism and its coupling with the response to oxidative stress. In *H. pylori*, Fur acts as a transcriptional commutator switch by repressing the expression of iron-inducible and iron-repressible genes, respectively in the *apo* and in the *holo* form, respectively. In iron depletion conditions, Fur binds to DNA in the *apo* form as a dimer, repressing the transcription of ferritin, a protein dedicated to iron storage. Conversely, when iron is in excess, *holo* Fur binds to different DNA operators as a tetramer and represses the expression of genes involved in iron uptake. Two models of Fur-DNA complexes in the *apo* and *holo* forms have been proposed by Agriesti and co-workers in 2014 through a knowledge-based macromolecular docking protocol[124] (Figure 3C). In the same year, Cissé et al. proposed the use of a peptide derived from *anti*-Fur peptide aptamers as a possible Fur inhibitor.[125] The peptide-Fur complex was built by using a docking protocol combined with MD simulations in implicit solvent. The calculations were validated with experimental information provided by single points mutations.

4.6. The CsoR/RcnR Family: A Target for Future Studies

The members of this family bind Cu^{+} (CsoR) or Ni^{2+} and Co^{2+} (RcnR) or only Ni^{2+} (InrS).[4,108] When the concentration of the cognate metal ion increases, it binds to the metal sensor and reduces its DNA affinity; this permits the expression of the proteins devoted to decreasing the metal concentration. The models of the metal-site structures of RcnR from *E. coli* bound to Ni^{2+}, Co^{2+}, and the non-cognate metal Zn^{2+} have been recently proposed on the basis of spectroscopic and functional data obtained both on the wild type protein and on several mutants.[126]

4.7 Common Aspects of the Investigation of Metal-Responsive Transcriptional Regulators

The present-day challenge is to understand the molecular basis of the change in DNA affinity upon metal ions binding/unbinding to metal-dependent transcriptional regulators. Together with the experimental techniques, molecular modelling is the key for gaining insights into this complex mechanism. The increasing number of crystal structures, the improvement of the computational tools, the availability of accurate models for the metal ions and, last but not least, the growth of computational power can offer a deep and detailed characterization of the molecular determinants underlying the conformational changes needed to specifically repress/derepress one or more genes. Applied to pathogenic bacteria, and in particular to those species known to be a danger for their increasing antibiotic resistance (see http://www.who.int/medicines/publications/global-prioritylist-antibiotic-resistant-bacteria/en/), these findings can pave the way to the development of new classes of drugs and antibiotics targeting gene expression.

5. Computing Force Field Parameters for the Metal Site

The force fields commonly used in MD simulations do not include d-block metal ions. Thus, ad hoc parametrization methodologies must be deployed. In this section we describe the most popular approaches used for the parameterization of metal ions (for a complete and exhaustive review see ref.[127]). Table 1 summarizes the main methods addressed here.

5.1. Classical Approaches: The Nonbonded Model

The 12–6 Lennard-Jones (LJ) potential is the most popular one for the nonbonded interactions.[128] The potential for two atoms *i* and *j* is given by the sum of the electrostatic and VDW terms, as described in Equation (3) using three different formulations

$$\begin{aligned} U(r_{ij}) = E_{ele} + E_{VDW} &= \frac{Q_i Q_j}{4\pi\varepsilon_0 r_{ij}} + \frac{A_{ij}}{r_{ij}^{12}} + \frac{B_{ij}}{r_{ij}^{6}} \\ &= \frac{Q_i Q_j}{4\pi\varepsilon_0 r_{ij}} + \varepsilon_{ij}\left[\left(\frac{R_{min,ij}}{r_{ij}}\right)^{12} - \left(\frac{R_{min,ij}}{r_{ij}}\right)^{6}\right] \\ &= \frac{Q_i Q_j}{4\pi\varepsilon_0 r_{ij}} + 4\varepsilon_{ij}\left[\left(\frac{\sigma_{ij}}{r_{ij}}\right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}}\right)^{6}\right] \end{aligned} \quad (3)$$

In this equation, r_{ij} is the distance between the two atoms, Q_i and Q_j are their partial charges. In this model, the metal ion is usually assigned an integer charge corresponding to its oxidation state. Just a pair of parameters must be evaluated, namely A_{ij} and B_{ij}, $R_{\min,ij}$ and ε_{ij}, or $\sigma_{\min,ij}$ and ε_{ij} subject to the specific notation used, i.e. two parameters per pair of atom types. Note that $R_{\min,ij}$ is the distance where the LJ potential reaches its minimum, σ_{ij} is the distance where the value of the potential becomes zero and ε_{ij} is the well depth of the functional profile. Because in a system with *N* atom types there will be as many as *N*(*N*+1)/2 unique atom pair types, which entails a considerable parametrization effort, different rules have been introduced to derive *ij* pair parameters from parameters specific of each individual atom type (e.g., ε_{ij} can be computed by combining ε_{ij} and ε_{ij} values according to a predefined rule). In this way, only two parameters per atom type (e.g., ε and σ_{min}) are needed (2 N in total). The resulting potential is a good approximation especially for low electronegative monovalent ions, where the polarization and charge transfer effects are small.[129,130]

The parametrization of ions is usually performed trying to reproduce one or more properties among: thermodynamic quantities, structural properties and kinetic or dynamic properties. The parametrization can be derived using different computational methods such as force field or single point energy calculations, Monte Carlo[131] or MD simulations.[132] The transferability of the parametrization determined by fitting against specific properties should be tested before the model is applied in different, often more complex, systems, e.g. in biological systems where the ion typically encounters dissimilar ion/atom types in combination.

Instead, one of the main drawbacks of potentials derived from QM fitting is that they typically consider solely the first water molecule; this can result in the effective potential for ion–water interactions being overestimated. The latter overestimation is in part due to the many-body effects being neglected. Therefore, the resulting hydration free energies (HFEs) and the coordination numbers (CNs) are in general higher than those observed experimentally. Working with small basis sets may reduce the error because they underestimate the interaction energy of the ion bound to a single water molecule, even though they are not effective on highly charged metal ions for which the many-body effect is noteworthy. Various approaches can be deployed to circumvent the many-body issue: (A) introducing many-body terms; (B) working with an effective potential; and (C) working with a combined model. Two further models were developed specifically for zinc-containing proteins: the short-long effective function (SLEF) model and the quantum calibrated polarizable-charge transfer (QPCT) model. SLEF was parametrized reproducing QM/MM-derived forces for zinc-enzymes with various coordination environments.[133,134] In QPCT, a CT term is added to the nonbonded model in order to match the interaction energy between the metal and one of the ligands.[135]

Parameterizations derived from experimental data are often effectively based on free energy. From the computational chemistry point of view, two methods are particularly noteworthy in estimating free-energy changes: the free-energy perturbation[136] (FEP) and thermodynamic integration (TI) methods.[137] Other approaches trying to evaluate the variation of free energy are the MM Poisson–Boltzmann surface area[138] (MM/PBSA), MM generalized Born surface area (MM/GBSA) and linear response approaches.[139]

As introduced above, the LJ model is a good approximation for the simulation of ions in solution but has also some limitations. For example, an important factor is the water model used. In general, the error of this model increases with the square of the ion charge. Thus, for metal ions having a charge of +2 or higher it is often hard or infeasible to achieve a parametrization that results in an acceptable agreement with experimental data such as HFE and the ion–oxygen distance (IOD) of the first water shell. An easy way to reduce errors in reproducing structural properties is to refit the parameters to satisfy specific experimental features. Obviously, this approach leads to bad performance in reproducing macroscopic or kinetic properties. A drawback particularly relevant for multivalent ions is that classical approaches neglect the charge-induced dipole and dipole-induced dipole interactions. To correct this, a new 12–6-4 LJ model was proposed in 2014 and applied to parametrize various metal ions in conjunction with three common water models.[130,140,141] The new 12–6-4 model has an additional term representing the ion-induced dipole interaction [compare Equation (3) above with Equation (4)].

$$
\begin{aligned}
U(r_{ij}) &= \frac{Q_iQ_j}{4\pi\varepsilon_0 r_{ij}} + \frac{A_{ij}}{r_{ij}^{12}} + \frac{B_{ij}}{r_{ij}^{6}} + \frac{C_{ij}}{r_{ij}^{4}} \\
&= \frac{Q_iQ_j}{4\pi\varepsilon_0 r_{ij}} + \varepsilon_{ij}\left[\left(\frac{R_{min,ij}}{r_{ij}}\right)^{12} - 2\left(\frac{R_{min,ij}}{r_{ij}}\right)^{6}\right] + \frac{C_{ij}}{r_{ij}^{4}} \\
&= \frac{Q_iQ_j}{4\pi\varepsilon_0 r_{ij}} \\
&\quad + \varepsilon_{ij}\left[\left(\frac{R_{min,ij}}{r_{ij}}\right)^{12} - 2\left(\frac{R_{min,ij}}{r_{ij}}\right)^{6} - 2kR_{min,ij}^{2}\left(\frac{R_{min\,ij}}{r_{ij}}\right)^{4}\right] with\ k = \frac{C_{ij}}{B_{ij}}
\end{aligned}
\tag{4}
$$

Similarly, classical force fields do not include polarization effects, because they use simple point charges. The charge scaling (CS) approach can be applied to introduce long-range screening effects into the picture. A successful application of this concept led to the MD electronic continuum (MDEC) approach, which can be applied to organic molecules,[142] water,[143] and ionized systems.[144] Furthermore, some common models of water underestimate its dipole moment. A model of water based on the CS approach features fluctuating charge and has the advantage to be suitable for use in conjunction with unpolarized force fields like CHARMM or AMBER.[145] This model can recapitulate the different polarization of water molecules binding metal ions, at the macromolecular surface, and in the bulk phase. Therefore, the CS approach can model some properties that the classical unpolarized force fields cannot model. As it does not introduce additional terms, it preserves the MD performance. However, the use of the model is still limited by the fact that it corresponds to an average perspective suitable to predict some statistical properties but unable to reproduce the microscopic environment.

5.2. Classical Approaches: The Bonded Model

5.2.1. Bond, Angle, and Torsion Parameters

Equation (5) describes the well-known combination of terms describing bonded and nonbonded interactions. In the first group

$$
\begin{aligned}
U = &\sum_{bonds} k_i(r - r_{eq})^2 + \sum_{angles} k_\theta(\theta - \theta_{eq})^2 \\
&+ \sum_{torsions}\sum_{n=1,2,3,4} V_n[1 + \cos(n\Phi - \gamma)] + \sum_{i,j>i}^{N}\left(\frac{q_iq_j}{4\pi\varepsilon_0 r_{ij}} + \frac{A_{ij}}{r_{ij}^{12}} + \frac{B_{ij}}{r_{ij}^{6}}\right)
\end{aligned}
\tag{5}
$$

we find bond, angle and dihedral terms; in the latter the electrostatic and VDW terms [grouped in the fourth summation of Equation (5)]. In addition to the nonbonded parameters discussed in the previous section, the force field contains several other parameters for the bonded terms: for each group of bonds, the k_r (force constant) and r_{eq} (equilibrium length); for each group of angles, k_θ (force constant) and θ_{eq} (equilibrium value); for each group of torsion angles, V_n (energy barrier), n (periodicity), and γ (phase); for each atom, its partial charge q.

The quality of a force field is largely determined by nonbonded terms, with atomic charges and VDW parameters assigned to represent different properties at the same time. The most widely used classical force fields for biological systems are based on the same functional form with some differences in the LJ terms. The parameters for dihedral and improper torsion angles have defined minima depending on the repulsions between bond electrons.[146] However, in the case of metal ions all angle values are considered accessible at physiological temperature; therefore the corresponding energy barriers are normally set equal to zero.[123,147,148]

The commonly used force fields share a harmonic representation of the bond and angle terms (with an additional 1–3 term in CHARMM[149]), resulting in a good parameter transferability among them. More sophisticated force fields contain higher power as well as cross terms.[150,151] The corresponding parameters can be obtained in multiple ways, but we will focus on methods effective for biological systems. Small representative molecules are normally used to derive the parameters based on agreement with QM or experimental data. The resulting parameters are then transferred on larger systems as macromolecules. As supported by various experiments, this basic transferability assumption is a good approximation for equilibrium bond and angle distances, whose values are often similar in different environments. However, in QM calculations the bonded and nonbonded interactions are coupled together, making it difficult in some cases to disentangle the two contributions.

The potential energy surface (PES) scanning method is one of the simplest approach to parametrize force fields. The force constants are optimized based on their match with PESs obtained from QM computations. The PES scanning method was applied to develop a force field suitable for use with AMBER for various heme species.[152] Despite the simplicity, this approach needs a PES scan for every bond and angle that are part of the metal environment, making the computational cost considerable.

The Z-matrix method relies on the calculation of the Hessian matrix. This step can be performed in different ways such as calculating the derivative of the energy either numerically or analytically, with the latter more widely used in combination with the DFT level of theory.[153,154] The Z-matrix method has some known limits. First, it is dependent on the internal coordinates chosen.[155] Second, the derived force constants might require corrections to fit the QM calculations because only the harmonic terms are considered in the method, thereby resulting in a partial match to the PES. Finally, the calculated dihedral force constants cannot be used within modern force fields models for two reasons: the harmonic potential form is not applicable to the Fourier expansion used in common classical force fields [Equation (5)] and the dihedral 1–4 interactions are already fully or partially considered. The Seminario method calculates force constants using a sub-matrix of the Cartesian Hessian matrix[155]. However, these constants incorporate the influence of the environment and may not be suitable for isolated bonded terms. Furthermore, the dihedral and improper torsion parameters are in the harmonic form. This limit is overcome by setting the dihedral and improper torsion terms to zero while keeping bond and angle parameters (see above). To date, several metal ion parameters are present in literature based on the use of Seminario method within the MCPB toolkit.[31] For instance, the zinc AMBER force field (ZAFF) was developed for 10 sites where the donor atoms around the metal ion featured tetrahedral geometry.[123] Furthermore, the Seminario method was recently used for the development of a python based metal center parameter builder[106] and the visual force field derivation toolkit[156] (VFFDT).

The automated parametrization method (APM) is a simple approach to obtain parameters directly scanning the parameter space and comparing the resulting properties with experimental or QM data. APM has the advantage of not presenting the double counting issue of bonded and nonbonded interactions already mentioned.[157] However, the number of parameter combinations raises exponentially with the parameter space dimension. This problem can be handled in different ways: (i) dividing the parameter set in subsets to fit separately on the basis of chemical and physical considerations or (ii) performing the parameter optimization as a geometry optimization. A number of freely available APM programs are available, such as Parmfit and ForceBalance.[158,159]

5.2.2. Charge Parameters

There are four main methods to produce charge parameters for classical force fields: empirical methods, fluctuating charge method (FQ), experimental derivation methods and approaches based on QM fitting. Empirical charge methods are time consuming and perform poorly when addressing transition and excited states. The FQ model is a relatively straightforward polarizable model (see later); its main shortcomings are related to charge transfer and application to macromolecules. In experiment-based methods, the charges are computed from XRD experiments and data on dipole moments.[160] This approach requires structures with very high resolution and low B-factors; in return, it provides a reference for scaling the point charges of atoms or for assessing the charges obtained by different theoretical methods.[161] QM-based methods are probably the most popular ones to infer charge parameters; below, we will focus only on them.

Partial charges can be obtained with QM methods by fitting the electrostatic potential (ESP) (e.g., CHELP,[162] CHELPG[163]), through molecular orbital analysis (e.g., Mulliken,[164–167] NPA[168]), from the electron density in real space (e.g., ADCH[169]), and by postprocessing or scaling approaches (e.g., AM1-BCC,[170] CM4,[171] CM5[172]). Despite their clear physical meaning, Mulliken charges cannot be used straightforwardly in common force fields because they fail to replicate molecular multipolar properties. Thus, these charges are often scaled to make them suitable for use in MD calculations.

Multipolar properties and electrostatic interactions of molecules are well replicated by the partial charges provided by ESP fitting. Thus the latter approach is one of most commonly used.[173] Several algorithms and weights are available to choose ESP points and their spatial location for fitting. Early works found that selecting points in a distance from the Van der Waals surface of the metal up to 3 or 7 Å provide reasonable results.[174,175] There are two equally good methods to pick ESP points: a cube-based or a layer/shell-based method. The CHarges from Electrostatic Potentials (CHELP) approach uses points that are symmetrically distributed on four spherical shells from the VDW surface separated by 1 Å, with 14 points on each shell.[162] The CHarges from ELectrostatic Potentials using a Grid-based (CHELPG) strategy picks the points between the VDW surface of the molecule and a surface 2.8 Å away based on a cubic grid where the points are separated by 0.3 Å.[163] ESP-derived charges have some limits. They are usually derived from a single configuration, thereby they do not take into accounts charge fluctuations caused by conformational dynamics. Moreover, deriving ESP charges for macromolecules has a high computational cost. In addition, this methodology may underestimate the charge of buried atoms and overestimate the charge of nonpolar atoms.

The restrained ESP (RESP) charge fitting procedure is the preferred method to compute the partial charges in metalloproteins. Using the RESP algorithm at the HF/6-31G* level of theory reproduces better solvation effects.[176] RESP charges for metal complexes can be derived using only one configuration, thanks to the rigidity of the metal–protein bonds. In doing this, the DFT functional B3LYP with the 6-31G* basis set has shown excellent performance. Nevertheless, an effective core potential basis set can be preferable, e.g. for the parametrization of clusters bearing an overall negative charge.[157,177,178]

5.3. Classical Approaches: The Cationic Dummy Atom and the Combined Model

The cationic dummy atom model (CDAM) representation distributes the total mass and charge of the metal ion among the metal center (core) and a number of dummy atoms surrounding the core. The exact number of dummy atoms depends on the coordination number of the metal under study. This model allows a great flexibility in the parametrization procedure while having the advantage to better reproduce the native coordination of ions in solution. The CDAM was originally described by Åqvist and Warshel[179] for Mg^{2+} and then extended to other divalent ions as Ca^{2+}, Mn^{2+}, Fe^{2+}, Co^{2+}, Cu^{2+}, Zn^{2+} and recently for Ni^{2+} ions.[180–182] In the classical octahedral representation, the core site has a −1.0 charge while each dummy atom bears a partial charge of +0.5. Furthermore, a weak repulsive LJ term was added to prevent coordinating atoms from getting too close to charge-carrying dummy sites. The CDAM approach can reproduce experimental HFE and IOD data together, improving upon nonbonded models.[183] Nevertheless, the fractional charges might reduce the energy of the interaction between the metal and surrounding amino acids. In this case the classical 12–6 LJ nonbonded model could work better.[184] More recently, consistent CC models have been developed for eleven transition metal ions and different water models through an elaborate procedure that in principle allows parameters to be compatible with both periodic and non-periodic simulation conditions.[185]

An alternative approach is the combined model, where the ion and the first water shell around it are combined in a single unit for MM calculations.[186] This approach has the advantage of immediately incorporating many-body effects in the first water shell, while taking into account electronic state crossing. The quality of the resulting potential is linked to the quality of QM calculations used to derive the parametrization. Moreover, the combined model representation works well solely for ions with a charge of +2 or higher and relatively short MD simulations, due to the long mean residence time of the water molecules in the first shell. To date, combined models were developed for various highly charged ions, including Mg^{2+},[187] Zn^{2+}[188,189] and Cu^{2+}.[190]

5.4. Classical Approaches: The Polarizable Model

Polarization is important for many interatomic interactions, especially involving highly charged systems, where it becomes more pronounced.[130] As consequence, the polarizable model was developed to account for charge delocalization as a function of the coordination environment. Three fundamental approaches to the development of polarizable models are popular: the FQ model,[160] the Drude oscillator model,[191] and the induced dipole model.[192] Polarizable force fields are applicable in simulations through various strategies: the self-consistent field (SCF),[193,194] matrix inversion,[195,196] and extended Lagrangian methods.[197,198] Currently, the latter is the preferred strategy to perform MD simulations as it has a good accuracy and a reasonable computational cost. A few polarizable models for various metal ions have been published so far.[199]

5.5. Different Approaches for Parametrization Applied to the Zn^{2+} Ion

There are several examples of applications of metal ion parametrization. Here we concentrate on the Zn^{2+} ion, one of the most common d-block metal ions found in proteins and, concomitantly, the ion for which most parametrization examples are available.

Tuccinardi et al. studied matrix metalloproteinases with a classical force field.[200] The parametrization of the metal ion was performed as follows: the partial charges were derived with the RESP algorithm; the force constants for the bonds between Zn^{2+} and the donor atoms were derived with the *Z*-matrix method; the force constants for the planar angles of the metal site were derived by reducing the relevant bond force constants by a factor 10. An optimization of the parameters against the relative energies of various configurations computed by QM methods was carried out to improve the parametrization.

Lin and Wang applied the Seminario approach to 18 complexes binding Zn^{2+} to parametrize the constants for bonds and angles, whereas the RESP approach provided the charge parameters.[147] By evaluating these parameters sets with various combinations of the bonded/nonbonded model, using also formal charges in addition to RESP charges, it appeared that the best performance was achieved by the bonded model in conjunction with RESP charges.

Zn^{2+} ions are also the focus of several papers related to the specific development of appropriate charge parameters. In particular, already in the '90s the ESP approach was applied to the investigation of human carbonic anhydrase II.[201–203] For this parametrization, the equilibrium values of bonds and angles were taken from QM semi-empirical calculations and the corresponding force constants from experimental data; the nonbonded parameters were adjusted to match QM calculations for the complexes of zinc with carbon dioxide and with water. These results highlighted the significant redistribution of electron density over the coordination environment of Zn^{2+}. Indeed, according to the data obtained for the $[(\text{imidazole})_3Zn\text{–}OH]^+$ complex, the partial charge on the divalent metal ion was as low as +0.5.[202]

Hoops et al. investigated different charge modelling approaches on the model of a zinc center coordinated by the nitrogen atoms of the imidazole rings of three histidine residues and by one hydroxide.[201] The ESP charges for two system representing the high and low pH forms of the system ($[Zn(NH_3)_3H_2O]^{2+}$ and $[Zn(NH_3)_3OH]^+$, respectively) were calculated at the MNDO, HF/STO-3G, HF/MINI-4, HF/MIDI-4, and HF/6-31G* levels of theory on the semi-empirical MNDO optimized geometries. The HF/STO-3G computations resulted in small partial positive charges for the Zn^{2+} ion (+0.68 and +0.48 for the high and low pH forms, respectively), while the HF/MINI-4 calculations gave partial charges more similar to the whole ionic picture (+1.64 and +1.72 for the high and low pH forms, respectively) Taken together, these results suggest that in the case of the bonded model is not advisable to assign a +2 point charge to the Zn^{2+} ion.

CDAM parameters for the Zn^{2+} ion have been derived in a study involving human carbonic anhydrase with the empirical valence bond approach.[204] The model involved six dummy atoms, each bearing a +0.5 charge. The simulations showed a good agreement with experimental data, supporting the crucial role played by electrostatic interactions in for the catalytic metal ion.

Also Pang and co-workers applied the CDAM approach to zinc-proteins in a series of studies.[205,206] As early as 1999, a CDAM model for the Zn^{2+} ion was developed by considering four dummy atoms sharing the +2 total charge placed in a tetrahedral geometry and a neutral nucleus. No VDW parameters were assigned to the dummy sites. This CDAM model was designed to tackle the underestimation issued observed in the case of Zn^{2+}–H_2O interactions in the nonbonded model. This model underestimated the Zn–S distances when compared to experimental data but achieved a significant improvement in the prediction of the HFE value. Moreover, the interaction energy of the Zn^{2+}–H_2O dimer was modelled more accurately by the CDAM model than the nonbonded model. The transferability of this CDAM model was demonstrated through calculations carried out on carboxypeptidase, carbonic anhydrase II, and rubredoxin.

Finally, Pappalardo and Marcos,[188] who fitted a potential for the Zn^{2+} hydrate ion based on QM calculations, proposed the first combined model of a metal. The obtained hydration energy for Zn^{2+} (–517.6 kcal/mol) is in excellent agreement with the experimental value (–491 kcal/mol). Interestingly, in the case of the combined model second shell water molecules exhibited less directionality with respect to the calculation executed employing a nonbonded model. The authors proposed that this effect can be reduced by using a flexible water model.[188]

Acknowledgments

We thank Dr. Andrea Giachetti for many useful discussions.

Keywords: Zinc · Iron · Molecular dynamics · Metal transport · Metal storage · Force field

[1] K. J. Waldron, J. C. Rutherford, D. Ford, N. J. Robinson, *Nature* **2009**, *460*, 823–830*)).*

[2] C. Andreini, I. Bertini, A. Rosato, *Acc. Chem. Res.* **2009**, *42*, 1471–1479.

[3] L. A. Finney, T. V. O'Halloran, *Science* **2003**, *300*, 931–936.

[4] B. Zambelli, F. Musiani, S. Ciurli, *Met. Ions Life Sci.* **2012**, *10*, 135–70.

[5] L. Banci, I. Bertini, K. S. McGreevy, A. Rosato, *Nat. Prod. Rep.* **2010**, *27*, 695.

[6] Y. Valasatava, A. Rosato, N. Furnham, J. M. Thornton, C. Andreini, *J. Inorg. Biochem.* **2018**, *179*, 40–53.

[7] I. Bertini, A. Rosato, *Cell Mol. Life Sci.* **2008**, *65*, 89–91.

[8] D. P. Giedroc, A. I. Arunkumar, *Dalton Trans.* **2007**, 3107–20.

[9] Z. Ma, F. E. Jacobsen, D. P. Giedroc, *Chem. Rev.* **2009**, *109*, 4644–81.

[10] A. J. Guerra, D. P. Giedroc, *Arch. Biochem. Biophys.* **2012**, *519*, 210–22.

[11] S. Silver, L. T. Phung, *Annu. Rev. Microbiol.* **1996**, *50*, 753–789.

[12] M. R. Bruins, S. Kapil, F. W. Oehme, *Ecotoxicol. Environ. Saf.* **2000**, *45*, 198–207.

[13] D. H. Nies, *FEMS Microbiol. Rev.* **2003**, *27*, 313–339.

[14] D. Osman, J. S. Cavet, *Nat. Prod. Rep.* **2010**, *27*, 668–80.

[15] K. Poole, *J. Antimicrob. Chemother.* **2012**, *67*, 2069–89.

[16] J. J. R. F. da Silva, R. J. P. Williams, *The Biological Chemistry of the Elements: The Inorganic Chemistry of Life*, Oxford University Press, **2001**, pp. 229–277.

[17] C. Andreini, A. Rosato, L. Banci, *PLoS One* **2017**, *12*, e0171279.

[18] C. Andreini, L. Banci, I. Bertini, A. Rosato, *J. Proteome Res.* **2006**, *5*, 196–201.

[19] T. Kambe, Y. Yamaguchi-Iwai, R. Sasaki, M. Nagao, *Cell Mol. Life Sci.* **2004**, *61*, 49–68.

[20] E. Z. Eisenmesser, O. Millet, W. Labeikovsky, D. M. Korzhnev, M. Wolf-Watz, D. A. Bosco, J. J. Skalicky, L. E. Kay, D, Kern, *Nature* **2005**, *438*, 117–121.

[21] R. Otten, L. Liu, L. R. Kenner, M. W. Clarkson, D. Mavor, D. S. Tawfik, D. Kern, J. S. Fraser, *Nat. Commun.* **2018**, *9*, 1314.

[22] M. Orozco, *Chem. Soc. Rev.* **2014**, *43*, 5051–5066.

[23] M. De Vivo, M. Masetti, G. Bottegoni, A. Cavalli, *J. Med. Chem.* **2016**, *59*, 4035–4061.

[24] I. M. Armitage, T. Drakenberg, B. Reilly, *Met. Ions Life Sci.* **2013**, *11*, 117–44.

[25] G. J. Kornhaber, D. Snyder, H. N. B. Moseley, G. T. Montelione, *J. Biomol. NMR Spectroscopy.* **2006**, *34*, 259–269.

[26] I. Anastassopoulou, L. Banci, I. Bertini, F. Cantini, E. Katsari, A. Rosato, *Biochemistry* **2004**, *43*, 13046–13053.

[27] L. Banci, I. Bertini, F. Cantini, N. Della-Malva, M. Migliardi, A. Rosato, *J. Biol. Chem.* **2007**, *282*, 23140–23146.

[28] L. Banci, I. Bertini, F. Cantini, C. Massagni, M. Migliardi, A. Rosato, *J. Biol. Chem.* **2009**, *284*, 9354–60.

[29] L. Banci, I. Bertini, V. Calderone, N. Della-Malva, I. C. Felli, S. Neri, A. Pavelkova, A. Rosato, *Biochem. J.* **2009**, *422*, 37–42.

[30] L. Banci, I. Bertini, S. Mangani, *J. Synchrotron Radiat.* **2005**, *12*, 94–97.

[31] D. K. Chakravorty, B. Wang, C. W. Lee, A. J. Guerra, D. P. Giedroc, K. M. Merz, *J. Biomol. NMR.* **2013**, *56*, 125–137.

[32] I. Bertini, C. Luchinat, A. Rosato, *Prog. Biophys. Mol. Biol.* **1996**, *66*, 43–80.

[33] M. Allegrozzi, I. Bertini, M. B. L. Janik, Y.-M. Lee, G. Liu, C. Luchinat, *J. Am. Chem. Soc.* **2000**, *122*, 4154–4161.

[34] G. Otting, *Annu. Rev. Biophys.* **2010**, *39*, 387–405.

[35] F. Arnesano, L. Banci, M. Piccioli, *Q. Rev. Biophys.* **2005**, *38*, 167.

[36] I. Bertini, C. Luchinat, G. Parigi, R. Pierattelli, *ChemBioChem* **2005**, *6*, 1536–1549.

[37] I. Solomon, *Phys. Rev.* **1955**, *99*, 559–565.

[38] I. Solomon, N. Bloembergen, *J. Chem. Phys.* **1956**, *25*, 261–266.

[39] Y. S. Cho-Chung, H. C. Pitot, *Eur. J. Biochem.* **1968**, *3*, 401–406.

[40] I. Bertini, A. Donaire, I. C. Felli, C. Luchinat, A. Rosato, *Inorg. Chem.* **1997**, *36*, 4798–4803.

[41] C. Schmitz, M. J. Stanton-Cook, X.-C. Su, G. Otting, T. Huber, *J. Biomol. NMR Spectroscopy.* **2008**, *41*, 179–189.

[42] M. Rinaldelli, A. Carlon, E. Ravera, G. Parigi, C. Luchinat, *J. Biomol. NMR Spectroscopy.* **2015**, *61*, 21–34.

[43] M. Gochin, H. Roder, *Protein Sci.* **2008**, *4*, 296–305.

[44] L. Banci, I. Bertini, K. L. Bren, M. A. Cremonini, H. B. Gray, C. Luchinat, P. Turano, *JBIC J. Biol. Inorg. Chem.* **1996**, *1*, 117–126.

[45] I. Bertini, M. B. L. Janik, Y.-M. Lee, C. Luchinat, A. Rosato, *J. Am. Chem. Soc.* **2001**, *123*, 4181–4188.

[46] I. Bertini, I. Gelis, N. Katsaros, C. Luchinat, A. Provenzani, *Biochemistry* **2003**, *42*, 8011–8021.

[47] K. Barthelmes, A. M. Reynolds, E. Peisach, H. R. A. Jonker, N. J. DeNunzio, K. N. Allen, B. Imperiali, H. Schwalbe, *J. Am. Chem. Soc.* **2011**, *133*, 808–819.

[48] M. A. Hass, M. Ubbink, *Curr. Opin. Struct. Biol.* **2014**, *24*, 45–53.

[49] F. Rodriguez-Castañeda, P. Haberz, A. Leonov, C. Griesinger, *Magn. Reson. Chem.* **2006**, *44*, S10–S16.

[50] X.-C. Su, G. Otting, *J. Biomol. NMR* **2010**, *46*, 101–112.

[51] S. Aono, D. Bentrop, I. Bertini, A. Donaire, C. Luchinat, Y. Niikura, A. Rosato, *Biochemistry* **1998**, *37*, 9812–9826.

[52] Y. Shen, F. Delaglio, G. Cornilescu, A. Bax, *J. Biomol. NMR* **2009**, *44*, 213–223.

[53] N. Tjandra, J. G. Omichinski, A. M. Gronenborn, G. M. Clore, A. Bax, *Nat. Struct. Biol.* **1997**, *4*, 732–738.

[54] B. Xia, V. Tsui, D. A. Case, H. J. Dyson, P. E. Wright, *J. Biomol. NMR Spectroscopy.* **2002**, *22*, 317–31.

[55] J. P. Linge, M. A. Williams, C. A. E. M. Spronk, A. M. J. J. Bonvin, M. Nilges, *Proteins Struct., Funct., Bioinf.* **2003**, *50*, 496–506.

[56] J. Chen, W. Im, C. L. Brooks, *J. Am. Chem. Soc.* **2004**, *126*, 16038–16047.

[57] M. Feig, J. Karanicolas, C. L. Brooks, *J. Mol. Graphics* **2004**, *22*, 377–395.

[58] C. C. Jao, B. G. Hegde, J. Chen, I. S. Haworth, R. Langen, *Proc. Natl. Acad. Sci. USA* **2008**, *105*, 19666–19671.

[59] J. Mao, N.-N. Do, F. Scholz, L. Reggie, M. Mehler, A. Lakatos, Y.-S. Ong, S. J. Ullrich, L. J. Brown, R. C. D. Brown, J. Becker-Baldus, J. Wachtveitl, C. Glaubitz, *J. Am. Chem. Soc.* **2014**, *136*, 17578–17590.

[60] J. H. Ryu, H. S. Wee, W. B. Lee, *Phys. Rev. E.* **2016**, *94*, 032501.

[61] S. B. Nabuurs, A. J. Nederveen, W. Vranken, J. F. Doreleijers, A. M. J. J. Bonvin, G. W. Vuister, G. Vriend, C. A. E. M. Spronk, *Proteins Struct., Funct., Bioinf.* **2004**, *55*, 483–486.

[62] A. J. Nederveen, J. F. Doreleijers, W. Vranken, Z. Miller, C. A. E. M. Spronk, S. B. Nabuurs, P. Güntert, M. Livny, J. L. Markley, M. Nilges, E. L. Ulrich, R. Kaptein, A. M. J. J. Bonvin, *Proteins Struct., Funct., Genet.* **2005**, *59*, 662–672.

[63] S. B. Nabuurs, C. A. E. M. Spronk, G. W. Vuister, G. Vriend, *PLoS Comput. Biol.* **2006**, *2*, e9.

[64] E. Saccenti, A. Rosato, *J. Biomol. NMR* **2008**, *40*, 251–261.

[65] D. A. Case, T. E. Cheatham, T. Darden, H. Gohlke, R. Luo, K. M. Merz, A. Onufriev, C. Simmerling, B. Wang, R. J. Woods, *J. Comput. Chem.* **2005**, *26*, 1668–1688.

[66] D. der Spoel, E. Lindahl, B. Hess, G. Groenhof, A. E. Mark, H. J. C. Berendsen, *J. Comput. Chem.* **2005**, *26*, 1701–1718.

[67] I. Bertini, D. A. Case, L. Ferella, A. Giachetti, A. Rosato, *Bioinformatics* **2011**, *27*, 2384–2390.

[68] T. A. Wassenaar, M. van Dijk, N. Loureiro-Ferreira, G. van der Schot, S. J. de Vries, C. Schmitz, J. van der Zwan, R. Boelens, A. Giachetti, L. Ferella, A. Rosato, I. Bertini, T. Herrmann, H. R. A. Jonker, A. Bagaria, V. Jaravine, P. Guntert, H. Schwalbe, W. F. Vranken, J. F. Doreleijers, G. Vriend, G. W. Vuister, D. Franke, A. Kikhney, D. I. Svergun, R. H. Fogh, J. Ionides, E. D. Laue, C. Spronk, S. Jurksa, M. Verlato, S. Badoer, S. Dal Pra, M. Mazzucato, E. Frizziero, A. M. J. J. Bonvin, *J. GRID Comput.* **2012**, *10*, 743–767.

[69] D. E. Tanner, J. C. Phillips, K. Schulten, *J. Chem. Theory Comput.* **2012**, *8*, 2521–2530.

[70] P. Andreetto, J. Astalos, M. Dobrucky, A. Giachetti, D. Rebatto, A. Rosato, V. Tran, M. Verlato, L. Zangrando, *EGI federated platforms supporting accelerated computing*, in: Sissa Medialab Srl, **2017**/jnl>.

[71] L. Banci, I. Bertini, G. Cavallaro, A. Giachetti, C. Luchinat, G. Parigi, *J. Biomol. NMR* **2004**, *28*, 249–261.

[72] L. Banci, I. Bertini, G. G. Savellini, A. Romagnoli, P. Turano, M. A. Cremonini, C. Luchinat, H. B. Gray, *Proteins* **1997**, *29*, 68–76.

[73] D. Sala, A. Giachetti, C. Luchinat, A. Rosato, *J. Biomol. NMR* **2016**, *66*, 175–185.

[74] J. M. Bradley, N. E. Le Brun, G. R. Moore, *J. Biol. Inorg. Chem.* **2016**, *21*, 13–28.

[75] L. Li, C. J. Fang, J. C. Ryan, E. C. Niemi, J. A. Lebrón, P. J. Björkman, H. Arase, F. M. Torti, S. V. Torti, M. C. Nakamura, W. E. Seaman, *Proc. Natl. Acad. Sci. USA* **2010**, *107*, 3505–10.

[76] P. M. Harrison, P. Arosio, *Biochim. Biophys. Acta - Bioenerg.* **1996**, *1275*, 161–203.

[77] P. Santambrogio, S. Levi, A. Cozzi, B. Corsi, P. Arosio, *Biochem. J.* **1996**, *314 (Pt 1)*, 139–44.

[78] E. C. Theil, *Inorg. Chem.* **2013**, *52*, 12223–12233.

[79] X. L, and, E. C. Theil*, **2005**.

[80] X. Yang, P. Arosio, N. D. Chasteen, *Biophys. J.* **2000**, *78*, 2049–59.

[81] N. D. Chasteen, *Met. Ions Biol. Syst.* **1998**, *35*, 479–514.

[82] M. C. Linder, *Nutrients* **2013**, *5*, 4022–4050.

[83] Y. Zhang, M. Mikhael, D. Xu, Y. Li, S. Soe-Lin, B. Ning, W. Li, G. Nie, Y. Zhao, P. Ponka, *Antioxid. Redox Signaling* **2010**, *13*, 999–1009.

[84] T. Z. Kidane, *AJP Cell Physiol.* **2006**, *291*, C445–C455.

[85] M. Rudeck, T. Volk, N. Sitte, T. Grune, *IUBMB Life* **2000**, *49*, 451–456.

[86] D. Finazzi, P. Arosio, *Arch. Toxicol.* **2014**, *88*, 1787–1802.

[87] M. Kim, Y. Rho, K. S. Jin, B. Ahn, S. Jung, H. Kim, M. Ree, *Biomacromolecules* **2011**, *12*, 1629–1640.

[88] T. Kurz, B. Gustafsson, U. T. Brunk, *Free Radical Biol. Med.* **2011**, *50*, 1647–1658.

[89] G. Melman, F. Bou-Abdallah, E. Vane, P. Maura, P. Arosio, A. Melman, *Biochim. Biophys. Acta, Biochim. Biophys. Acta Gen. Subj.* **2013**, *1830*, 4669–4674.

[90] M. J. Hynes, M. Ó. Coinceanainn, *J. Inorg. Biochem.* **2002**, *90*, 18–21.

[91] R. Anandakrishnan, B. Aguilar, A. V. Onufriev, *Nucleic Acids Res.* **2012**, *40*, W537–W541.

[92] D. Sala, S. Ciambellotti, A. Giachetti, P. Turano, A. Rosato, *J. Chem. Inf. Model.* **2017**, *57*, 2112–2118.

[93] B. Roux, *Comput. Phys. Commun.* **1995**, *91*, 275–282.

[94] O. Kolaj-Robin, D. Russell, K. A. Hayes, J. T. Pembroke, T. Soulimane, *FEBS Lett.* **2015**, *589*, 1283–1295.

[95] M. Lu, J. Chai, D. Fu, *Nat. Struct. Mol. Biol.* **2009**, *16*, 1063–1067.

[96] Y. Wei, D. Fu, *J. Biol. Chem.* **2005**, *280*, 33716–33724.

[97] N. Coudray, S. Valvo, M. Hu, R. Lasala, C. Kim, M. Vink, M. Zhou, D. Provasi, M. Filizola, J. Tao, J. Fang, P.A. Penczek, I. Ubarretxena-Belandia, D.L. Stokes, *Proc. Natl. Acad. Sci. USA* **2013**, *110*, 2140–2145.

[98] S. Gupta, J. Chai, J. Cheng, R. D'Mello, M. R. Chance, D. Fu, *Nature* **2014**, *512*, 101–104.

[99] B. Gold, G. M. Rodriguez, S. A. E. Marras, M. Pentecost, I. Smith, *Mol. Microbiol.* **2008**, *42*, 851–865.

[100] I. Brune, H. Werner, A. T. Hüser, J. Kalinowski, A. Pühler, A. Tauch, *BMC Genomics* **2006**, *7*, 21.

[101] N. Bhattacharya, M. Yi, H.-X. Zhou, T. M. Logan, *J. Mol. Biol.* **2007**, *374*, 977–92.

[102] G. P. Wylie, V. Rangachari, E. A. Bienkiewicz, V. Marin, N. Bhattacharya, J. F. Love, J. R. Murphy, T. M. Logan, *Biochemistry* **2005**, *44*, 40–51.

[103] S. Ghosh, N. Chandra, S. Vishveshwara, *PLoS Comput. Biol.* **2015**, *11*, 1–24.

[104] G. M. Rodriguez, M. I. Voskuil, B. Gold, G. K. Schoolnik, I. Smith, *Infect. Immun.* **2002**, *70*, 3371–81.

[105] G. Wisedchaisri, R. K. Holmes, W. G. J. Hol, *J. Mol. Biol.* **2004**, *342*, 1155–69.

[106] P. Li, K. M. Merz, *J. Chem. Inf. Model.* **2016**, *56*, 599–604.

[107] B. Zambelli, F. Musiani, S. Benini, S. Ciurli, *Acc. Chem. Res.* **2011**, *44*, 520–30.

[108] F. Musiani, B. Zambelli, M. Bazzani, L. Mazzei, S. Ciurli, *Metallomics* **2015**, *7*, 1305–1318.

[109] K. De Pina, V. Desjardin, M. A. Mandrand-Berthelot, G. Giordano, L. F. Wu, *J. Bacteriol.* **1999**, *181*, 670–674.

[110] P. T. Chivers, R. T. Sauer, *J. Biol. Chem.* **2000**, *275*, 19735–41.

[111] A. H. M. van Vliet, F. D. Ernst, J. G. Kusters, *Trends Microbiol.* **2004**, *12*, 489–94.

[112] P. T. Chivers, R. T. Sauer, *Protein Sci.* **1999**, *8*, 2494–500.

[113] E. R. Schreiter, S. C. Wang, D. B. Zamble, C. L. Drennan, *Proc. Natl. Acad. Sci. USA* **2006**, *103*, 13676–81.

[114] L. Mazzei, O. Dobrovolska, F. Musiani, B. Zambelli, S. Ciurli, *J. Biol. Inorg. Chem.* **2015**, *20*, 1021–37.

[115] M. J. Bradley, P. T. Chivers, N. A. Baker, *J. Mol. Biol.* **2008**, *378*, 1155–73.

[116] D. J. Sindhikara, A. E. Roitberg, K. M. Merz, *Biochemistry* **2009**, *48*, 12024–33.

[117] F. Musiani, B. Bertoša, A. Magistrato, B. Zambelli, P. Turano, V. Losasso, C. Micheletti, S. Ciurli, P. Carloni, *J. Chem. Theory Comput.* **2010**, *6*, 3503–3515.

[118] H.-B. B. Guo, A. Johs, J. M. Parks, L. Olliff, S. M. Miller, A. O. Summers, L. Liang, J. C. Smith, *J. Mol. Biol.* **2010**, *398*, 555–568.

[119] A. C. Vaiana, Z. Cournia, I. B. Costescu, J. C. Smith, *Comput. Phys. Commun.* **2005**, *167*, 34–42.

[120] A. P. Morby, J. S. Turner, J. W. Huckle, N. J. Robinson, *Nucleic Acids Res.* **1993**, *21*, 921–925.

[121] C. W. Lee, D. K. Chakravorty, F.-M. J. Chang, H. Reyes-Caballero, Y. Ye, K. M. Merz, D. P. Giedroc, *Biochemistry* **2012**, *51*, 2619–29.

[122] D. K. Chakravorty, B. Wang, C. W. Lee, D. P. Giedroc, K. M. Merz, *J. Am. Chem. Soc.* **2012**, *134*, 3367–3376.

[123] M. B. Peters, Y. Yang, B. Wang, L. Füsti-Molnár, M. N. Weaver, K. M. Merz, *J. Chem. Theory Comput.* **2010**, *6*, 2935–2947.

[124] F. Agriesti, D. Roncarati, F. Musiani, C. Del Campo, M. Iurlaro, F. Sparla, S. Ciurli, A. Danielli, V. Scarlato, *Nucleic Acids Res.* **2014**, *42*, 3138–51.

[125] C. Cissé, S. V. Mathieu, M. B. O. O. Abeih, L. Flanagan, S. Vitale, P. Catty, D. Boturyn, I. Michaud-Soret, S. Crouzy, *ACS Chem. Biol.* **2014**, *9*, 2779–2786.

[126] C. E. Carr, F. Musiani, H.-T. Huang, P. T. Chivers, S. Ciurli, M. J. Maroney, *Inorg. Chem.* **2017**, *56*, 6459–6476.

[127] P. Li, K. M. Merz, *Chem. Rev.* **2017**, *117*, 1564–1686.

[128] J. O. Hirschfelder, R. B. Ewell, J. R. Roebuck, *J. Chem. Phys.* **1938**, *6*, 205–218.

[129] P. Li, B. P. Roberts, D. K. Chakravorty, K. M. Merz, *J. Chem. Theory Comput.* **2013**, *9*, 2733.

[130] P. Li, L. F. Song, K. M. Merz, *J. Phys. Chem. B* **2015**, *119*, 883.

[131] N. Metropolis, S. Ulam, *J. Am. Stat. Assoc.* **1949**, *44*, 335.

[132] B. J. Alder, T. E. Wainwright, *J. Chem. Phys.* **1957**, *27*, 1208.

[133] R. Wu, Z. Lu, Z. Cao, Y. Zhang, *J. Chem. Theory Comput.* **2011**, *7*, 433–443.

[134] W. Gong, R. Wu, Y. Zhang, *J. Comput. Chem.* **2015**, *36*, 2228–2235.

[135] T. Zhu, X. Xiao, C. Ji, J. Z. H. Zhang, *J. Chem. Theory Comput.* **2013**, *9*, 1788.

[136] R. W. Zwanzig, *J. Chem. Phys.* **1954**, *22*, 1420–1426.

[137] T. P. Straatsma, H. J. C. Berendsen, *J. Chem. Phys.* **1988**, *89*, 5876.

[138] F. Fogolari, A. Brigo, H. Molinari, *Biophys. J.* **2003**, *85*, 159.

[139] C. J. Cramer, D. G. Truhlar, *Chem. Rev.* **1999**, *99*, 2161.

[140] P. Li, L. F. Song, K. M. Merz, *J. Chem. Theory Comput.* **2015**, *11*, 1645.

[141] P. Li, K. M. Merz, *J. Chem. Theory Comput.* **2014**, *10*, 289–297.

[142] I. Leontyev, A. Stuchebrukhov, *J. Chem. Phys.* **2009**, *130*, 85102.

[143] I. Leontyev, A. Stuchebrukhov, *J. Chem. Theory Comput.* **2010**, *6*, 3153.

[144] I. Leontyev, A. Stuchebrukhov, *J. Chem. Theory Comput.* **2010**, *6*, 1498.

[145] I. V. Leontyev, A. A. Stuchebrukhov, *J. Chem. Theory Comput.* **2012**, *8*, 3207.

[146] R. D. Hancock, A. E. Martell, *Chem. Rev.* **1989**, *89*, 1875.

[147] F. Lin, R. Wang, *J. Chem. Theory Comput.* **2010**, *6*, 1852.

[148] MS Minkara, M. N. Ucisik, M. N. Weaver, K. M. Merz, *J. Chem. Theory Comput.* **2014**, *10*, 1852.

[149] A. D. MacKerell, M. Karplus, *J. Phys. Chem.* **1991**, *95*, 10559–10560.

[150] P. S. Nerenberg, T. Head-Gordon, *Curr. Opin. Struct. Biol.* **2018**, *49*, 129–138.

[151] L. Duan, T. Zhu, C. Ji, Q. Zhang, J. Z. H. Zhang, *Phys. Chem. Chem. Phys.* **2017**, *19*, 15273–15284.

[152] K. Shahrokh, A. Orendt, G. S. Yost, T. E. Cheatham, *J. Comput. Chem.* **2012**, *33*, 119.

[153] J. A. Pople, R. Krishnan, H. Schlegel, J. S. Binkley, *Int. J. Quantum Chem.* **1979**, *16*, 225.

[154] C. J. Cramer, D. G. Truhlar, *Phys. Chem. Chem. Phys.* **2009**, *11*, 10757.

[155] J. M. Seminario, *Int. J. Quantum Chem.* **1996**, *60*, 1271.

[156] S. Zheng, Q. Tang, J. He, S. Du, S. Xu, C. Wang, Y. Xu, F. Lin, *J. Chem. Inf. Model.* **2016**, *56*, 811.

[157] L. Hu, U. Ryde, *J. Chem. Theory Comput.* **2011**, *7*, 2452.

[158] R. M. Betz, R. C. Walker, *J. Comput. Chem.* **2015**, *36*, 79.

[159] L.-P. Wang, T. J. Martinez, V. S. Pande, *J. Phys. Chem. Lett.* **2014**, *5*, 1885.

[160] H. Heinz, U. W. Suter, *J. Phys. Chem. B* **2004**, *108*, 18341.

[161] C. Jelsch, V. Pichon-Pesme, C. Lecomte, A. Aubry, *Acta Crystallogr., Sect. D Biol. Crystallogr.* **1998**, *54*, 1306.

[162] L. E. Chirlian, M. M. Francl, *J. Comput. Chem.* **1987**, *8*, 894.

[163] C. M. Breneman, K. B. Wiberg, *J. Comput. Chem.* **1990**, *11*, 361.

[164] R. Mulliken, *J. Chem. Phys.* **1955**, *23*, 1833.

[165] R. Mulliken, *J. Chem. Phys.* **1955**, *23*, 1841.

[166] R. Mulliken, *J. Chem. Phys.* **1955**, *23*, 2338.

[167] R. Mulliken, *J. Chem. Phys.* **1955**, *23*, 2343.

[168] A. E. Reed, R. B. Weinstock, F. Weinhold, *J. Chem. Phys.* **1985**, *83*, 735.

[169] T. Lu, F. Chen, *J. Theor. Comput. Chem.* **2012**, *11*, 163.

[170] A. Jakalian, D. B. Jack, C. I. Bayly, B. L. Bush, D. B. Jack, C. I. Bayly, *J. Comput. Chem.* **2002**, *23*, 1623.

[171] C. P. Kelly, C. J. Cramer, D. G. Truhlar, *J. Chem. Theory Comput.* **2005**, *1*, 1133.

[172] A. V. Marenich, S. V. Jerome, C. J. Cramer, D. G. Truhlar, *J. Chem. Theory Comput.* **2012**, *8*, 527.

[173] U. C. Singh, P. A. Kollman, *J. Comput. Chem.* **1984**, *5*, 129.

[174] P. Kollman, J. McKelvey, A. Johansson, S. Rothenberg, *J. Am. Chem. Soc.* **1975**, *97*, 955.

[175] S. Cox, D. Williams, *J. Comput. Chem.* **1981**, *2*, 304.

[176] C. I. Bayly, P. Cieplak, W. Cornell, P. A. Kollman, *J. Phys. Chem.* **1993**, *97*, 10269.

[177] U. Ryde, *Eur. Biophys. J.* **1996**, *24*, 213.

[178] Y. Yang, M. N. Weaver, K. M. Merz, *J. Phys. Chem. A* **2009**, *113*, 9843.

[179] J. Åqvist, A. Warshel, *J. Am. Chem. Soc.* **1990**, *112*, 2860.

[180] P. Oelschlaeger, M. Klahn, W. A. Beard, S. H. Wilson, A. Warshel, *J. Mol. Biol.* **2007**, *366*, 687.

[181] M. Masetti, F. Musiani, M. Bernetti, F. Falchi, A. Cavalli, S. Ciurli, M. Recanatini, *J. Comput. Chem.* **2017**, *38*, 1834–1843.

[182] A. Saxena, A. E. García, *J. Phys. Chem. B* **2015**, *119*, 219–227.

[183] F. Duarte, P. Bauer, A. Barrozo, B. A. Amrein, M. Purg, J. Åqvist, S. C. L. Kamerlin, *J. Phys. Chem. B* **2014**, *118*, 4351.

[184] S. Y. Lu, Z. M. Huang, W. K. Huang, X. Y. Liu, Y. Y. Chen, T. Shi, J. Zhang, *Proteins Struct., Funct., Genet.* **2013**, *81*, 740.

[185] Y. Jiang, H. Zhang, T. Tan, *J. Chem. Theory Comput.* **2016**, *12*, 3250.

[186] J. M. Martínez, J. Hernández-Cobos, H. Saint-Martin, R. R. Pappalardo, I. Ortega-Blake, E. S. Marcos, *J. Chem. Phys.* **2000**, *112*, 2339.

[187] J. M. Martínez, R. R. Pappalardo, E. Sánchez Marcos, *J. Am. Chem. Soc.* **1999**, *121*, 3175.

[188] R. R. Pappalardo, E. Sanchez Marcos, *J. Phys. Chem.* **1993**, *97*, 4500.

[189] M. Bernal-Uruchurtu, J. Hernández-Cobos, I. Ortega-Blake, *J. Chem. Phys.* **1998**, *108*, 1750.

[190] J. Chaboy, A. Muñoz-Páez, P. J. Merkling, E. S. Marcos, *J. Chem. Phys.* **2006**, *124*, 64509.

[191] G. Lamoureux, E. A. Orabi, *Mol. Simul.* **2012**, *38*, 704.

[192] A. Warshel, M. Levitt, *J. Mol. Biol.* **1976**, *103*, 227.

[193] I. M. Svishchev, P. G. Kusalik, J. Wang, R. J. Boyd, *J. Chem. Phys.* **1996**, *105*, 4742.

[194] P. J. van Maaren, D. van der Spoel, *J. Phys. Chem. B* **2001**, *105*, 2618.

[195] M. Medeiros, M. E. Costas, *J. Chem. Phys.* **1997**, *107*, 2012.

[196] T. Campbell, R. K. Kalia, A. Nakano, P. Vashishta, S. Ogata, S. Rodgers, *Phys. Rev. Lett.* **1999**, *82*, 4866.

[197] M. E. Tuckerman, G. J. Martyna, *J. Phys. Chem. B* **2001**, *104*, 159.

[198] M. Souaille, H. Loirat, D. Borgis, M. Gaigeot, *Comput. Phys. Commun.* **2009**, *180*, 276.

[199] Y. Shi, P. Ren, M. Schnieders, J.-P. Piquemal, *Polarizable Force Fields for Biomolecular Modeling*, in: Wiley-Blackwell, **2015**, pp. 51–86.

[200] T. Tuccinardi, A. Martinelli, E. Nuti, P. Carelli, F. Balzano, G. Uccello-Barretta, G. Murphy, A. Rossello, *Bioorg. Med. Chem.* **2006**, *14*, 4260.

[201] S. C. Hoops, K. W. Anderson, K. M. Merz, *J. Am. Chem. Soc.* **1991**, *113*, 8262.

[202] K. M. Merz, *J. Am. Chem. Soc.* **1991**, *113*, 406.

[203] K. M. Merz, M. A. Murcko, P. A. Kollman, *J. Am. Chem. Soc.* **1991**, *113*, 4484.

[204] J. Åqvist, A. Warshel, *J. Mol. Biol.* **1992**, *224*, 7.

[205] Y.-P. Pang, *J. Mol. Model.* **1999**, *5*, 196.

[206] Y.-P. Pang, J. Davis, S. Wang, J. G. Park, M. P. Nambiar, J. J. Schmidt, C. B. Millard, *PLoS One* **2010**, *5*, e10129.

Received: June 4, 2018

BIBLIOGRAPHY

1. Nielsen, F. H. Evolutionary events culminating in specific minerals becoming essential for life. *Eur. J. Nutr.* **39**, 62–66 (2000).
2. Williams, J. J. R. Frausto da Silva, R. J. P. *The Biological Chemistry of the Elements: The Inorganic Chemistry of Life*. *Proteins* (Oxford University Press, 2001).
3. Potassium and Sodium (In Biological Systems). in *Van Nostrand's Scientific Encyclopedia* (John Wiley & Sons, Inc., 2009). doi:10.1002/0471743984.vse5720
4. Romani, A. M. P. Magnesium Homeostasis in Mammalian Cells. in 69–118 (2013). doi:10.1007/978-94-007-5561-1_4
5. Erulkar, S. D. The Versatile Role of Calcium in Biological Systems. *Interdiscip. Sci. Rev.* **6**, 322–332 (2014).
6. Crichton, R. The Importance of Iron for Biological Systems. in *Iron Metabolism* 17–58 (John Wiley & Sons, Ltd, 2009). doi:10.1002/9780470010303.ch2
7. Maret, W. Zinc and Zinc Ions in Biological Systems. in *Encyclopedia of Metalloproteins* 2396–2400 (Springer New York, 2013). doi:10.1007/978-1-4614-1533-6_185
8. Finney, L. A. & O'Halloran, T. V. Transition metal speciation in the cell: Insights from the chemistry of metal ion receptors. *Science (80-.).* **300**, 931–936 (2003).
9. Bertini, I. & Rosato, A. Menkes disease. *Cell. Mol. Life Sci.* **65**, 89–91 (2008).
10. Boros, E. & Packard, A. B. Radioactive Transition Metals for Imaging and Therapy. *Chem. Rev.* **119**, 870–901 (2019).
11. Andreini, C., Bertini, I., Cavallaro, G., Holliday, G. L. & Thornton, J. M. Metal ions in biological catalysis: From enzyme databases to general principles. *J. Biol. Inorg. Chem.* **13**, 1205–1218 (2008).
12. Singh Sekhon, B. Metallochaperones - an Overview. *Curr. Chem. Biol.* **4**, 173–186 (2012).
13. Valasatava, Y., Rosato, A., Furnham, N., Thornton, J. M. & Andreini, C. To what extent do structural changes in catalytic metal sites affect enzyme function? *J. Inorg. Biochem.* **179**, 40–53 (2018).
14. Guerra, A. J. & Giedroc, D. P. Metal site occupancy and allosteric switching in bacterial metal sensor proteins. *Arch. Biochem. Biophys.* **519**, 210–222 (2012).
15. Thirumoorthy, N., Manisenthil Kumar, K. T., Sundar, A. S., Panayappan, L. & Chatterjee, M. Metallothionein: An overview. *World Journal of Gastroenterology* **13**, 993–996 (2007).
16. Koppenol, W. H. The centennial of the Fenton reaction. *Free Radic. Biol. Med.* **15**, 645–51 (1993).
17. Baltaci, A. K. & Yuce, K. Zinc Transporter Proteins. *Neurochem. Res.* **0**, 0 (2017).
18. Andreini, C., Banci, L., Bertini, I. & Rosato, A. Counting the zinc-proteins encoded in the human genome. *J. Proteome Res.* **5**, 196–201 (2006).
19. Beyersmann, D. & Haase, H. Functions of zinc in signaling, proliferation and differentiation of mammalian cells. *BioMetals* **14**, 331–341 (2001).
20. Eisenmesser, E. Z. *et al.* Intrinsic dynamics of an enzyme underlies catalysis. *Nature* **438**, 117–121 (2005).
21. Orozco, M. A theoretical view of protein dynamics. *Chem. Soc. Rev.* **43**, 5051–5066 (2014).
22. Li, P. & Merz, K. M. Metal Ion Modeling Using Classical Mechanics. *Chem. Rev.* **117**, 1564–1686 (2017).
23. Hirschfelder, J. O., Ewell, R. B. & Roebuck, J. R. Determination of intermolecular forces from the Joule-Thomson coefficients. *J. Chem. Phys.* **6**, 205–218 (1938).
24. Li, P., Roberts, B. P., Chakravorty, D. K. & Merz, K. M. Rational Design of Particle Mesh Ewald Compatible Lennard-Jones Parameters for + 2 Metal Cations in Explicit

Solvent. *J. Chem. Theory Comput.* **9**, 2733 (2013).

25. Zwanzig, R. W. High†ï¿½Temperature Equation of State by a Perturbation Method. I. Nonpolar Gases. *J. Chem. Phys.* **22**, 1420–1426 (1954).
26. Straatsma, T. P. & Berendsen, H. J. C. Free Energy of Ionic Hydration: Analysis of a Thermodynamic Integration Technique to Evaluate Free Energy Differences by Molecular Dynamics Simulations. *J. Chem. Phys.* **89**, 5876 (1988).
27. Fogolari, F., Brigo, A. & Molinari, H. Protocol for Mm/Pbsa Molecular Dynamics Simulations of Proteins. *Biophys. J.* **85**, 159 (2003).
28. Cramer, C. J. & Truhlar, D. G. Implicit Solvation Models: Equilibria, Structure, Spectra, and Dynamics. *Chem. Rev.* **99**, 2161 (1999).
29. Li, P., Song, L. F. & Merz, K. M. Parameterization of Highly Charged Metal Ions Using the 12–6-4 Lj-Type Nonbonded Model in Explicit Water. *J. Phys. Chem. B* **119**, 883 (2015).
30. Li, P., Song, L. F. & Merz, K. M. Systematic Parameterization of Monovalent Ions Employing the Nonbonded Model. *J. Chem. Theory Comput.* **11**, 1645 (2015).
31. Li, P. & Merz, K. M. Taking into account the ion-induced dipole interaction in the nonbonded model of ions. *J. Chem. Theory Comput.* **10**, 289–297 (2014).
32. Leontyev, I. & Stuchebrukhov, A. Electronic Continuum Model for Molecular Dynamics Simulations. *J. Chem. Phys.* **130**, 85102 (2009).
33. Leontyev, I. V & Stuchebrukhov, A. A. Polarizable Mean-Field Model of Water for Biological Simulations with Amber and Charmm Force Fields. *J. Chem. Theory Comput.* **8**, 3207 (2012).
34. Hancock, R. D. & Martell, A. E. Ligand Design for Selective Complexation of Metal Ions in Aqueous Solution. *Chem. Rev.* **89**, 1875 (1989).
35. Minkara, M. S., Ucisik, M. N., Weaver, M. N. & Merz, K. M. Molecular Dynamics Study of Helicobacter Pylori Urease. *J. Chem. Theory Comput.* **10**, 1852 (2014).
36. Nerenberg, P. S. & Head-Gordon, T. New developments in force fields for biomolecular simulations. *Curr. Opin. Struct. Biol.* **49**, 129–138 (2018).
37. Duan, L., Zhu, T., Ji, C., Zhang, Q. & Zhang, J. Z. H. Direct folding simulation of helical proteins using an effective polarizable bond force field. *Phys. Chem. Chem. Phys.* **19**, 15273–15284 (2017).
38. Shahrokh, K., Orendt, A., Yost, G. S. & Cheatham, T. E. Quantum Mechanically Derived Amber-Compatible Heme Parameters for Various States of the Cytochrome P450 Catalytic Cycle. *J. Comput. Chem.* **33**, 119 (2012).
39. Cramer, C. J. *Essentials of Computational Chemistry Theories and Models*. *Essentials of Computational Chemistry* **42**, (2004).
40. Seminario, J. M. Calculation of Intramolecular Force Fields from Second-Derivative Tensors. *Int. J. Quantum Chem.* **60**, 1271 (1996).
41. Merz, K. M. NIH Public Access. **56**, 125–137 (2014).
42. Li, P. & Merz, K. M. MCPB.py: A Python Based Metal Center Parameter Builder. *J. Chem. Inf. Model.* **56**, 599–604 (2016).
43. Zheng, S. *et al.* Vffdt: A New Software for Preparing Amber Force Field Parameters for Metal-Containing Molecular Systems. *J. Chem. Inf. Model.* **56**, 811 (2016).
44. Betz, R. M. & Walker, R. C. Paramfit: Automated Optimization of Force Field Parameters for Molecular Dynamics Simulations. *J. Comput. Chem.* **36**, 79 (2015).
45. Wang, L.-P., Martinez, T. J. & Pande, V. S. Building Force Fields: An Automatic, Systematic, and Reproducible Approach. *J. Phys. Chem. Lett.* **5**, 1885 (2014).
46. Singh, U. C. & Kollman, P. A. An Approach to Computing Electrostatic Charges for Molecules. *J. Comput. Chem.* **5**, 129 (1984).
47. Bayly, C. I., Cieplak, P., Cornell, W. & Kollman, P. A. A Well-Behaved Electrostatic Potential Based Method Using Charge Restraints for Deriving Atomic Charges: The Resp Model. *J. Phys. Chem.* **97**, 10269 (1993).
48. Chirlian, L. E. & Francl, M. M. Atomic Charges Derived from Electrostatic Potentials: A Detailed Study. *J. Comput. Chem.* **8**, 894 (1987).
49. Breneman, C. M. & Wiberg, K. B. Determining Atom-Centered Monopoles from

Molecular Electrostatic Potentials. The Need for High Sampling Density in Formamide Conformational Analysis. *J. Comput. Chem.* **11**, 361 (1990).

50. Yang, Y., Weaver, M. N. & Merz, K. M. Assessment of the "6-31+ G**+ Lanl2dz" Mixed Basis Set Coupled with Density Functional Theory Methods and the Effective Core Potential: Prediction of Heats of Formation and Ionization Potentials for First-Row-Transition-Metal Complexes. *J. Phys. Chem. A* **113**, 9843 (2009).
51. Heinz, H. & Suter, U. W. Atomic Charges for Classical Simulations of Polar Systems. *J. Phys. Chem. B* **108**, 18341 (2004).
52. Lamoureux, G. & Roux, B. Absolute Hydration Free Energy Scale for Alkali and Halide Ions Established from Simulations with a Polarizable Force Field. *J. Phys. Chem. B* **110**, 3308 (2006).
53. Warshel, A. & Levitt, M. Theoretical Studies of Enzymic Reactions: Dielectric, Electrostatic and Steric Stabilization of the Carbonium Ion in the Reaction of Lysozyme. *J. Mol. Biol.* **103**, 227 (1976).
54. Souaille, M., Loirat, H., Borgis, D. & Gaigeot, M. Mdvry: A Polarizable Classical Molecular Dynamics Package for Biomolecules. *Comput. Phys. Commun.* **180**, 276 (2009).
55. Åqvist, J. & Warshel, A. Free Energy Relationships in Metalloenzyme-Catalyzed Reactions. Calculations of the Effects of Metal Ion Substitutions in Staphylococcal Nuclease. *J. Am. Chem. Soc.* **112**, 2860 (1990).
56. Duarte, F. *et al.* Force Field Independent Metal Parameters Using a Nonbonded Dummy Model. *J. Phys. Chem. B* **118**, 4351 (2014).
57. Martínez, J. M. *et al.* Coupling a Polarizable Water Model to the Hydrated Ion–Water Interaction Potential: A Test on the Cr3+ Hydration. *J. Chem. Phys.* **112**, 2339 (2000).
58. Liao, Q., Pabis, A., Strodel, B. & Kamerlin, S. C. L. Extending the Nonbonded Cationic Dummy Model to Account for Ion-Induced Dipole Interactions. *J. Phys. Chem. Lett.* **8**, 5408–5414 (2017).
59. Armitage, I. M., Drakenberg, T. & Reilly, B. Use of113Cd NMR to probe the native metal binding sites in metalloproteins: An overview. *Met. Ions Life Sci.* **11**, 117–144 (2013).
60. Chakravorty, D. K. *et al.* Solution NMR refinement of a metal ion bound protein using metal ion inclusive restrained molecular dynamics methods. *J. Biomol. NMR* **56**, 125–137 (2013).
61. Otting, G. Protein NMR Using Paramagnetic Ions. *Annu. Rev. Biophys.* **39**, 387–405 (2010).
62. Rinaldelli, M., Carlon, A., Ravera, E., Parigi, G. & Luchinat, C. FANTEN: A new web-based interface for the analysis of magnetic anisotropy-induced NMR data. *J. Biomol. NMR* **61**, 21–34 (2015).
63. Bertini, I. *et al.* Magnetic Susceptibility Tensor Anisotropies for a Lanthanide Ion Series in a Fixed Protein Matrix. *J. Am. Chem. Soc.* **123**, 4181–4188 (2001).
64. Bertini, I., Gelis, I., Katsaros, N., Luchinat, C. & Provenzani, A. Tuning the affinity for lanthanides of calcium binding proteins. *Biochemistry* **42**, 8011–8021 (2003).
65. Rodriguez-Castañeda, F., Haberz, P., Leonov, A. & Griesinger, C. Paramagnetic tagging of diamagnetic proteins for solution NMR. *Magn. Reson. Chem.* **44**, S10–S16 (2006).
66. Allegrozzi, M. *et al.* Lanthanide-induced pseudocontact shifts for solution structure refinements of macromolecules in shells up to 40 ?? from the metal ion. *J. Am. Chem. Soc.* **122**, 4154–4161 (2000).
67. Güntert, P. & Buchner, L. Combined automated NOE assignment and structure calculation with CYANA. *J. Biomol. NMR* **62**, 453–471 (2015).
68. Schwieters, C. D., Kuszewski, J. J., Tjandra, N. & Clore, G. M. The Xplor-NIH NMR molecular structure determination package. *J. Magn. Reson.* **160**, 65–73 (2003).
69. Tjandra, N., Omichinski, J. G., Gronenborn, A. M., Marius Clore, G. & Bax, A. Use of dipolar1H-15N and1H-13C couplings in the structure determination of

magnetically oriented macromolecules in solution. *Nat. Struct. Biol.* **4**, 732–738 (1997).

70. Mao, J. *et al.* Structural basis of the green-blue color switching in proteorhodopsin as determined by NMR spectroscopy. *J. Am. Chem. Soc.* **136**, 17578–17590 (2014).

71. Ryu, J. H., Wee, H. S. & Lee, W. B. Molecular dynamics study on microstructures of diblock copolymer melts with soft potential and potential recovery. *Phys. Rev. E* **94**, 032501 (2016).

72. Case, D. A. *et al.* The Amber biomolecular simulation programs. *J. Comput. Chem.* **26**, 1668–1688 (2005).

73. der Spoel, D. *et al.* GROMACS: Fast, flexible, and free. *J. Comput. Chem.* **26**, 1701–1718 (2005).

74. Bertini, I., Case, D. A., Ferella, L., Giachetti, A. & Rosato, A. A grid-enabled web portal for NMR structure refinement with AMBER. *Bioinformatics* **27**, 2384–2390 (2011).

75. Takegoshi, K., Nakamura, S., Terao, T. & Nakamura, S. 13C–1H dipolar-assisted rotational resonance in magic-angle spinning NMR. *Chem. Phys. Lett.* **344**, 631–637 (2001).

76. Salinas, V. H. & Ranganathan, R. Coevolution-based inference of amino acid interactions underlying protein function. *Elife* **7**, (2018).

77. Uguzzoni, G. *et al.* Large-scale identification of coevolution signals across homo-oligomeric protein interfaces by direct coupling analysis. *Proc. Natl. Acad. Sci.* **114**, E2662–E2671 (2017).

78. Huang, Y. J., Tejero, R., Powers, R. & Montelione, G. T. A topology-constrained distance network algorithm for protein structure determination from NOESY data. *Proteins* **62**, 587–603 (2006).

79. Huang, Y. J., Powers, R. & Montelione, G. T. Protein NMR recall, precision, and F-measure scores (RPF scores): Structure quality assessment measures based on information retrieval statistics. *J. Am. Chem. Soc.* **127**, 1665–1674 (2005).

80. Klepeis, J. L., Lindorff-Larsen, K., Dror, R. O. & Shaw, D. E. Long-timescale molecular dynamics simulations of protein structure and function. *Current Opinion in Structural Biology* **19**, 120–127 (2009).

81. Lane, T. J., Shukla, D., Beauchamp, K. A. & Pande, V. S. To milliseconds and beyond: Challenges in the simulation of protein folding. *Current Opinion in Structural Biology* **23**, 58–65 (2013).

82. Piana, S., Lindorff-Larsen, K. & Shaw, D. E. Protein folding kinetics and thermodynamics from atomistic simulation. *Proc. Natl. Acad. Sci.* **109**, 17845–17850 (2012).

83. Best, R. B. Atomistic molecular simulations of protein folding. *Current Opinion in Structural Biology* **22**, 52–61 (2012).

84. Suárez, E. *et al.* Simultaneous computation of dynamical and equilibrium information using a weighted ensemble of trajectories. *J. Chem. Theory Comput.* **10**, 2658–2667 (2014).

85. Kubelka, J., Hofrichter, J. & Eaton, W. A. The protein folding ‘speed limit’. *Current Opinion in Structural Biology* **14**, 76–88 (2004).

86. Lindorff-Larsen, K., Piana, S., Dror, R. O. & Shaw, D. E. How fast-folding proteins fold. *Science (80-.).* **334**, 517–520 (2011).

87. Li, W., Wang, J., Zhang, J. & Wang, W. Molecular simulations of metal-coupled protein folding. *Curr. Opin. Struct. Biol.* **30**, 25–31 (2015).

88. Bentrop, D. *et al.* Structural and dynamical properties of a partially unfolded Fe4S4 protein: Role of the cofactor in protein folding. *Biochemistry* **38**, 4669–4680 (1999).

89. Miao, Y., Feixas, F., Eun, C. & McCammon, J. A. Accelerated molecular dynamics simulations of protein folding. *J. Comput. Chem.* **36**, 1536–1549 (2015).

90. Bradley, J. M., Le Brun, N. E. & Moore, G. R. Ferritins: Furnishing proteins with iron Topical Issue in Honor of R.J.P. Williams. *J. Biol. Inorg. Chem.* **21**, 13–28 (2016).

91. Chen, T. T. *et al.* TIM-2 is expressed on B cells and in liver and kidney and is a receptor for H-ferritin endocytosis. *J. Exp. Med.* **202**, 955–65 (2005).

92. Li, L. *et al.* Binding and uptake of H-ferritin are mediated by human transferrin receptor-1. *Proc. Natl. Acad. Sci.* **107**, 3505–3510 (2010).

93. Santambrogio, P., Levi, S., Cozzi, A., Corsi, B. & Arosio, P. Evidence that the specificity of iron incorporation into homopolymers of human ferritin L- and H-chains is conferred by the nucleation and ferroxidase centres. *Biochem. J.* **314 (Pt 1**, 139–44 (1996).

94. Theil, E. C. Ferritin: The protein nanocage and iron biomineral in health and in disease. *Inorg. Chem.* **52**, 12223–12233 (2013).

95. Finazzi, D. & Arosio, P. Biology of ferritin in mammals: an update on iron storage, oxidative damage and neurodegeneration. *Arch. Toxicol.* **88**, 1787–1802 (2014).

96. Liu, X. & Theil, E. C. Ferritins: Dynamic management of biological iron and oxygen chemistry. *Acc. Chem. Res.* **38**, 167–175 (2005).

97. Chasteen, N. D. Ferritin. Uptake, storage, and release of iron. *Met Ions Biol Syst* **35**, 479–514 (1998).

98. Melman, G. *et al.* Iron release from ferritin by flavin nucleotides. *Biochim. Biophys. Acta - Gen. Subj.* **1830**, 4669–4674 (2013).

99. Rudeck, M., Volk, T., Sitte, N. & Grune, T. Ferritin oxidation in vitro: implication of iron release and degradation by the 20S proteasome. *IUBMB Life* **49**, 451–6 (2000).

100. Kim, M. *et al.* PH-dependent structures of ferritin and apoferritin in solution: Disassembly and reassembly. *Biomacromolecules* **12**, 1629–1640 (2011).

101. Kolaj-Robin, O., Russell, D., Hayes, K. A., Pembroke, J. T. & Soulimane, T. Cation diffusion facilitator family: Structure and function. *FEBS Lett.* **589**, 1283–1295 (2015).

102. Wei, Y. & Fu, D. Selective metal binding to a membrane-embedded aspartate in the Escherichia coli metal transporter YiiP (FieF). *J. Biol. Chem.* **280**, 33716–33724 (2005).

103. Lu, M., Chai, J. & Fu, D. Structural basis for autoregulation of the zinc transporter YiiP. *Nat. Struct. Mol. Biol.* **16**, 1063–1067 (2009).

104. Coudray, N. *et al.* Inward-facing conformation of the zinc transporter YiiP revealed by cryoelectron microscopy. *Proc. Natl. Acad. Sci. U. S. A.* **110**, 2140–5 (2013).

105. Lopez-Redondo, M. L., Coudray, N., Zhang, Z., Alexopoulos, J. & Stokes, D. L. Structural basis for the alternating access mechanism of the cation diffusion facilitator YiiP. *Proc. Natl. Acad. Sci.* 201715051 (2018). doi:10.1073/pnas.1715051115

106. Shi, Y. Common Folds and Transport Mechanisms of Secondary Active Transporters. *Annu. Rev. Biophys.* **42**, 51–72 (2013).

107. Gupta, S. *et al.* Visualizing the kinetic power stroke that drives proton-coupled zinc(II) transport. *Nature* **512**, 101–4 (2014).

108. Rutter, G. A. & Chimienti, F. SLC30A8 mutations in type 2 diabetes. *Diabetologia* **58**, 31–36 (2015).

109. Solomou, A. *et al.* Over-expression of Slc30a8/ZnT8 selectively in the mouse α cell impairs glucagon release and responses to hypoglycemia. *Nutr. Metab. (Lond).* **13**, 46 (2016).

110. Davidson, H. W., Wenzlau, J. M. & O'Brien, R. M. Zinc transporter 8 (ZnT8) and β cell function. *Trends Endocrinol. Metab.* **25**, 415–424 (2014).

111. Flannick, J. *et al.* Loss-of-function mutations in SLC30A8 protect against type 2 diabetes. *Nat. Genet.* **46**, 357–363 (2014).

112. Parsons, D. S., Hogstrand, C. & Maret, W. The C-terminal cytosolic domain of the human zinc transporter ZnT8 and its diabetes risk variant. *FEBS J.* **285**, 1237–1250 (2018).

113. Merriman, C., Huang, Q., Rutter, G. A. & Fu, D. Lipid-tuned zinc transport activity of human ZnT8 protein correlates with risk for type-2 diabetes. *J. Biol. Chem.* **291**, 26950–26957 (2016).

114. Huang, J. *et al.* CHARMM36m: An improved force field for folded and intrinsically disordered proteins. *Nat. Methods* (2016). doi:10.1038/nmeth.4067

115. Robustelli, P., Piana, S. & Shaw, D. E. Developing a molecular dynamics force field for both folded and disordered protein states. *Proc. Natl. Acad. Sci.* (2018). doi:10.1073/pnas.1800690115

116. Zanzoni, S. *et al.* Unsaturated Long-Chain Fatty Acids Are Preferred Ferritin Ligands That Enhance Iron Biomineralization. *Chem. - A Eur. J.* **23**, 9879–9887 (2017).