

Pseudocontact Shift-Driven Iterative Resampling for 3D Structure Determinations of Large Proteins

Kala Bharath Pilla, Gottfried Otting and Thomas Huber

Research School of Chemistry, Australian National University, Canberra, ACT 2601, Australia

Correspondence to Thomas Huber: t.huber@anu.edu.au http://dx.doi.org/10.1016/j.jmb.2016.01.007 Edited by C. Kalodimos

Abstract

Pseudocontact shifts (PCSs) induced by paramagnetic lanthanides produce pronounced effects in nuclear magnetic resonance spectra, which are easily measured and which deliver valuable long-range structure restraints. Even sparse PCS data greatly enhance the success rate of 3D (*3-d*imensional) structure predictions of proteins by the modeling program Rosetta. The present work extends this approach to 3D structures of larger proteins, comprising more than 200 residues, which are difficult to model by Rosetta without additional experimental restraints. The new algorithm improves the fragment assembly method of Rosetta by utilizing PCSs generated from paramagnetic lanthanide ions attached at four different sites as the only experimental restraints. The sparse PCS data are utilized at multiple stages, to identify native-like local structures, to rank the best structural models and to rebuild the fragment libraries. The fragment libraries are refined iteratively until convergence. The PCS-driven iterative resampling algorithm is strictly data dependent and shown to generate accurate models for a benchmark set of eight different proteins, ranging from 100 to 220 residues, using solely PCSs of backbone amide protons.

© 2016 Elsevier Ltd. All rights reserved.

Introduction

The assembly of short peptide fragments is the most widely adopted approach for de novo 3D (3-dimensional) structure predictions of proteins. Biennial CASP experiments have shown that, although this approach is very powerful for small proteins, it suffers from low success rates for medium (>100 amino acid residues) to large proteins (>200 residues) [1]. The failure with large proteins can be attributed to the difficulty of sampling the very large conformational space associated with the search for the global minimum in a highdimensional energy function. To attain efficient sampling, different structure prediction methods resort to different resampling algorithms. The QUARK method iteratively reshuffles short to large fragments during fragment assembly [2]. The I-TAS-SER method adopts iterative template fragment assembly [3]. Rosetta incorporates multiple different iterative approaches such as resampling of β-strand pairings [4], resampling of local structures identified from initial sampling [5], identification of starting models

0022-2836/© 2016 Elsevier Ltd. All rights reserved.

with correct topology followed by iterative rebuilding and refinement of the local regions of the structure that diverged the most in the ensemble [6] and, more recently, resolution-adapted structural recombination (RASREC). RASREC is a special genetic algorithm that iteratively resamples supersecondary and secondary structural features [7].

While iterative resampling improves the conformational search, inclusion of sparse experimental restraints has a marked effect in guiding the conformational sampling, starting from an extended polypeptide chain, toward the native 3D protein structure [8]. RASREC performs reliably in 70% of the proteins with less than 100 residues by the inclusion of sparse backbone chemical shift information [9]. Significantly improved performance is achieved with the combination of sparse distance restraints from nuclear Overhauser effects (NOEs) and orientation restraints from residual dipolar couplings, allowing structure determination of proteins greater than 150 amino acids [10,11]. The RASREC approach has recently proven to be useful where traditional methods had limited success [12,13].

The RASREC algorithm is designed to identify native-like features from intermediate models, even in the absence of experimental restraints, and it neither takes explicit advantage of experimental structural information nor uses such information to select or identify specific structural features. In view of the powerful long-range structural information inherent in even sparse pseudocontact shift (PCS) datasets and the ease with which PCSs can be measured for large proteins, we developed a new iterative resampling method that relies on the structural information encoded by PCSs.

PCSs are induced by paramagnetic metal ions associated with anisotropic susceptibility (χ) tensors. They are measured as the difference in chemical shift between a sample containing a paramagnetic ion and the corresponding sample containing a diamagnetic metal. Lanthanide ions offer distinct advantages for PCS measurements [14] and, in some metalloproteins, can replace natural metal ions [15]. Much more generally, however, nonmetalloproteins can be engineered with single lanthanide binding sites, mostly by site-specific labeling with a synthetic lanthanide tag, enabling PCS measurements not only in solution [16,17] but also in the solid state [18]. The PCS of a nuclear spin (measured in ppm) arising from a paramagnetic metal center is given by:

$$\delta^{\text{PCS}} = \frac{1}{12\pi r^3} \left[\Delta \chi_{\text{ax}} (3\cos^2\theta - 1) + \frac{3}{2} \Delta \chi_{\text{rh}} (\sin^2\theta \cos^2\varphi) \right]$$
(1)

where r, θ and φ are the polar coordinates of the nuclear spin with respect to the principal axes of the χ tensor. $\Delta \chi_{ax}$ and $\Delta \chi_{rh}$ are the axial and rhombic components of the χ tensor [19] and a $\Delta \chi$ tensor can be defined as the χ tensor minus its average isotropic component. Equation (1) shows that PCSs are both orientation and distance dependent. The potentially large anisotropic magnetic susceptibility of lanthanides in combination with the relatively weak r^{-3} distance dependence makes it possible to observe PCSs over a distance range of up to 80 Å (40 Å from the metal center). The PCS of a nuclear spin therefore provides direct long-range information about the spin's location in the $\Delta \chi$ -tensor frame, so long as the location of the metal center and the Δx -tensor orientation with respect to the protein are known or can be determined by fitting to a subset of PCSs from spins with defined atom positions.

The long-range nature of PCSs makes them superbly suitable as experimental restraints for modeling protein folds. We have shown previously that the Rosetta fragment assembly method can be combined with PCSs to yield reliable 3D structure determinations of proteins with less than 150 residues, using PCSs generated from a single metal center [20]. Structure determinations of larger proteins, however, face three major limiting factors. Firstly, if the protein is larger than the range of sizeable PCSs, only parts of the protein will be structurally defined by the PCS restraints. Secondly, PCSs of spins close to the metal center experience strong paramagnetic relaxation enhancements. which broaden the nuclear magnetic resonance (NMR) signals beyond detection and result in missing data. Thirdly, PCS data produced by different paramagnetic lanthanides are strongly correlated if the chemical structure of the tag is unchanged and therefore add only limited amount of new information. In previous work, we overcame these restrictions by extending the use of PCS restraints from a single metal center to PCSs from multiple metal centers. $\Delta \chi$ tensors from multiple tags ensure complete coverage of the protein with PCSs and allow restraining the location of nuclear spins in 3D space in a manner analogous to the global positioning system (GPS). The implementation of this algorithm in Rosetta was termed "GPS-Rosetta" [21]. GPS-Rosetta has since been shown to be superior for 3D structure determinations of proteins compared with traditional NMR approaches both in solution [21] and in the solid state [22]. More recently, we have demonstrated that GPS-Rosetta can be used to discriminate between distinct conformational states based on sparse PCS data generated from four different metal centers in the dengue virus NS2B/ NS3 protease [23].

The GPS-Rosetta approach is in principle applicable for structure determinations of larger proteins, but the inherent sampling limitation in Rosetta makes it difficult to generate correct models for proteins over 150 amino acids [11]. Additional time constraints arise from computing the $\Delta \chi$ tensors needed to score the structures. In GPS-Rosetta, calculation of a Δx tensor involves a search for the best location of the metal ion on a cubic grid and the $\Delta \chi$ -tensor computation must be repeated for each fragment move during a Monte Carlo assembly, typically involving over 100,000 moves per structure [20]. This computational overhead slows down a GPS-Rosetta simulation with four different metal centers and PCSs from two different metal ions at each site approximately 10-fold when compared with an unrestrained Rosetta simulation.

To overcome sampling and time constraints, we developed a new iterative resampling algorithm, which depends only on sparse PCSs measured from multiple metal centers. With the use of these PCSs, the algorithm automatically identifies good intermediate structures, extracts local structural elements that agree with the experimental data and rebuilds new fragment libraries. By iteratively resampling and rebuilding new fragment libraries, we direct the conformational search to the energetically favorable minimum while generating no more than a few thousand sample structures. We benchmark our new "iterative GPS-Rosetta" algorithm on a larger, 218-residue, seven-transmembrane α-helical microbial integral membrane protein, phototactic receptor sensory rhodopsin II (pSRII) from *Natronomonas pharaonis*, where experimental PCSs were measured from four different metal centers [24]. Furthermore, we assess the performance of the iterative GPS-Rosetta algorithm on an additional set of seven proteins, which contain 100–200 residues and comprise different folds, including membrane-bound, α-helical, β-barrel and α/β topologies.

Results

Assessment of iterative GPS-Rosetta using the integral membrane protein pSRII

The iterative GPS-Rosetta algorithm was applied to pSRII generating 2000 models in each iteration except for the zeroth iteration, where 3000 models were sampled. The structures were assembled from three-residue and nine-residue fragment libraries, each containing 200 fragments for any given window along the amino acid sequence. The calculations took about 4000 CPU hours per iteration. Populating the libraries with fragments in agreement with the PCS data in an iterative manner dramatically enhanced the chances of finding the correct protein fold. The results are summarized in Fig. 1. The scatter plots (Fig. 1a and b) show how the combined Rosetta and PCS energy of the final models decreased with iterations, while the C^{α} RMSD relative to the crystal structure [25] in both centroid decoys and all-atom refined structures improved simultaneously.

The improvement in the local fragments by the PCS-based selection is particularly striking, showing a substantial enhancement in the selection of native-like fragments over successive iterations (Fig. 1c). The very first iteration alone (shown in blue) already produced much more native-like fragments than the standard fragment library, which is computed based on sequence information and chemical shift data (shown in black). As a result, the median RMSD of the structures sampled in the first iteration shifted by 8 Å from 13 Å in the zeroth iteration to 5 Å in the first iteration (Fig. 1d). The PCS energy converged in about six iterations, at which point 90% of the sampled structures were within 3.5 Å RMSD of the crystal structure. Although further iterations no longer reduced the PCS and Rosetta energies, the probability of generating structures with lower RMSD values continued to increase because of an increase in the number of PCS-identified fragments. For example, 97% of the structures sampled in the tenth iteration had an RMSD below 3.2 Å, compared with 90% in the sixth iteration (Fig. 1d).

In the last four iterations, the combined Rosetta and PCS energies ranged between -390 REU and -434 REU (Fig. 1b). This large spread can be attributed to the existence of multiple local minima and the high sensitivity of Rosetta's all-atom energy function to small structural changes. Interestingly, there is an almost linear correlation between PCS energy and C^{α} RMSD for both centroid and all-atom structures (Fig. 2a and b), suggesting that the PCS energy acts as a better selection filter than the Rosetta energy function. The structure with the lowest PCS energy in the converged sixth iteration had a C^{α} RMSD of 2.7 Å to the crystal structure and was chosen as the final representative structure of the calculation (Fig. 1e). The back-calculated PCSs correlated closely with the experimental PCSs for this structure (Fig. 2c-f), with a low guality-factor Q[26] of 0.09 and only 2.6% of the PCSs deviating by more than the error bound of 0.05 ppm. The axial and rhombic components of the Δx tensors (Table S1) are also resolved to similar magnitudes compared to the previously determined values [24,27].

Performance benchmark of iterative GPS-Rosetta algorithm

We benchmarked the performance of the iterative GPS-Rosetta algorithm on an additional set of seven proteins. Three targets that are considered difficult targets for *de novo* structure determination were chosen from the Protein Structure Initiative project [28]. The simulation setup for all proteins was identical the one employed for pSRII. For all seven targets B-H, the energy scatter plots, improvement in local fragment libraries and density plots and the similarity of the final calculated structure to the target structure all exhibited similar traits as observed for pSRII (Figs. S1–S7). Target B was the smallest of all of the targets and the energy converged within three iterations. For targets C and H, convergence took four iterations. In contrast, the PCS energy for targets D and G continued to drop until the tenth iteration. The structures with the lowest PCS energy after convergence or after the tenth iteration were chosen as the representative structure to assess the model quality. Table 1 summarizes the results for all benchmark proteins including pSRII. All have Q factors below 0.12, indicating excellent agreement of the experimental data with the structural model [26]. The RMSD to the reference structures was as low as 1.3 Å (target E). The highest RMSD (6.2 Å) was observed for target G. The high RMSD is, however, entirely due to differences in the structures of loop regions. Excluding the loop regions from the RMSD calculation lowers the value to 1.1 Å.

The fragment libraries rebuilt using PCSs had a marked effect on sampling. In all targets, every iteration sampled structures with lower RMSD values compared to the previous iteration as shown in Fig. 3.



Fig. 1. Results from PCS-driven iterative GPS-Rosetta applied to pSRII. (a) Scatter plot of structures sampled by GPS-Rosetta. The combined Rosetta centroid energy and PCS energy is plotted *versus* the C^{α} RMSD of the crystal structure (PDB ID 1H68 [25]). The results from the different iterations are colorcoded, with the zeroth iteration in black and the next 10 iterations in blue to red as shown in the color bar on the right. The same colorcoding is used throughout the manuscript. (b) Same as (a), but after all-atom refinement. (c) Improvement in the quality of fragments identified by overlapping $\Delta \chi$ tensors in the PCS-driven iterative scheme. The plot shows the RMSD calculated between each nine-residue fragment and its corresponding native fragment in the crystal structure. The zeroth iteration (black) used the standard fragment library of the Robetta server, while subsequent iterations took the PCSs into account. (d) Probability density plots illustrating how consecutive iterations shift the conformational sampling toward structures with lower C^{α} RMSD to the crystal structure. Vertical bars shown for the 0th, 1st and 7th iterations identify the respective medians. (e) Superimposition of the structure with the lowest PCS energy (green) with the crystal structure (gray).

The effect is very prominent in the first iteration, which highlights the capacity to identify native-like local structure using PCS datasets from multiple metal sites. In all targets, more than 60% of the structures sampled in the converged iteration had RMSD values below 5 Å to the native structure and more than 85% reached this value in the tenth iteration. For target E

after the tenth iteration, 99% of the structures had an RMSD below 1.8 Å relative to the native structure.

Finally, we also tested the iterative GPS-Rosetta protocol with fewer PCS datasets. As an example, we restricted the experimental data to PCSs from three metal centers and two metals per center in the targets C, E and F. Targets C and E performed



Fig. 2. PCS assessment of the PCS-driven iterative GPS-Rosetta calculation applied to pSRII. (a) Scatter plot of centroid structures sampled by GPS-Rosetta as in Fig. 1a, but showing the PCS energy only *versus* the C^{α} RMSD of the crystal structure. (b) Same as (a), but for the structures after all-atom refinement. (c) Correlation between experimental and back-calculated PCSs for the tag at position 56 in the representative structure determined with iterative GPS-Rosetta (Fig. 1e). The data from the four different metals are represented in gray (Dy³⁺), cyan (Tm³⁺), brown (Tb³⁺) and pink (Yb³⁺), respectively. (d–f) Same as (c), except for the mutants I121C, S154C and V169C, respectively.

similarly well as in the situation of four tags with four metals used in the benchmark set (Figs. S8 and S9), but the PCS-based identification of improved fragments failed for target F (Fig. S10). With an increase in the number of PCS data to four metals per center, the iterative GPS-Rosetta method again produced a successful result for target F (Fig. S11) but needed four additional iterations to converge. This indicates that the structure of target F is intrinsically more difficult to predict. Target F has a complex α/β topology consisting of one α -helix and seven β -strands that form two antiparallel β -sheets. In this case, availability of PCS datasets from four metals per center was clearly crucial to increase the coverage and selection of a larger number of native-like fragments. The algorithm requires PCSs

Targets	PDB ID	N _{res} ^a	C ^α RMSD ^b 0th iteration (Å)	C ^α RMSD (iteration) (Å)	C ^α RMSD ^c (ordered residues) (Å)	Q factor ^d	C ^α RMSD 10th iteration (Å)	Biological Magnetic Resonance Bank ID	Reference
A (pSRII)	1H68	218	3.6	2.7 (6)	2.4 (185)	0.09	2.7	16678	[25]
B (ERp29-C)	2M66	106	6.6	3.0 (3)	2.2 (90)	0.12	3.4	4920	[21]
C (OmpX)	2M06	148	6.1	3.3 (4)	2.5 (100)	0.10	3.3	4936	[44]
D (polyketide cyc-like protein)	2M47	157	4.7	3.5 (5)	2.1 (111)	0.09	3.9	18989	Unpublished
E (CAP protein)	1S0P	160	3.6	1.3 (10)	1.0 (136)	0.05	1.3	5393	[45]
F (LEA protein)	1YYC	167	19.8	3.7 (6)	3.0 (112)	0.13	3.4	6515	Unpublished
G (OprH)	2LHF	179	13.2	6.2 (10)	1.1 (92)	0.10	6.2	17842	[46]
H (human leukocyte function-associated antigen 1)	1DGQ	188	6.0	3.5 (4)	3.1 (123)	0.11	3.5	4553	[47]

Table 1. Benchmark performance of the iterative GPS-Rosetta protocol.

^a Number of amino acid residues.

^b The C^a RMSD was calculated between the structure with the lowest PCS energy and the corresponding reference structure determined by X-ray crystallography or NMR.

^c The C^{α} RMSD was calculated as in the previous column, but including only ordered residues, identified by the STRIDE secondary structure assignment program [48].

^d The *Q* factor was calculated as the RMSD between experimental and back-calculated PCSs divided by the root mean square of the experimental PCSs.

from at least two different metal centers to identify an improved fragment and availability of PCSs from four metal centers, as used in the benchmark set, increases the coverage and allows the algorithm to select from six different pair-wise combinations, whereas three metal centers allow only three combinations. around the structural information encoded in PCS data. PCS data from multiple tags have major advantages for structure determination, as they can pinpoint the location of atoms in space. PCSs recorded for a nuclear spin from two or more metal centers restrict the location of the spin to the intersection of the isosurfaces defined by two or more $\Delta \chi$ tensors. This approach of using lanthanide tags in a manner analogous to GPS satellites has previously been shown to identify the global fold of a protein with high accuracy [21,22,29] and to discriminate between different conformational states [23,30]. Here we extended this concept by taking

Discussion

The success of the iterative GPS-Rosetta approach lies in building the computational algorithm



Fig. 3. Cumulative probability density plots of all the eight proteins in the benchmark set (a–h). The targets are labeled as in Table 1.

advantage of the restraint information associated with overlapping PCS isosurfaces to populate fragment libraries with native-like local structural elements. Reliable identification of local structure greatly boosts the performance of fragment assembly-based algorithms, which hinge on the assumption that the global fold of the protein is dictated by the local structure adopted by any given amino acid sequence [31]. Enriching the fragment library with local fragments of correct structure very much reduces the amount of conformational sampling, which is critically important for large proteins. In the case of target B, the iterative GPS-Rosetta protocol only took 7000 structures to achieve convergence compared to the standard GPS-Rosetta protocol, which is required to generate over 100,000 models [21]. In the present work, up to 23,000 structures were sampled per target.

A most advantageous feature of our PCS-based fragment selection is the identification of not only ordered secondary structure elements but also loop regions, which is manifested as a drop in overall energy with successive iterations in either centroid or all-atom modes. The largest effect is seen in a clear and distinct drop in energy in the very first iteration. Our PCS-driven resampling technique is in stark contrast to RASREC, which attempts to rebuild fragments by systematically biasing toward generalized structural features of known proteins [7].

Using PCSs from only two rather than three or more metal centers results in a lesser quality of the selected fragments and often leads to the inclusion of fragments with non-native conformation. This brings about higher RMSD values of the fragments selected for the first iteration (Fig. 1c and Figs. S1– S7). Nonetheless, less precise fragments tend to be quickly removed in subsequent fragment assembly stages and the accumulation of correct fragments in later iterations is reflected in lower RMSD values.

The number of iterations required for the PCS energy to converge varies between different targets. This is expected, as the protein topology and the quality of fragments present in the fragment libraries differ for different proteins. The eight different proteins chosen in the present study represent different fold families and native and homologous fragments were explicitly excluded from the fragment libraries to avoid any bias that could have enhanced convergence. Much greater convergence rates can probably be achieved, if structures of homologous proteins are available to populate the initial fragment library.

In this work, the convergence criterion and selection of the best structural elements were based on PCS energy only. The Rosetta all-atom score was not used for three reasons: (i) the PCS scores correlated better with fragment structural similarity than the Rosetta all-atom score. Rosetta all-atom energies are highly sensitive to small local structural variations, whereas the long-range effect of PCSs constitutes a more global measure of structural similarity. (ii) The PCS energy is a meaningful metric, as the PCS score directly indicates agreement with experiment. (iii) By not relying on the Rosetta built-in energy function, neither for fragment selection nor for judging convergence, it is straightforward to implement our approach with any other experimental parameter imbued with structural information.

Membrane-bound proteins constitute nearly 30% of the human genome [32], many of which are potential drug targets [33]. Three of the proteins in the benchmark set are membrane bound; pSRII (target A) has an α -helical topology, while OmpX (target C) and OprH (target G) form β -barrels. Novel methodologies in solution and solid-state NMR have advanced the field of membrane protein structure determination [34]. Nonetheless, it is still difficult to measure a large number of NOEs in a suitable membrane mimetic environment. In contrast, PCSs can be measured with high sensitivity in simple 2D (*2-d*imensional) NMR experiments and their long-range nature offers an excellent experimental underpinning of the final structural model.

The 3D structure of target A (pSRII) has previously been solved by two different approaches based on sparse NMR restraints. The first approach used RASREC Rosetta [7,10] with NOE restraints generated using perdeuterated samples in combination with ¹³C-methyl labeling of the amino acids isoleucine. leucine, valine, alanine, methionine and threonine [35]. The results of this approach [10] were very similar to the structure obtained by the iterative GPS-Rosetta protocol. The second and more recent approach utilized a combination of NMR-derived restraints including PCSs [24]. The PCSs were obtained from four different metal centers with fixed Δx-tensor parameters, sparse NOEs were generated using ILVA (isoleucine, leucine, valine and alanine) labeled deuterated samples, backbone dihedral angles were predicted using TALOS [36] and hydrogen-bond networks were predicted from slow exchange observed for amide protons in solvent accessibility experiments in combination with secondary structure analysis using chemical shift information. Using the combined restraints in Xplor-NIH [37,38] generated a structure with 2.6 Å RMSD to the reference structure, whereas using only PCS data produced a structure with 5.0 Å RMSD [24]. Remarkably, using the PCS data from the same study, our iterative GPS-Rosetta protocol produced a quite similar result (2.7 Å RMSD) without using any other restraints and without making any assumptions about any of the $\Delta \chi$ -tensor parameters, instead optimizing them dynamically during fragment assembly.

Methods

The iterative GPS-Rosetta algorithm

The iterative GPS-Rosetta protocol is divided into two stages (Fig. 4). The first stage generates a small number (e.g., 3000) of structural decoys. The second stage rebuilds new fragments guided by PCSs. The two stages are iterated until the PCS energy has converged or until a maximum number of iterations is reached. Convergence in sampling is considered to be attained if there is no further decrease in the PCS energy for the lowest-energy structure compared to the lowest-energy structure in the previous iteration.

Stage 1: GPS-Rosetta sampling

The Rosetta fragment assembly protocol employs Metropolis Monte Carlo assembly of nine-residue and three-residue fragments, which are generated using sequence and backbone diamagnetic chemical shift information of the target protein [39,40]. PCS scores for each of the different metal centers are weighted relative to Rosetta's centroid scoring function. The weighting factors (w) for each of the metal centers used to score the PCSs



Fig. 4. Flowchart of the iterative GPS-Rosetta protocol.

relative to the Rosetta scoring function are calculated by generating 1000 structures without PCS restraints. The weighting factors are then calculated for each of the n metal centers independently using

$$w = \left(\frac{a_{\text{high}} - a_{\text{low}}}{c_{\text{high}} - c_{\text{low}}}\right) / n \tag{2}$$

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 the PCS scores obtained by rescoring 1000 decoys with a unity PCS weighting factor.

All $\Delta \chi$ tensors for the individual metal centers are optimized simultaneously during the folding simulation in Rosetta. For each metal center, all of the eight parameters, as defined in Eq. (1), are fitted and the fit quality is scored as

$$s_{\rm k} = Rc \sum_{q=1}^{m} \sqrt{\sum_{p=1}^{n_{\rm PCS}} \left({\rm PCS}_{\rm calc}^{pq} - {\rm PCS}_{\rm exp}^{pq} \right)^2} \tag{3}$$

where *m* is the number of PCS datasets (one dataset per metal ion) and n_{PCS} is the number of PCSs in the dataset. *Rc* is a unity constant in units of $\frac{REU}{ppm}$ to convert PCS root-mean-square deviations into Rosetta energy units (REU). The total PCS energy (E_{PCS}) is given by:

$$E_{\rm PCS} = \sum_{k=1}^{n \rm tag} s_k \tag{4}$$

For the Rosetta centroid fragment assembly phase, PCS fit quality scores for each of the metal centers are independently weighted and the total weighted sum score, S_{total} , is added to the low-resolution centroid energy function of Rosetta:

$$S_{\text{total}} = \sum_{k=1}^{n \text{tag}} s_k . w_k \tag{5}$$

In the zeroth iteration, which uses the standard fragment libraries from the Robetta server [41], 3000 structures are generated. These structures are ranked according to their combined PCS energy [Eq. (4)] from all of the metal centers, and the top 200 structures are selected and refined as full-atom models using Rosetta's *Relax* protocol. For each of these top 200 structures, five different *Relax* simulations are performed, generating 1000 structures. These structures are again ranked according to their total PCS energy [using Eq. (4)] and the top 100 structures are used to build new fragment libraries.

Stage 2: Identification of new fragments based on PCS

Each of the top 100 structures generated in stage 1 is scanned, in overlapping nine-residue windows, for regions that strictly satisfy two conditions: Firstly, a nine-residue window must contain at least four PCSs per metal ion. Secondly, PCSs from at least two different metal centers must be within the error margin (e.g., ± 0.05 ppm) of the experimental value. The windows that fail to comply are discarded. A new fragment library is then generated and populated in a ratio of at least 12% new *versus* old fragments. At any given iteration, new fragments selected from the top 100 structures can populate at most 50% of the fragment library (which, by default, comprises 200 fragments) so that 50% of the original fragments are always retained. New sampling is then performed as described for stage 1 except that fewer structures, 2000 models per iteration, are generated.

The algorithm is designed to run on a computer cluster and is automated. The user can modify the individual steps in the algorithm if needed^{\dagger}. The algorithm requires the Rosetta software suite^{\ddagger}.

PCS data

Experimental PCS data

Currently, there are only two proteins with published PCS datasets that have been measured from at least four different metal centers: pSRII, which is a seventransmembrane α -helical integral membrane protein containing 218 residues [24,35], and the C-terminal domain of the endoplasmic reticulum protein 29 (ERp29-C), which contains 106 residues. ERp29-C was previously used to demonstrate the GPS-Rosetta protocol [21].

In this study, pSRII was used to demonstrate the PCS-driven iterative GPS-Rosetta algorithm. The PCSs for this protein were obtained using C2 lanthanide tags [29,27] ligated to the four different cysteine mutants L56C, I121C, S154C and V169C. Residues 56 and 121 are in the extracellular loop regions of the membrane protein, S154 is on the cytosolic side and V169 is in the transmembrane region. A total of 737 PCSs have been measured with Dy³⁺, Tb³⁺ and Tm³⁺ in a membrane-mimicking micelle environment with an experimental error of 0.02 ppm, but only 66% of the residues have at least one measured PCS value [24].

In ERp29-C, 212 PCSs have been measured for Tb³⁺ and Tm³⁺ at four different sites [21], using IDA-SH tags [42] ligated to the mutants C157S/S200C/K204D, C157S/A218C/A222D and C157S/Q241C/N245D, as well as the C1 tag [27] ligated to the wild-type protein.

Simulated PCS data

For other benchmark proteins devoid of experimental PCS data, datasets were generated mimicking real experimental conditions by computationally grafting the coordinates of the C2 tag [29,27] onto the target structure at four randomly chosen solvent-exposed residues. For each site, a rotamer library was generated for the tag to sample all physically possible 3D conformations of the C2 tag without steric clashes to the protein and a single rotamer was picked randomly to define the coordinates of metal position of the $\Delta \chi$ tensor. Euler angles, which determine the orientation of the $\Delta \chi$ -tensor frame relative to the protein frame, were also chosen randomly. PCS data were generated for Dy^{3+} , Tb^{3+} , Tm^{3+} and Yb^{3+} , using the $\Delta \chi_{ax}$ and $\Delta \chi_{rh}$ values determined for the L56C mutant of pSRII [24] by fitting the experimental PCS data to the pSRII iterative GPS-Rosetta model. PCS data were generated only for the backbone amide protons using pyParaTools [43]. PCSs of spins within a 12-Å radius from the metal centers were excluded from the datasets to account for the

loss of signal due to the paramagnetic relaxation enhancement effect. A random error of ± 0.04 ppm, which is twice the standard deviation found in the fits of experimental PCSs for pSRII, was added to all PCS data. To account for incomplete data, we randomly deleted PCSs from each of the datasets until the total coverage was 66%. In total, the four metal centers, each carrying four different lanthanide metals, resulted in sixteen datasets.

Starting fragment library

The Robetta server [41] was used to create the initial three- and nine-residue fragment libraries, explicitly omitting homologous proteins in the fragment generation. Fragment selection was aided by ¹H, ¹⁵N and ¹³C diamagnetic chemical shifts of the backbone atoms, which were taken from the Biological Magnetic Resonance Bank (Table 1).

Conclusion

This work demonstrates that PCS-driven preselection of local fragments presents a practical route to the calculation of 3D protein structures of medium to large size. By iterative fragment sampling and rebuilding guided by PCSs from different metal centers, we generated near-native models for all of the eight different protein folds in the benchmark set. This procedure overcomes the prohibitively large amount of sampling required in other Rosetta fragment assembly methods that determine the structures of larger proteins with the help short-range restraints.

Acknowledgements

We thank Professor Daniel Nietlispach for the PCS data of pSRII. Financial support to T.H. and G.O. by the Australian Research Council is gratefully acknowledged. This research was undertaken with the assistance of resources from the National Computational Infrastructure, which is supported by the Australian Government.

Appendix A. Supplementary data

Supplementary data to this article can be found online at http://dx.doi.org/10.1016/j.jmb.2016.01.007.

Received 9 November 2015; Received in revised form 8 January 2016; Accepted 8 January 2016 Available online 15 January 2016

Keywords:

pseudocontact shifts; NMR spectroscopy; conformational sampling; membrane proteins; 3D structure determination †The scripts to implement the algorithm are available for download from https://github.com/kalabharath/ pcs_driven_iterative_resampling. ‡which is available for download from http://www. rosettacommons.org.

Abbreviations used:

PCS, pseudocontact shift; RASREC, resolution-adapted structural recombination; NOE, nuclear Overhauser effect; GPS, global positioning system; pSRII, phototactic receptor sensory rhodopsin II; ERp29-C, C-terminal

domain of the endoplasmic reticulum protein 29.

References

- [1] C.H. Tai, H. Bai, T.J. Taylor, B. Lee, Assessment of templatefree modeling in CASP10 and ROLL, Proteins: Struct., Funct., Bioinf. 82 (2014) 57–83, http://dx.doi.org/10.1002/ prot.24470.
- [2] D. Xu, Y. Zhang, *Ab initio* protein structure assembly using continuous structure fragments and optimized knowledgebased force field, Proteins 80 (2012) 1715–1735, http://dx. doi.org/10.1002/prot.24065.
- [3] A. Roy, A. Kucukural, Y. Zhang, I-TASSER: A unified platform for automated protein structure and function prediction, Nat. Protoc. 5 (2010) 725–738, http://dx.doi.org/ 10.1038/nprot.2010.5.
- [4] P. Bradley, D. Baker, Improved beta-protein structure prediction by multilevel optimization of nonlocal strand pairings and local backbone conformation, Proteins 65 (2006) 922–929, http://dx.doi.org/10.1002/prot.21133.
- [5] B. Blum, M.I. Jordan, D. Baker, Feature space resampling for protein conformational search, Proteins: Struct., Funct., Bioinf. 78 (2010) 1583–1593, http://dx.doi.org/10.1002/prot. 22677.
- [6] B. Qian, S. Raman, R. Das, P. Bradley, A.J. McCoy, R.J. Read, et al., High-resolution structure prediction and the crystallographic phase problem, Nature 450 (2007) 259–264, http://dx.doi.org/10.1038/nature06249.
- [7] O.F. Lange, D. Baker, Resolution-adapted recombination of structural features significantly improves sampling in restraint-guided structure calculation, Proteins: Struct., Funct., Bioinf. 80 (2012) 884–895, http://dx.doi.org/10.1002/ prot.23245.
- [8] S.J. Fleishman, D. Baker, Role of the biomolecular energy gap in protein design, structure, and evolution, Cell 149 (2012) 262–273, http://dx.doi.org/10.1016/j.cell.2012.03.016.
- [9] G. van der Schot, Z. Zhang, R. Vernon, Y. Shen, W.F. Vranken, D. Baker, et al., Improving 3D structure prediction from chemical shift data, J. Biomol. NMR 57 (2013) 27–35, http://dx.doi.org/10.1007/s10858-013-9762-6.
- [10] O.F. Lange, P. Rossi, N.G. Sgourakis, Y. Song, H.-W. Lee, J.M. Aramini, 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. 109 (2012) 10873–10878, http://dx.doi.org/10.1073/pnas. 1203013109.
- [11] S. Raman, O.F. Lange, P. Rossi, M. Tyka, X. Wang, J.M. Aramini, et al., NMR structure determination for larger proteins using backbone-only data, Science 327 (2010) 1014–1018, http://dx.doi.org/10.1126/science.1183649.

- [12] T. Rao, J.W. Lubin, G.S. Armstrong, T.M. Tucey, V. Lundblad, D.S. Wuttke, Structure of Est3 reveals a bimodal surface with differential roles in telomere replication, Proc. Natl. Acad. Sci. U.S.A. 111 (2014) 214–218, http://dx.doi.org/ 10.1073/pnas.1316453111.
- [13] N.G. Sgourakis, K. Natarajan, J. Ying, B. Vogeli, L.F. Boyd, D.H. Margulies, et al., The structure of mouse cytomegalovirus m04 protein obtained from sparse NMR data reveals a conserved fold of the m02-m06 viral immune modulator family, Structure 22 (2014) 1263–1273, http://dx.doi.org/10. 1016/j.str.2014.05.018.
- [14] G. Otting, Protein NMR using paramagnetic ions, Annu. Rev. Biophys. 39 (2010) 387–405, http://dx.doi.org/10.1146/ annurev.biophys.093008.131321.
- [15] I. Bertini, C. Luchinat, G. Parigi, R. Pierattelli, Perspectives in paramagnetic NMR of metalloproteins, Dalton Trans. (2008) 3782–3790, http://dx.doi.org/10.1039/b719526e.
- [16] W.-M. Liu, M. Overhand, M. Ubbink, The application of paramagnetic lanthanoid ions in NMR spectroscopy on proteins, Coord. Chem. Rev. 273–274 (2014) 2–12, http://dx. doi.org/10.1016/j.ccr.2013.10.018.
- [17] P.H.J. Keizers, M. Ubbink, Paramagnetic tagging for protein structure and dynamics analysis, Prog. Nucl. Magn. Reson. Spectrosc. 58 (2011) 88–96, http://dx.doi.org/10.1016/j.pnmrs. 2010.08.001.
- [18] C.P. Jaroniec, Structural studies of proteins by paramagnetic solid-state NMR spectroscopy, J. Magn. Reson. 253 (2015) 50–59, http://dx.doi.org/10.1016/j.jmr.2014.12.017.
- [19] I. Bertini, C. Luchinat, G. Parigi, Magnetic susceptibility in paramagnetic NMR, Prog. Nucl. Magn. Reson. Spectrosc. 40 (2002) 249–273, http://dx.doi.org/10.1016/S0079-6565(02)00002-X.
- [20] C. Schmitz, R. Vernon, G. Otting, D. Baker, T. Huber, Protein structure determination from pseudocontact shifts using ROSETTA, J. Mol. Biol. 416 (2012) 668–677, http://dx.doi. org/10.1016/j.jmb.2011.12.056.
- [21] H. Yagi, K.B. Pilla, A. Maleckis, B. Graham, T. Huber, G. Otting, Three-dimensional protein fold determination from backbone amide pseudocontact shifts generated by lanthanide tags at multiple sites, Structure 21 (2013) 883–890, http://dx.doi.org/10.1016/j.str.2013.04.001.
- [22] J. Li, K.B. Pilla, Q. Li, Z. Zhang, X. Su, T. Huber, et al., Magic angle spinning NMR structure determination of proteins from pseudocontact shifts, J. Am. Chem. Soc. 135 (2013) 8294–8303, http://dx.doi.org/10.1021/ja4021149.
- [23] K.B. Pilla, J.K. Leman, G. Otting, T. Huber, Capturing conformational states in proteins using sparse paramagnetic NMR data, PLoS One 10 (2015) e0127053, http://dx.doi.org/ 10.1371/journal.pone.0127053.
- [24] D.J. Crick, J.X. Wang, B. Graham, J.D. Swarbrick, H.R. Mott, D. Nietlispach, Integral membrane protein structure determination using pseudocontact shifts, J. Biomol. NMR 61 (2015) 1–9, http://dx.doi.org/10.1007/s10858-015-9899-6.
- [25] a Royant, P. Nollert, K. Edman, R. Neutze, E.M. Landau, E. Pebay-Peyroula, et al., X-ray structure of sensory rhodopsin II at 2.1-Å resolution, Proc. Natl. Acad. Sci. U.S.A. 98 (2001) 10131–10136, http://dx.doi.org/10.1073/pnas.181203898.
- [26] G. Cornilescu, J.L. Marquardt, M. Ottiger, A. Bax, Validation of protein structure from anisotropic carbonyl chemical shifts in a dilute liquid crystalline phase, J. Am. Chem. Soc. 120 (1998) 6836–6837, http://dx.doi.org/10.1021/ja9812610.
- [27] B. Graham, C.T. Loh, J.D. Swarbrick, P. Ung, J. Shin, H. Yagi, et al., DOTA-amide lanthanide tag for reliable generation of pseudocontact shifts in protein NMR spectra,

Bioconjug. Chem. 22 (2011) 2118–2125, http://dx.doi.org/10. 1021/bc200353c.

- [28] H.M. Berman, J.D. Westbrook, M.J. Gabanyi, W. Tao, R. Shah, A. Kouranov, et al., The protein structure initiative structural genomics knowledgebase, Nucleic Acids Res. 37 (2009) D365–D368, http://dx.doi.org/10.1093/nar/gkn790.
- [29] L. de la Cruz, T.H.D. Nguyen, K. Ozawa, J. Shin, B. Graham, T. Huber, et al., Binding of low molecular weight inhibitors promotes large conformational changes in the dengue virus NS2B-NS3 protease: Fold analysis by pseudocontact shifts, J. Am. Chem. Soc. 133 (2011) 19205–19215, http://dx.doi. org/10.1021/ja208435s.
- [30] S.P. Skinner, W.-M. Liu, Y. Hiruma, M. Timmer, A. Blok, M.A.S. Hass, et al., Delicate conformational balance of the redox enzyme cytochrome P450cam, Proc. Natl. Acad. Sci. (2015), http://dx.doi.org/10.1073/pnas.1502351112.
- [31] K.A. Dill, S.B. Ozkan, T.R. Weikl, J.D. Chodera, V.A. Voelz, The protein folding problem: When will it be solved? Curr. Opin. Struct. Biol. 17 (2007) 342–346, http://dx.doi.org/10. 1016/j.sbi.2007.06.001.
- [32] E. Wallin, G. von Heijne, Genome-wide analysis of integral membrane proteins from eubacterial, archaean, and eukaryotic organisms, Protein Sci. 7 (1998) 1029–1038, http://dx. doi.org/10.1002/pro.5560070420.
- [33] M.A. Yildirim, K.-I. Goh, M.E. Cusick, A.-L. Barabási, M. Vidal, Drug-target network, Nat. Biotechnol. 25 (2007) 1119–1126, http://dx.doi.org/10.1038/nbt1338.
- [34] R. Kaptein, G. Wagner, NMR studies of membrane proteins, J. Biomol. NMR 61 (2015) 181–184, http://dx.doi.org/10. 1007/s10858-015-9918-7.
- [35] A. Gautier, H.R. Mott, M.J. Bostock, J.P. Kirkpatrick, D. Nietlispach, Structure determination of the seven-helix transmembrane receptor sensory rhodopsin II by solution NMR spectroscopy, Nat. Struct. Mol. Biol. 17 (2010) 768–774, http://dx.doi.org/10.1038/nsmb.1807.
- [36] G. Cornilescu, F. Delaglio, A. Bax, Protein backbone angle restraints from searching a database for chemical shift and sequence homology, J. Biomol. NMR 13 (1999) 289–302, http://dx.doi.org/10.1023/A:1008392405740.
- [37] C.D. Schwieters, J.J. Kuszewski, N. Tjandra, G.M. Clore, The Xplor-NIH NMR molecular structure determination package, J. Magn. Reson. 160 (2003) 65–73, http://dx.doi.org/10.1016/ S1090-7807(02)00014-9.
- [38] L. Banci, I. Bertini, G. Cavallaro, A. Giachetti, C. Luchinat, G. Parigi, Paramagnetism-based restraints for Xplor-NIH, J. Biomol. NMR 28 (2004) 249–261, http://dx.doi.org/10.1023/ B:JNMR.0000013703.30623.f7.
- [39] C.A. Rohl, C.E.M. Strauss, K.M.S. Misura, D. Baker, Protein structure prediction using Rosetta, Methods Enzymol. 383 (2004) 66–93, http://dx.doi.org/10.1016/S0076-6879(04)83004-0.
- [40] P.M. Bowers, C.E.M. Strauss, D. Baker, *De novo* protein structure determination using sparse NMR data, J. Biomol. NMR 18 (2000) 311–318, http://dx.doi.org/10.1023/A: 1026744431105.
- [41] D.E. Kim, D. Chivian, D. Baker, Protein structure prediction and analysis using the Robetta server, Nucleic Acids Res. 32 (2004) W526–W531, http://dx.doi.org/10.1093/nar/gkh468.
- [42] J.D. Swarbrick, P. Ung, S. Chhabra, B. Graham, An iminodiacetic acid based lanthanide binding tag for paramagnetic exchange NMR spectroscopy, Angew. Chem. Int. Ed. Engl. 50 (2011) 4403–4406, http://dx.doi.org/10.1002/anie.201007221.
- [43] M. Stanton-Cook, X.-C. Su, G. Otting, T. Huber, pyParaTools v0.8-alpha, 2011, http://dx.doi.org/10.5281/zenodo.10313.