Chapter 1

3D Computational Modeling of Proteins Using Sparse Paramagnetic NMR Data

Kala Bharath Pilla, Gottfried Otting, and Thomas Huber

Abstract

Computational modeling of proteins using evolutionary or de novo approaches offers rapid structural characterization, but often suffers from low success rates in generating high quality models comparable to the accuracy of structures observed in X-ray crystallography or nuclear magnetic resonance (NMR) spectroscopy. A computational/experimental hybrid approach incorporating sparse experimental restraints in computational modeling algorithms drastically improves reliability and accuracy of 3D models. This chapter discusses the use of structural information obtained from various paramagnetic NMR measurements and demonstrates computational algorithms implementing pseudocontact shifts as restraints to determine the structure of proteins at atomic resolution.

Key words Pseudocontact shifts, PCS, Paramagnetic NMR, Rosetta, GPS-Rosetta, Sparse restraints, 3D structure determination

1 Introduction

Nuclear magnetic resonance (NMR) spectroscopy has for decades facilitated structure determination in solution or solid-state. NMR exploits the nuclear spin properties in strong constant magnetic fields. The nuclear spins are manipulated by radiofrequency pulses and their free induction decay is recorded. These are then Fourier transformed to produce a frequency spectrum of the NMR experiment. Two spins that are close in space have a direct magnetic interaction between them, referred as dipole–dipole coupling. When these two spins are aligned, the interaction energy becomes minimal resulting in nuclear Overhauser effect (NOE). Intermolecular and intramolecular NOEs are observed for spins that are typically separated by 3–6 Å. By resolving a dense network of NOEs [1], the 3D structures of proteins and nucleic acids can be determined. This conventional method is relied upon in structure determination of a large number of proteins; however, assigning spin resonances of all spins in the system typically requires various 3D or

Jonathan M. Keith (ed.), *Bioinformatics: Volume II: Structure, Function, and Applications*, Methods in Molecular Biology, vol. 1526, DOI 10.1007/978-1-4939-6613-4_1, © Springer Science+Business Media New York 2017

4D NMR experiments to be applied. In addition, with increasing molecular weight of proteins, they tend to produce poor spectra and determining 3D structures becomes increasingly difficult.

As an alternative to short range restraints using NOEs, paramagnetic NMR generates versatile structural restraints. Proteins carrying paramagnetic metal ions induce significant effects in NMR experiments. These effects arise from the unpaired electrons of the paramagnetic metals, as electrons have a magnetic moment that is three orders of magnitude larger than that of a proton. Metalloproteins, which make up to 25 % of proteins in any organism's proteome [2], offer natural metal centers that potentially can be directly exploited in paramagnetic NMR experiments. Further, Mn^{2+}, Fe2^{+}, Cu^{2+}, and Co^{2+} are naturally paramagnetic and found in native biological samples.

Lanthanide ions are highly useful for paramagnetic NMR experiments, as their paramagnetism varies greatly while their physicochemical properties are highly similar. This makes it possible for different lanthanides to be used interchangeably in different NMR experiments [3]. Proteins that lack a natural metal center can be engineered to carry lanthanides. Figure 1 illustrates different ways to introduce metal ions into proteins. Small peptides, containing 12–18 residues, are designed to bind lanthanide ion to their side chain atoms and these peptides are attached to either a thiol-reactive cysteine or at an N- or C-terminus of a protein [4]. The most popular means of attaching lanthanide ions is through metal chelating chemical tags. These chemical tags are site specifically attached either through cysteine ligation or more recently using unnatural amino acids which can be reacted via bio-orthogonal click chemistry [5]. Several reviews [6–9] provide a comprehensive overview of the chemistries to functionalise proteins with lanthanide tags.

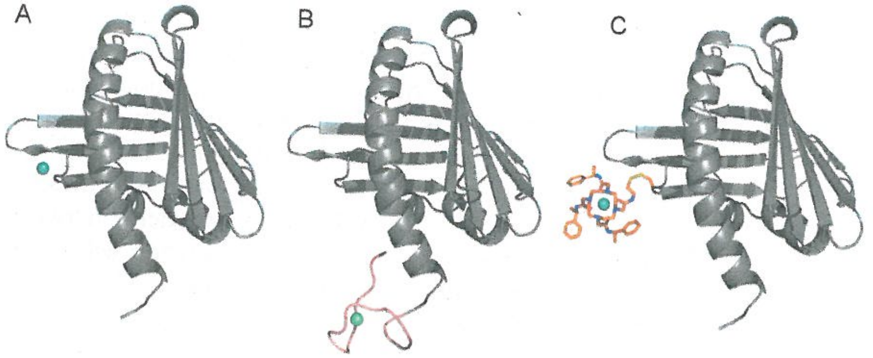

Fig. 1 Illustration of various modes to introduce metal ions into proteins. (**a**) Replacing a native metal with paramagnetic lanthanide ion in metalloproteins. (**b**) Lanthanide binding peptides attached at C-terminus of a protein. (**c**) Lanthanide carrying chemical tag site specifically attached to a cysteine

1.1 Paramagnetic Effects in NMR

The unpaired electrons in a paramagnetic metal ion strongly interact with nuclear spins and the NMR spectrum changes due to induced paramagnetic effects. These paramagnetic effects are quantified by comparing with a diamagnetic (reference) spectra and then translated into structural restraints. The resulting structural restraints can be either distance dependent or orientation dependent or both. One can measure four distinct paramagnetic observables from NMR experiments, namely:

1.1.1 Pseudocontact Shift (PCS)

PCS is a contribution to the chemical shift experienced by a spin caused by the presence of centers of unpaired electrons. PCS of a nucleus influenced by a paramagnetic center can be calculated from a $\Delta\chi$-tensor, shown in Fig. 2a, given by:

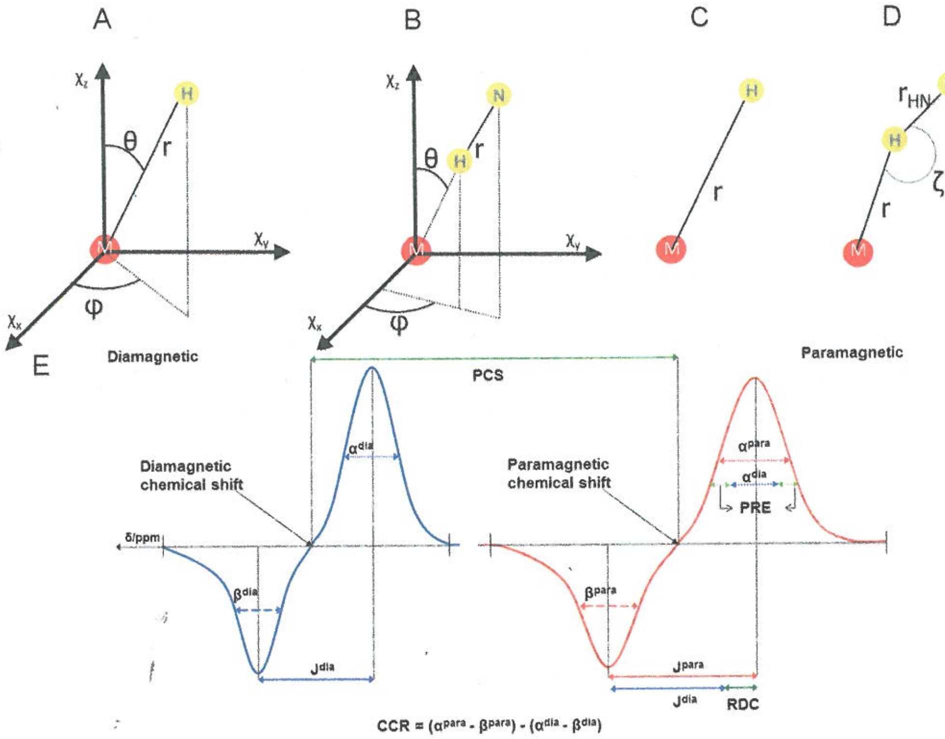

Fig. 2 The four distinct paramagnetic effects represented geometrically. (**a**) The pseudocontact shift (PCS) between metal center (M) and amide hydrogen (H). (**b**) The residual dipolar coupling (RDC) between two spins H and N. (**c**) The Paramagnetic relaxation enhancement (PRE) between m and H. (**d**) The cross correlation between Curie spin and dipole–dipole relaxation (CCR) between m and H. (**e**) Measurement of the four different paramagnetic effects, illustrated with two 1D undecoupled spectra, showing the diamagnetic and paramagnetic antiphase doublets. PCS is measured as the change in chemical shift between paramagnetic and diamagnetic states. RDC is measured as the difference in line splitting. PRE and CCR can be determined from the differential line broadening. Adapted from Schmitz (2009) [49]

$$\mathrm{PCS}_i^{\mathrm{calc}} = \frac{1}{12\pi r_{\mathrm{MH}}^3}\left[\Delta\chi_{\mathrm{ax}}\left(3\cos^2\theta_{\mathrm{MH}} - 1\right) + \frac{3}{2}\Delta\chi_{\mathrm{rh}}\sin^2\theta_{\mathrm{MH}}\cos 2\varphi_{\mathrm{MH}}\right] \tag{1}$$

where, r, θ, φ define the polar coordinates of the nuclear spin with respect to principal axis of the $\Delta\chi$-tensor (centered on the paramagnetic ion) and $\Delta\chi_{\mathrm{ax}}$, $\Delta\chi_{\mathrm{rh}}$ define the axial and rhombic component of the magnetic susceptibility tensor χ and $\Delta\chi$-tensor is defined as χ-tensor minus its isotropic component [10]. PCS is measured as change in the chemical shift of a spin's paramagnetic and diamagnetic states, illustrated in Fig. 2e.

1.1.2 Residual Dipolar Coupling (RDC)

Presence of paramagnetic metal weakly aligns the protein to an external magnetic field resulting in observable RDCs, which are manifested as an increase or decrease in magnitude of multiplet of splits that can be observed in undecoupled spectra, illustrated in Fig. 2e. The RDC is given by Eq. (2) shown in Fig. 2b:

$$D_{\mathrm{NH}} = -\frac{B_0^2}{15kT}\cdot\frac{\gamma_{\mathrm{H}}\gamma_{\mathrm{N}}\hbar}{8\pi^2 r_{\mathrm{NH}}^3}\left[\Delta\chi_{\mathrm{ax}}(3\cos^2\theta_{\mathrm{NH}} - 1) + \frac{3}{2}, \Delta\chi_{\mathrm{rh}}\sin^2\theta_{\mathrm{NH}}, \cos 2\varphi_{\mathrm{NH}}\right] \tag{2}$$

where B_0 is the magnetic field strength, γ_{H} and γ_{N} are the gyromagnetic ratios of the proton and nitrogen spin, $\hbar = h/2\pi$ with h being Planck's constant, r_{NH} is the distance between the nitrogen and proton nuclei [11].

1.1.3 Paramagnetic Relaxation Enhancement (PRE)

PREs give distance restraints between the paramagnetic lanthanide and spin of interest from peak intensity ratios between paramagnetic and diamagnetic states (Fig. 2e). The PRE is given by Eq. (1) shown in Fig. 2c.

$$\lambda^{\mathrm{PRE}} = \frac{K}{r^6}\left(4\tau_{\mathrm{r}} + \frac{3\tau_{\mathrm{r}}}{1 + \omega_{\mathrm{H}}^2\tau_{\mathrm{r}}^2}\right) \tag{3}$$

with,

$$K = \frac{1}{5}\left(\frac{\mu_0}{4\pi}\right)^2 \frac{B_0^2\gamma_{\mathrm{H}}^2\left(g_j\mu_{\mathrm{B}}\right)^4 J^2(J+1)^2}{(3k_{\mathrm{B}}T)^2} \tag{4}$$

where τ_{r} is the rotational correlation time, ω_{H} is the Larmor frequency of the proton, μ_0 is the vacuum permeability, g_j the g-factor, μ_{B} the Bohr magneton, and J the total spin moment [11].

1.1.4 Cross Correlated Relaxation (CCR)

This effect is measured by comparing the line width between the two components of the antiphase doublet (Fig. 2e) [11]. This effect combines distance and angle dependence given by Eq. (3) shown in Fig. 2d.

$$\eta^{\mathrm{CCR}} = K \frac{3\cos^2\eta - 1}{r^3}\left(4\tau_{\mathrm{r}} + \frac{3\tau_{\mathrm{r}}}{1 + \omega_{\mathrm{H}}^2\tau_{\mathrm{r}}^2}\right) \tag{5}$$

with,

$$K = \frac{1}{30}\left(\frac{\mu_0}{4\pi}\right)^2 \frac{B_0^2\gamma_{\mathrm{H}}^2\left(g_J\mu_{\mathrm{B}}\right)^4 J^2(J+1)^2}{(3k_{\mathrm{B}}T)^2} \tag{6}$$

1.2 Structural Information from Paramagnetic Effects

RDCs, which are defined from the molecular alignment tensor (Eq. 2, Fig. 2b), give the orientation of spin pairs relative to the external magnetic field in a distance independent fashion. RDCs by themselves can be directly used to determine the structure of small proteins only when a large number of experimental RDCs are available. Measurement of heteronuclear RDCs becomes difficult for proteins that exhibit limited solubility or produce broad NMR line widths due to tag mobility.

PREs on the other hand give distance information from the paramagnetic center (Eq. 1, Fig. 2c). PREs induced by lanthanide ions range up to 20 Å, but the effect is heavily influenced by the motion of the metal carrying tag [12]. Direct usage of PREs in structure determination is limited but chemically inert paramagnetic probes when added as co-solvents can be quantitatively used to characterize interfaces in protein–protein complexes.

1.2.1 Uniqueness of PCS

In comparison to RDCs and PREs, PCSs are the most potent structural restraints. A PCS defined by the $\Delta\chi$-tensor is both orientation and distance dependent (Eq. 1, Fig. 2a). The PCS effect has the longest range among all the paramagnetic effects and extends up to 80 Å (40 Å from the paramagnetic center) and can be precisely measured even at low protein concentrations (<20 μM) [13]. It can be easily seen from the $\Delta\chi$-tensor defined in Eq. (1) that PCS influenced by a spin is proportional to r^{-3} from the metal center, which decays slower with distance than PRE with r^{-6} dependence (Eq. 1). RDCs, in stark contrast to PCSs and PREs, are only orientation dependent (Eq. 2) brought about by the weak alignment from the inserted paramagnetic metal [8].

Experimentally PCSs are easy to measure in proteins by taking the difference in chemical shifts of a protein's paramagnetic and diamagnetic states from simple 2D NMR spectra (shown in Fig. 3a). PCS can also be measured with higher accuracy and sensitivity compared to other paramagnetic effects, such as measuring coupling constants between nuclei for RDCs and measuring peak intensities for PREs. The induced PCS described within the $\Delta\chi$-tensor can be visualized as isosurfaces of constant PCS (shown in Fig. 3b). The $\Delta\chi$-tensor is fully defined by eight parameters, the origin of the tensor frame which coincides with the coordinates of

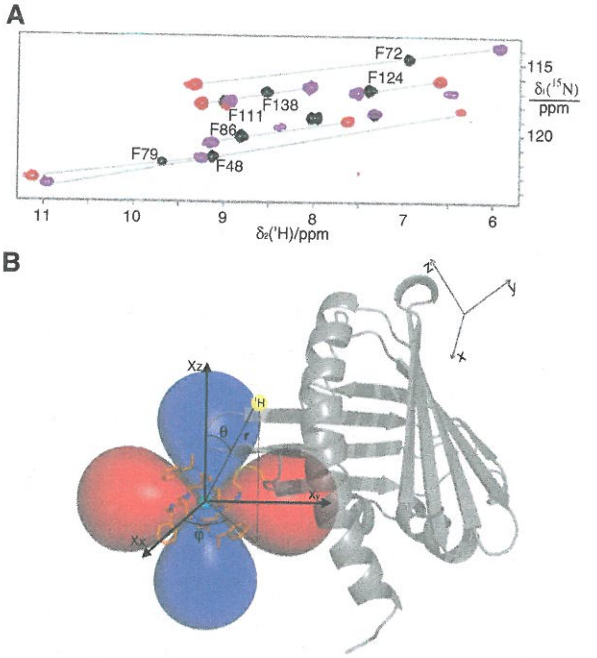

Fig. 3 Measurement of pseudocontact shift (PCSs) and display of PCS as isosurfaces. (**a**) An illustration of three superimposed ^{15}N-HSQC spectra, showing the chemical shift changes due to presence of paramagnetic metal ions in the protein. Black resonances come from the diamagnetic reference (Y^{3+}) sample, while red (Dy^{3+}) and magenta (Er^{3+}) resonances show chemical shift changes due to the paramagnetic lanthanide ions attached in the sample. (**b**) Visualization of induced PCS as isosurfaces calculated from the $\Delta\chi$-tensor

the metal (x,y,z), orientation of $\Delta\chi$-tensor frame (three Euler angles α, β, γ) with respect to the coordinate frame of the protein, and two components of the $\Delta\chi$-tensor, $\Delta\chi_{\mathrm{ax}}$ (axial) and $\Delta\chi_{\mathrm{rh}}$ (rhombic). To solve for the full mathematical description of a $\Delta\chi$-tensor one needs to measure a minimum of eight PCSs.

2 PCSs in Protein Structure Characterization

2.1 Paramagnetic NMR Spectrum Assignment

Accurate assignment of resonances in the NMR spectrum is the essential first step in extracting restraints. Especially for large proteins (>20 kDa), assignment of multidimensional NMR spectra becomes increasingly difficult due to spectral overlap and increased transverse relaxation of spins. If 3D atomic coordinates of nuclear spins are known, the NMR resonance assignments of both paramagnetic and diamagnetic spectra can be assigned with software algorithms. Several software algorithms are available to assist with NMR assignments, including Numbat [14], Possum [15], Echidna [16], and PARAssign [17].

2.2 Protein–Ligand Interactions

PCSs can be measured not only on the protein's nuclear spins but also on the spins of the bound ligands. With the availability of a diverse range of metal binding chemical tags, the orientation and location of the ligand can be easily identified with the help of PCSs [3, 9]. This ability has major implications for rational drug design. John et al. [18] have demonstrated this concept using *E. coli*'s $\varepsilon 186/\theta$ (a natural lanthanide binding protein) in complex with the ligand thymidine, where the ligand affinity and its binding orientation was entirely determined using only PCSs. Saio et al. [19] showed that a combination of PCSs and PREs generated from two point anchored lanthanide binding peptide can be used to screen for ligands for protein Grb2. Guan et al. [20] showed that even in the absence of isotope labeled protein samples, the location of the ligand bound to the protein can be determined in low resolution with predicted $\Delta\chi$-tensor parameters.

2.3 Protein–Protein Complexes

Protein–protein complexes are fundamental to the function of cellular signaling and function. If 3D structures of the interacting protein partners are known, then the directionality and distance dependence of the $\Delta\chi$-tensor can be exploited in docking the interacting partners in the right orientation. Pintacuda et al. [21] reported the first demonstration of the use of PCSs to compute the structure of a protein complex, using the interacting partners of *E. coli* DNA polymerase complex's N-terminal domain of the subunits ε and θ. Recent studies involving a large PCS data set (446 PCSs) have been used to characterize cytochrome P450cam in complex with putidaredoxin using double cysteine anchored tag [22]. PCS restraints are incorporated into protein–protein docking program Haddock, where the orientation of interacting partners and $\Delta\chi$-tensors are simultaneously fitted for finding optimized interacting surfaces [23].

2.4 Protein Structure Refinement

If the coordinates of atoms in the protein are known, PCSs can be effectively used to refine protein structures. Allegrozzi et al. [24] showed that NOE derived structural models can be further refined using PCSs that are measured using three different lanthanides (Ce^{3+}, Yb^{3+}, and Dy^{3+}), which have different coverage range over the protein. Supplementing PCS restraints on the protein calbindin decreased the overall RMSD over NOE derived NMR structures. Gaponenko et al. [25] showed that using PCS data generated from three different lanthanide attachment sites extended the refinement approach to proteins larger than 30 kDa. PCS refined structures showed improvement over an Ångström RMSD when compared to NOE only structures, and this improved accuracy is also validated using RDCs. Other paramagnetic restraints also have been used in a similar manner. Sparse datasets of RDCs combined with sparse NOEs have been used to identify the best models from a pool of structures generated using homology modeling [26] and de novo methods [27].

To directly use PCSs for structure calculation is challenging as one needs to determine the eight parameters to describe the $\Delta\chi$-tensor, which are difficult to estimate as they depend on the chemical environment of the metal. Without the knowledge of 3D coordinates of the protein it is not possible to fit the $\Delta\chi$-tensor to reproduce the experimentally observed PCSs. However, one can use PCSs as restraints in de novo structure prediction methods such as Rosetta [28]. Rosetta's forcefield accurately describes the protein state and the software algorithms are designed to robustly search the conformational space accessible to the protein.

3 Protein Structure Determination Using PCS and Rosetta

Incomplete or sparse structural data generated from NMR experiments can be used as structural restraints in Rosetta calculations to facilitate structure determination. Unlike traditional methods where structure calculation is mainly determined by the completeness of experimental data which defines the position of atomic coordinates in a protein structure, the sparse NMR data is used to guide the conformational search which directs the sampling towards the global minimum. Different types of NMR measurements have been incorporated as additional scoring restraints in Rosetta. Chemical shift measurements combined with predicted backbone dihedrals and secondary structure elements can be used in picking fragments that match the prediction, a procedure known as CS-Rosetta [29, 30]. Backbone NOEs in combination with RDCs have also been included in protein structure determination [28] using an advanced genetic algorithm [31]. Incorporation of sparse NMR data in structure calculations has been shown to improve protein structure predictions.

3.1 Rosetta Structure Calculation Algorithm

Based on folding studies of small proteins, Rosetta's algorithms are built on the assumption that the ensemble of local structures sampled by a sequence fragment can be approximated by a small number of local structures that a similar fragment adopts in known protein structures [32]. For a given protein sequence whose structure is to be determined, the sequence is decomposed into overlapping windows of nine and three residues. The fragment libraries are constructed for each of the nine and three residue windows by searching through 3D structure databases for protein fragments whose sequences or secondary structures have high similarity to that of the query. The corresponding backbone dihedral angles of the matched protein fragments are bundled up into fragment libraries The search for the lowest energy structure is carried out by assembling the fragments into protein-like structures using Metropolis Monte-Carlo and simulated annealing algorithms [33]. Starting from a linear polypeptide, the search is carried out in

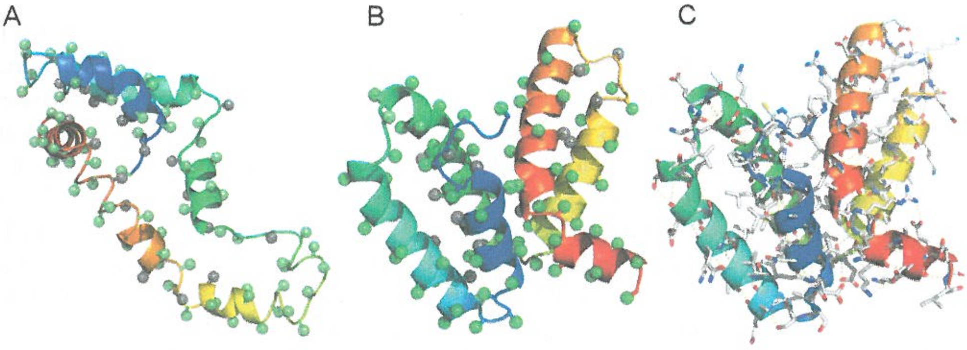

Fig. 4 Illustration of Rosetta's ab initio fragment assembly. (**a**) A protein decoy in an intermittent state during fragment assembly. The backbone atoms are shown in a cartoon representation and the side chain atoms are represented as spheres attached to Cβ atoms. The hydrophobic residues represented in *grey* and the solvent accessible residues represented in *green*. (**b**) Final fold of the protein shown in (**a**) after the low resolution fragment assembly. (**c**) All-atom representation of the final fold of the protein with complete side-chain atoms

two distinct phases, a low resolution centroid mode and a high resolution all-atom mode [34].

3.1.1 Centroid Mode

In this mode, the conformational search is carried out in a low-resolution phase, in which the amino acid residues are represented in a stripped down version that lacks complete side chain detail. The side chains are represented as spheres attached to the backbone (Cβ and beyond) at their centroid point as shown in Fig. 4a. The fragment assembly follows Monte-Carlo moves starting from an arbitrary position from a random nine residue fragment window. For every move, which replaces the coordinates of a protein segment from that of a fragment library, the energy of the resultant protein decoy is evaluated. The scoring function in the centroid phase is a coarse-grained description of probabilistic functions which favors the formation of globular compact structures. This scoring function explicitly scores for electrostatic and solvation effects among residues which are based on the observed distributions in known proteins. Formation of secondary structural elements in the folding pathway is encouraged with distinct function terms that favor helix–helix, helix–sheet, and sheet–sheet pairing. This low resolution centroid mode generates protein like decoy structures, in which the polar amino acids are exposed to the solvent while burying the hydrophobic residues in the core of the protein (shown in Fig. 4b). Multiple folding pathways are independently sampled, generating tens to hundreds of thousands of protein decoy structures to sample the vast conformational space.

3.1.2 All-Atom Mode

This mode generates complete and optimized placement of side-chain coordinates (shown in Fig. 4c). Here side chains are modeled by searching through discrete combinations of amino acid rotamers by simulated annealing. To further optimize the geometry, multistep Monte Carlo minimisation is enforced on each decoy; steps include torsion angle perturbations, one-at-a-time rotamer optimization and continuous gradient based minimisation of backbone torsion angles and side chain coordinates. The scoring function during this stage is more detailed, physically realistic, accurate to the atomic level and computationally expensive. Hydrogen bonding is explicitly included in the analysis. Hydrogen bonding terms are knowledge-based terms which are orientation and secondary structure dependent and were derived from high resolution protein structures. Typically, multiple independent trajectories are first clustered and atomic details are generated on the desired cluster [33].

3.1.3 PCS Restraints in Rosetta

In the centroid mode, at each instance of a fragment move, $\Delta\chi$-tensors are fitted to the assembled structure and PCSs are back-calculated. The difference between the input and back-calculated PCSs are then used as a quality score to guide assembly to the right fold of the protein. It has been shown that using PCSs from a single metal center, 3D protein structures up to 150 amino acid residues can be determined at atomic resolution [35]. However, this method is limited in its application for proteins larger than 150 amino acids.

The primary limitation associated with the PCSs measured from a single metal center is the reduction in quality of PCS data. Lanthanide tags attached to a single metal center often fail to induce significantly large PCS for most of the spins in the protein. This loss of data is pronounced in large molecular weight proteins. Secondly, there is additional loss of data due to induced PRE effect by the lanthanide ions, where NMR signals of the spins near the vicinity of the lanthanides are broadened beyond detection.

3.2 Extending PCS Scoring to Multiple Metal Centers

To resolve the ambiguities associated with the PCS data generated from a single metal center and to achieve complete coverage, the approach has been extended from a single metal center to multiple metal centers. A second PCS measured for the same nucleus from a lanthanide attached at a different site restricts the spin to lie on intersecting isosurfaces. A third PCS measured from a lanthanide attached at a site different from the first two would further restrict the location of the spin in space. This technique, which is analogous to the method of finding a location on Earth from three or more GPS satellites, is incorporated into the Rosetta framework and was dubbed GPS-Rosetta [36].

3.2.1 Scoring PCS Data from Multiple Metal Centers in Rosetta

The $\Delta\chi$-tensor from Eq. (1) can be rewritten as

$$\mathrm{PCS}_i^{\mathrm{calc}} = \frac{1}{12\pi r_i^5} \cdot \mathrm{Trace}\left[\begin{pmatrix} 3x_i^2 - r_i^2 & 3x_i y_i & 3x_i z_i \\ 3x_i y_i & 3y_i^2 - r_i^2 & 3y_i z_i \\ 3x_i z_i & 3y_i z_i & 3z_i^2 - r_i^2 \end{pmatrix} \begin{pmatrix} \Delta\chi_{xx} & \Delta\chi_{xy} & \Delta\chi_{xz} \\ \Delta\chi_{xy} & \Delta\chi_{yy} & \Delta\chi_{yz} \\ \Delta\chi_{xz} & \Delta\chi_{yz} & \Delta\chi_{zz} \end{pmatrix}\right] \tag{7}$$

where, r_i is the distance between the spin i and the paramagnetic center M; x_i, y_i, and z_i are the Cartesian coordinates of the vector between the metal ion and the spin i in an arbitrary frame f; and $\Delta\chi_{xx}$, $\Delta\chi_{yy}$, $\Delta\chi_{zz}$, $\Delta\chi_{xy}$, $\Delta\chi_{xz}$, and $\Delta\chi_{yz}$ are the $\Delta\chi$-tensor components in the frame f (as $\Delta\chi_{zz} = -\Delta\chi_{xx} - \Delta\chi_{yy}$, there are only five independent parameters). The determination of $\mathrm{PCS}_i^{\mathrm{calc}}$ (Eq. 5) poses a nonlinear least-square fit problem, which can be divided into its linear and nonlinear parts. $\mathrm{PCS}_i^{\mathrm{calc}}$ is linear with respect to the five $\Delta\chi$-tensor components which can be optimized efficiently using singular value decomposition. With the knowledge of the location of the chemical tag used, search over the metal coordinates x_{M}, y_{M}, and z_{M} of the paramagnetic center can be carried out on a 3D grid. The 3D grid is defined with parameters which include center of the grid search (cg), step size between two nodes (sg), an outer cutoff radius (co) which limits the search to a minimal distance from cg and an inner cutoff radius (ci) to avoid a search too close to cg [35].

PCSs recorded from multiple lanthanide carrying chemical tags are given as input into Rosetta by constructing multiple 3D grids for individual tag site. For each PCS dataset per metal and chemical tag, the $\Delta\chi$-tensor components are fitted at each node of the 3D grid and the PCSs are back-calculated. The grid node with the lowest score obtained from Eq. (6) is then taken as the starting point to further optimize the metal position and the five components of the $\Delta\chi$-tensor to reach the minimum cost for all the metal centers.

$$s_k = \sum_{q=1}^{m} \sqrt{\sum_{p=1}^{n_{\mathrm{pcs}}} \left(\mathrm{PCS}_{\mathrm{calc}}^{pq} - \mathrm{PCS}_{\mathrm{exp}}^{pq}\right)^2} \tag{8}$$

where m is the number of PCS data sets (one dataset per metal ion) per binding site k and n_{pcs} is the number of PCSs in the dataset. A total weighted sum of square deviations are used as PCS scoring S_{total} and added to the low-resolution energy function of Rosetta:

$$S_{\mathrm{total}} = \sum_{k=1}^{n} s_k \cdot w_k \tag{9}$$

where n is the total number of metal binding centers and w denotes the weighting factor relative to the Rosetta ab initio scoring function. The weighting factor w for each of the n centers was calculated independently by

$$w = \left(\frac{a_{\text{high}} - a_{\text{low}}}{c_{\text{high}} - c_{\text{low}}} \right) / n \tag{10}$$

where a_{high} and a_{low} are the averages of the highest and lowest 10 % of the values of the Rosetta ab initio score, and c_{high} and c_{low} are the averages of the highest and lowest 10 % of PCS score obtained by rescoring 1000 decoys with unity weighting factor.

4 The GPS-Rosetta Algorithm

The algorithm incorporating PCS scoring from multiple metal centers in Rosetta's structure determination protocol is described as a flow chart in Fig. 5. The $\Delta\chi$-tensors for each dataset from multiple sites are simultaneously optimized and the weighted PCS scores for individual metal sites are added to the centroid scoring function. Side chain atoms are then added to all the structural decoys and scored using Rosetta's all-atom scoring function. The PCS scoring is not used in this mode, because only minor changes in the backbone structure are generated. The side chain optimized structural models are rescored with PCS data from multiple metal centers with new weights generated using Eq. (8), except that they are now weighted against Rosetta's all-atom scoring function. The top structures are selected based on lowest combined scores of Rosetta's all-atom score and weighted PCS score from all the tag sites.

GPS-Rosetta protocol has been implemented in determining 3D structures from PCSs data generated from two different NMR experiments, solution state NMR and magic angle spinning (MAS) solid-state NMR experiments. C-terminal domain of endoplasmic reticulum protein 29, ERp29-C (106 residues) from rat, is determined from the PCS data generated at 4 different metal centers in solution state and Immunoglobulin Binding Domain of Protein G, GB1 (56 residues) from *Streptococcus* spp, is determined from PCS data generated at three different metal centers in microcrystalline state.

4.1 Fold Determination Using PCSs from Solution NMR Experiments

ERp29-C is a chaperone protein expressed in the endoplasmic reticulum of a mammalian cell, where it facilitates the folding and transport of other protein molecules. The 3D structure was first determined by solution NMR using a conventional NOE approach, and the result is referred to as the NOE structure [37]. However, the crystal structure of human ERp29-C [38] shows a significantly different fold with Cα root mean squared deviation (RMSD) of 4.5 Å when compared to the NOE structure. GPS-Rosetta protocol was employed to reassess the structure in solution [36]. Four different sites on the protein were chosen to bind two different lanthanide tags. The cysteine ligated, C1 tag [39] was chosen to

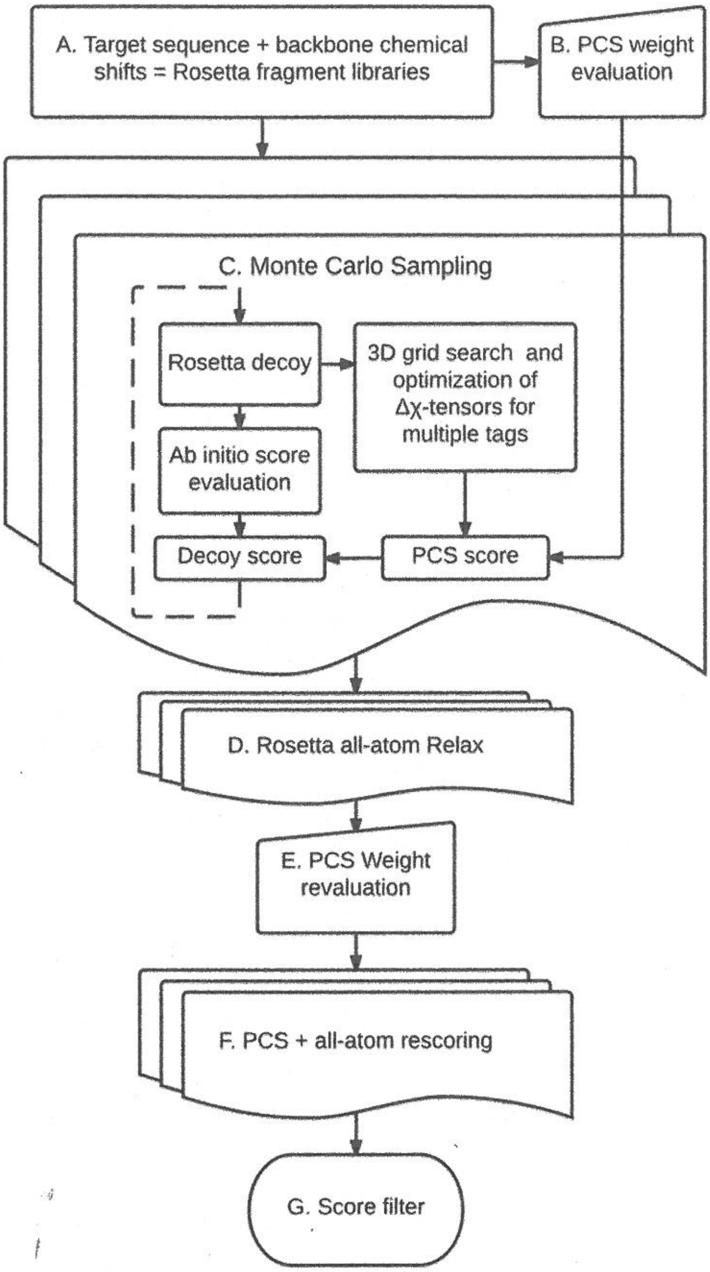
A. Target sequence + backbone chemical shifts = Rosetta fragment libraries
B. PCS weight evaluation
C. Monte Carlo Sampling
Rosetta decoy
3D grid search and optimization of Δχ-tensors for multiple tags
Ab initio score evaluation
Decoy score
PCS score
D. Rosetta all-atom Relax
E. PCS Weight revaluation
F. PCS + all-atom rescoring
G. Score filter

Fig. 5 Flowchart illustrating series of steps involved in running GPS-Rosetta protocol. (**a**) Short nine and three residue fragments are generated based on target sequence and secondary structure prediction based on backbone chemical shifts. (**b**) PCS weights are calculated using Eq. (8). (**c**) Centroid models are generated by fragment assembly following Metropolis Monte-Carlo sampling algorithm. PCS scores for individual tag sites are independently optimized and the PCS scores are added to Rosetta's scoring function. (**d**) Side chain generation and optimization to centroid models. (**e**) PCS score for individual tag sites are reweighted from all-atom models. (**f**) Models are rescored with PCSs and Rosetta's all-atom scoring function. (**g**) Final structure is selected based on lowest combined score value

bind at the native cysteine (C157) and IDA-SH tag [40] was attached at double mutants S200C/K204D, A218C/A222D, and Q241C/N245D. All the double mutations were on α-helices and the aspartate residue at $(i + 4)$th position forming a specific lanthanide binding site. The side chain carboxyl-oxygen of the aspartate served as an additional coordination site to immobilize the lanthanide ion. The PCS dataset from eight paramagnetic samples is composed of a total of 212 PCSs measured using lanthanides Tb^{3+}, Tm^{3+}, and Y^{3+}, where Y^{3+} served as diamagnetic reference.

The unique coordination feature of IDA-SH enabled determination of the position of the metal ion at 5.9 Å from the Cα of $(i + 4)$th residue, lying on a vector that joins the backbone amide nitrogen at $(i + 6)$ and Cα of $(i + 4)$th aspartate. The lanthanide position defined by C1 tag at C157 was dynamically optimized during the folding simulation. More than 100,000 all-atom models were generated using GPS-Rosetta protocol and multiple structures satisfying combined Rosetta and PCS score and experimental data were selected. The final structure was selected for the model that has the lowest Rosetta's all-atom and weighted PCS energy. The final selected structure, which is represented by the red point in Fig. 6a, has a backbone Cα RMSD of 2.4 Å to the crystal structure (Fig. 6b) [PDBID: 2QC7;[38]], and is referred to as the GPS-Rosetta model. The top five structures that are lowest in PCS RMSD are shown in blue points and the top five models with an arbitrary low combined score and low PCS RMSD are represented as green points (Fig. 6a). The GPS-Rosetta structure was compared against the crystal structure and top 10 selected structures. Superposition structures with low PCS RMSD are represented in shades of blue (Fig. 6b), and low scoring in PCS and Rosetta energy and low PCS RMSD are represented in shades of green (Fig. 6c). The Cα RMSD of all the selected structures lies in the range 2.0–2.9 Å to the crystal structure with the exception of small variations in the orientation of the C-terminal residues which were reported to be disordered [37]. The GPS-Rosetta structure, in red, (PDBID: 2M66) clearly resembles the crystal structure more closely (Fig. 6b, RMSD of 2.4 Å) than the NOE structure (Fig. 6d, RMSD 6 Å), effectively overruling it.

4.2 High Resolution Protein Structure Determination Using PCSs from MAS Solid-State NMR Experiments

MAS solid-state NMR spectroscopy has been routinely employed to determine structure of membrane biomolecules and proteins that are difficult to study by solution NMR or X-ray crystallography [41]. 3D structures are determined by resolving large number of dipolar couplings between ^{1}H, ^{13}C, and ^{15}N nuclei [42, 43]; however, the spectrum resolves in densely packed cross-peaks which are highly difficult to assign. Moreover, the peaks arising from long range correlations in dipolar couplings produce low signal-to-noise ratio and the time required to acquire a 2D spectra is several days [44, 45].

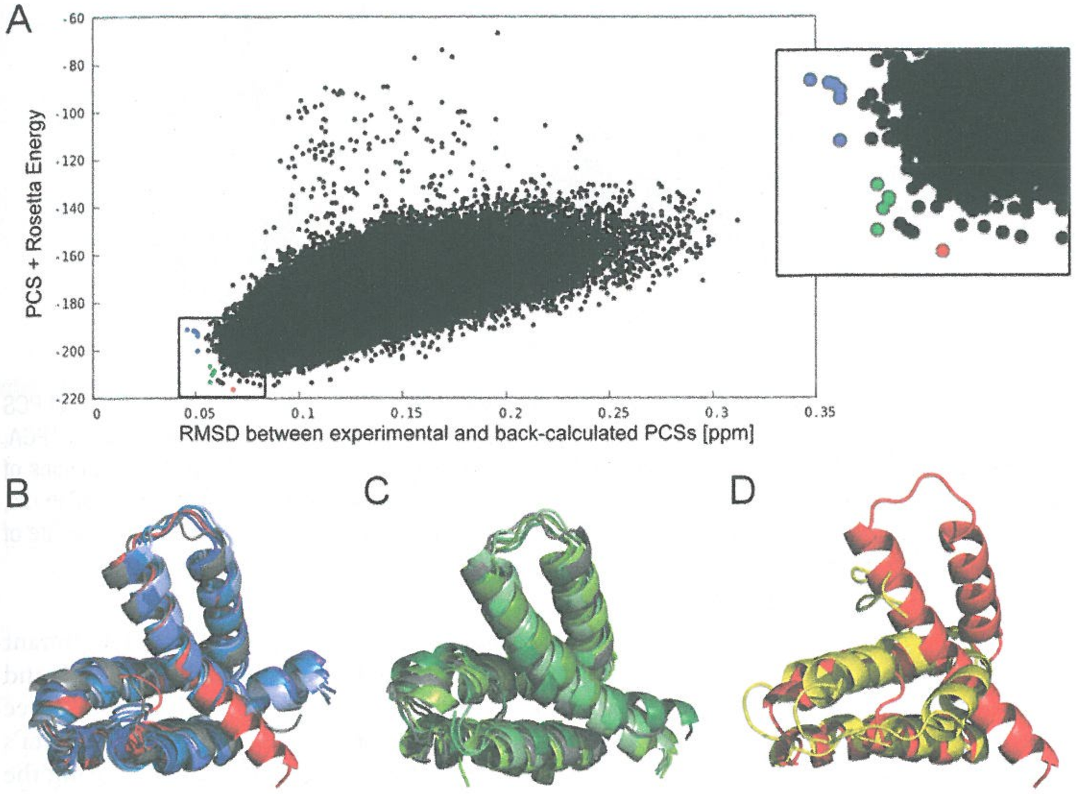

Fig. 6 Structure determination using GPS-Rosetta protocol for ERp29-C. (**a**) Combined score of weighted PCS and Rosetta energy is plotted against the PCS RMSD for each of the 100,000 generated structures. The final selected structure is represented in *red* has the lowest combined score. Structures with lowest PCS RMSD are represented in *blue* and the models with an arbitrary low combined score and low PCS RMSD are represented in *green*. (**b**) Superimposed cartoon representations of top structures selected using the GPS-Rosetta protocol. The crystal structure [PDBID: 2QC7] is shown in grey and the GPS-Rosetta structure is represented in red has 2.4 Å Cα RMSD to the crystal structure (residues 158–228 and 230–244). Top five models with low PCS RMSD represented in shades of blue have a Cα RMSD range of 2.0–2.9 Å to the crystal structure (residues 158–228 and 230–244). (**c**) Top five models with low PCS and Rosetta energy and also low in PCS RMSD are represented in shades of *green* have a Cα RMSD range of 2.2–2.6 Å to the crystal structure (residues 158–228 and 230–244). (**d**) The NOE structure [PDBID: 1G7D] represented in *yellow* has Cα RMSD of 6 Å to the GPS-Rosetta structure (residues 158–244) represented in *red*

Here we demonstrate the implementation of PCSs recorded in solid state for structure calculation. GB1 protein (56 amino acids) served as a model system. GB1 was covalently ligated to 4-mercaptomethyl-dipicolinic acid (4MMDPA) tag [46] at three different sites by generating three cysteine mutants at K28C, D40C, and E42C. The tags were loaded with paramagnetic metal ions Co^{2+}, Yb^{3+}, and Tm^{3+}, while Zn^{2+} and Lu^{3+} served as diamagnetic references. A total of 244 PCSs were measured from five paramagnetic datasets [47]. GB1 being a small protein, a stripped down version of GPS-Rosetta protocol was employed. Three

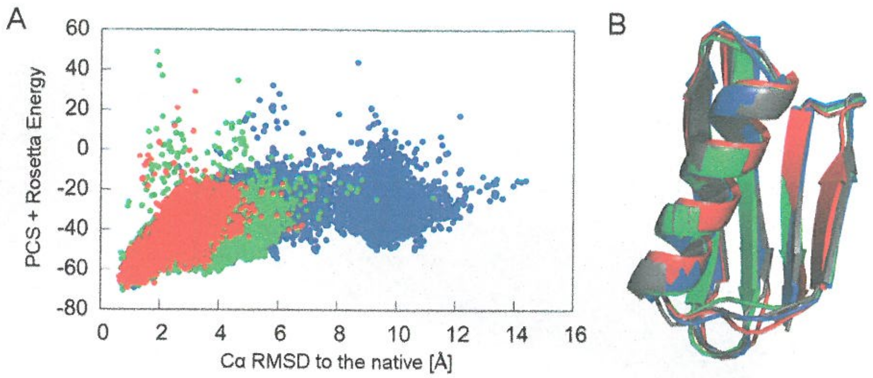

Fig. 7 Structure determination using GPS-Rosetta protocol with MAS-NMR PCSs. (**a**) Combined score of PCS energy from three tags and Rosetta energy versus the RMSD to the crystal structure of GB1 [PDBID:1PGA, [48]]. Sampling from K28C is represented in red, D40C in green and E42C in *blue*. (**b**) 3D superpositions of calculated models using GPS-Rosetta. The crystal structure of GB1 is represented in *grey*, mutant K28C in *red*, D40C in *green*, and E42C in *blue*. The three lowest scored structures have an RMSD to the crystal structure of 0.9, 0.7, and 1.1 Å respectively

independent Rosetta simulations were carried out for each mutant with nonhomologous fragment libraries. Around 4500, 8400, and 10,000 all-atom models were generated for each of the three mutants. To take advantage of all three datasets for GB1, Rosetta's all-atom structures for each of the mutants were rescored using the GPS-Rosetta protocol and the final structures were selected based on low Rosetta energy and combined low PCS score from all three datasets (Fig. 7a). The lowest combined energy structure was found to have RMSD of 0.7 Å when superimposed over the crystal structure (Fig. 7b) [PDBID: 1PGA, [48]], at atomic resolution.

The GPS-Rosetta protocol along with demonstration tutorials is available for download with the current Rosetta release.

5 Conclusion

Here GPS-Rosetta protocol's success in determining 3D structures using PCS data from multiple tags from both solution and solid-state NMR experiments has been demonstrated. This method offers great promise in resolving structures of large proteins. PCSs are obtained from simple ^{15}N-HSQC measurements which are highly accurate and sensitive compared to traditional NOE measurements and versatile PCS datasets can be generated by swapping a diverse range of available paramagnetic metals, metal carrying tags, and peptide sequences.

In computational modeling, incorporation of PCS data as structural restraints has enabled the computationally intractable conformational space to be explored in finite time. Inaccuracies in molecular force fields always posed a challenge in identifying native protein fold from well-formed structural decoys and PCSs being long range in nature effectively discriminates the native from non-native folds. PCSs can be also complemented with other sparse restraints such as RDCs, PREs, and NOEs, enhancing the structural information which can be efficiently exploited in computational modeling. In conclusion, the hybrid approach of incorporating experimental PCSs with structure determination algorithms forms a more efficient alternative approach to solve protein structures than traditional methods.

References

1. Wuthrich K (1986) NMR of proteins and nucleic acids. The George Fisher Baker Non-resident Lectureship in Chemistry at Cornell University
2. Andreini C, Bertini I, Rosato A (2009) Metalloproteomes: a bioinformatic approach. Acc Chem Res 42:1471–1479
3. Otting G (2010) Protein NMR using paramagnetic ions. Annu Rev Biophys 39:387–405
4. Su X-C, McAndrew K, Huber T, Otting G (2008) Lanthanide-binding peptides for NMR measurements of residual dipolar couplings and paramagnetic effects from multiple angles. J Am Chem Soc 130:1681–1687
5. Loh CT, Ozawa K, Tuck KL, Barlow N, Huber T, Otting G, Graham B (2013) Lanthanide tags for site-specific ligation to an unnatural amino acid and generation of pseudocontact shifts in proteins. Bioconjug Chem 24:260–268
6. Rodriguez-Castañeda, F., Haberz, P., Leonov, A., Griesinger, C. (2006) Paramagnetic tagging of diamagnetic proteins for solution NMR. *Magn Reson Chem* 44 Spec No, S10–S16.
7. Su X-C, Otting G (2010) Paramagnetic labelling of proteins and oligonucleotides for NMR. J Biomol NMR 46:101–112
8. Koehler J, Meiler J (2011) Expanding the utility of NMR restraints with paramagnetic compounds: background and practical aspects. Prog Nucl Magn Reson Spectrosc 59:360–389
9. Liu W-M, Overhand M, Ubbink M (2014) The application of paramagnetic lanthanoid ions in NMR spectroscopy on proteins. Coord Chem Rev 273–274:2–12
10. Bertini I, Luchinat C, Parigi G (2002) Magnetic susceptibility in paramagnetic NMR. Prog Nucl Magn Reson Spectrosc 40:249–273
11. Bertini I, Luchinat C, Parigi G, Pierattelli R (2008) Perspectives in paramagnetic NMR of metalloproteins. Dalton Trans 29:3782–3790
12. Iwahara J, Schwieters CD, Clore GM (2004) Ensemble approach for NMR structure refinement against (1)H paramagnetic relaxation enhancement data arising from a flexible paramagnetic group attached to a macromolecule. J Am Chem Soc 126:5879–5896
13. Keizers PHJ, Mersinli B, Reinle W, Donauer J, Hiruma Y, Hannemann F, Overhand M, Bernhardt R, Ubbink M (2010) A solution model of the complex formed by adrenodoxin and adrenodoxin reductase determined by paramagnetic NMR spectroscopy. Biochemistry 49:6846–6855
14. Schmitz C, Stanton-Cook M, Su X-C, Otting G, Huber T (2008) Numbat: an interactive software tool for fitting Deltachi-tensors to molecular coordinates using pseudocontact shifts. J Biomol NMR 41:179–189
15. John M, Schmitz C, Park AY, Dixon NE, Huber T, Otting G (2007) Sequence-specific and stereospecific assignment of methyl groups using paramagnetic lanthanides. J Am Chem Soc 129:13749–13757
16. Schmitz C, John M, Park AY, Dixon NE, Otting G, Pintacuda G, Huber T (2006) Efficient chi-tensor determination and NH assignment of paramagnetic proteins. J Biomol NMR 35:79–87
17. Skinner SP, Moshev M, Hass MAS, Keizers PHJ, Ubbink M (2013) PARAssign—paramagnetic NMR assignments of protein nuclei on the basis of pseudocontact shifts. J Biomol NMR 55:379–389
18. John M, Pintacuda G, Park AY, Dixon NE, Otting G (2006) Structure determination of protein-ligand complexes by transferred paramagnetic shifts. J Am Chem Soc 128:12910–12916
19. Saio T, Ogura K, Shimizu K, Yokochi M, Burke TR, Inagaki F (2011) An NMR strategy for fragment-based ligand screening utilizing a paramagnetic lanthanide probe. J Biomol NMR 51:395–408

20. Guan J-Y, Keizers PHJ, Liu W-M, Loehr F, Skinner SP, Heeneman EA, Schwalbe H, Ubbink M, Siegal GD, Löhr F, Skinner SP, Heeneman EA, Schwalbe H, Ubbink M, Siegal GD (2013) Small molecule binding sites on proteins established by paramagnetic NMR spectroscopy. J Am Chem Soc 135:5859–5868

21. Pintacuda G, Park AY, Keniry MA, Dixon NE, Otting G (2006) Lanthanide labeling offers fast NMR approach to 3D structure determinations of protein-protein complexes. J Am Chem Soc 128:3696–3702

22. Hiruma Y, Gupta A, Kloosterman A, Olijve C, Olmez B, Hass MA, Ubbink M (2014) Hot-spot residues in the cytochrome P450cam-putidaredoxin binding interface. Chembiochem 15:80–86

23. Schmitz C, Bonvin AMJJ (2011) Protein-protein HADDocking using exclusively pseudocontact shifts. J Biomol NMR 50:263–266

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

25. 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

26. Song Y, Dimaio F, Wang RY-R, Kim D, Miles C, Brunette T, Thompson J, Baker D (2013) High-resolution comparative modeling with RosettaCM. Structure 21:1735–1742

27. Meiler J, Baker D (2003) Rapid protein fold determination using unassigned NMR data. Proc Natl Acad Sci U S A 100:15404–15409

28. Raman S, Lange OF, Rossi P, Tyka M, Wang X, Aramini JM, Liu G, Ramelot TA, Eletsky A, Szyperski T, Kennedy MA, Prestegard J, Montelione GT, Baker D (2010) NMR structure determination for larger proteins using backbone-only data. Science 327:1014–1018

29. Shen Y, Vernon R, Baker D, Bax A (2009) De novo protein structure generation from incomplete chemical shift assignments. J Biomol NMR 43:63–78

30. 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 U S A 105:4685–4690

31. Lange OF, Baker D (2012) Resolution-adapted recombination of structural features significantly improves sampling in restraint-guided structure calculation. Proteins Struct Funct Bioinforma 80:884–895

32. Baker D (2014) Centenary award and Sir Frederick Gowland Hopkins Memorial Lecture. Protein folding, structure prediction and design. Biochem Soc Trans 42:225–229

33. Rohl CA, Strauss CEM, Misura KMS, Baker D (2004) Protein structure prediction using Rosetta. Methods Enzymol 383:66–93

34. Das R, Baker D (2008) Macromolecular modeling with Rosetta. Annu Rev Biochem 77:363–382

35. 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

36. 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

37. Liepinsh E, Baryshev M, Sharipo A, Ingelman-Sundberg M, Otting G, Mkrtchian S (2001) Thioredoxin fold as homodimerization module in the putative chaperone ERp29: NMR structures of the domains and experimental model of the 51 kDa dimer. Structure 9:457–471

38. Barak NN, Neumann P, Sevvana M, Schutkowski M, Naumann K, Malesević M, Reichardt H, Fischer G, Stubbs MT, Ferrari DM (2009) Crystal structure and functional analysis of the protein disulfide isomerase-related protein ERp29. J Mol Biol 385:1630–1642

39. Graham B, Loh CT, Swarbrick JD, Ung P, Shin J, Yagi H, Jia X, Chhabra S, Barlow N, Pintacuda G, Huber T, Otting G (2011) DOTA-amide lanthanide tag for reliable generation of pseudocontact shifts in protein NMR spectra. Bioconjug Chem 22:2118–2125

40. Swarbrick JD, Ung P, Chhabra S, Graham B (2011) An iminodiacetic acid based lanthanide binding tag for paramagnetic exchange NMR spectroscopy. Angew Chem 123:4495–4498

41. Hong M, Zhang Y, Hu F (2012) Membrane protein structure and dynamics from NMR spectroscopy. Annu Rev Phys Chem 63:1–24

42. De Paepe G, Lewandowski JR, Loquet A, Bockmann A, Griffin RG, De Paëpe G, Lewandowski JR, Loquet A, Böckmann A, Griffin RG (2008) Proton assisted recoupling and protein structure determination. J Chem Phys 129:245101

43. Korukottu J, Schneider R, Vijayan V, Lange A, Pongs O, Becker S, Baldus M, Zweckstetter M (2008) High-resolution 3D structure

determination of kaliotoxin by solid-state NMR spectroscopy. PLoS One 3:e2359

44. Wasmer C, Lange A, Van Melckebeke H, Siemer AB, Riek R, Meier BH (2008) Amyloid fibrils of the HET-s(218-289) prion form a beta solenoid with a triangular hydrophobic core. Science 319:1523–1526

45. Loquet A, Lv G, Giller K, Becker S, Lange A (2011) 13C spin dilution for simplified and complete solid-state NMR resonance assignment of insoluble biological assemblies. J Am Chem Soc 133:4722–4725

46. Su X-C, Man B, Beeren S, Liang H, Simonsen S, Schmitz C, Huber T, Messerle BA, Otting G (2008) A dipicolinic acid tag for rigid lanthanide tagging of proteins and paramagnetic NMR spectroscopy. J Am Chem Soc 130:10486–10487

47. 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

48. Gallagher T, Alexander P, Bryan P, Gilliland GL (1994) Two crystal structures of the B1 immunoglobulin-binding domain of streptococcal protein G and comparison with NMR. Biochemistry 33:4721–4729

49. Schmitz C (2009) Computational study of proteins with paramagnetic NMR: automatic assignments of spectral resonances, determination of protein-protein and protein-ligand complexes, and structure determination of proteins. Ph.D. thesis, University of Queensland