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

Kala Bharath Pilla[1], Gottfried Otting[1] and Thomas Huber[1]*

[1] Research School of Chemistry, Australian National University, Canberra, ACT 2601, Australia.

*Corresponding author

E-mail: t.huber@anu.edu.au

Abstract

Pseudocontact shifts (PCS) induced by paramagnetic lanthanides produce pronounced effects in nuclear magnetic resonance (NMR) spectra, which are easily measured and deliver valuable long-range structure restraints. Even sparse PCS data greatly enhance the success rate of 3D structure predictions of proteins by the modeling program Rosetta. The present work extends this approach to 3D structures of larger proteins, 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.

Introduction

The assembly of short peptide fragments is the most widely adopted approach for *de novo* 3D 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 high-dimensional 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-TASSER 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 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 super-secondary 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, towards 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 (NOE) and orientation restraints from residual dipolar couplings (RDC), 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, but it neither takes explicit advantage of experimental structural information nor does it use such information to select or identify specific structural features. In view of the powerful long-range structural information inherent in even

sparse PCS data sets 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 [15] and, in some metalloproteins, can replace natural metal ions [14]. Much more generally, however, non-metalloproteins 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:

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

where r, θ, 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 $\Delta\chi$-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 modelling 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 (PRE), which broaden the 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 $\Delta\chi$ 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 ten-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. Utilizing 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 (7 TM) 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.

Methods

The iterative GPS-Rosetta algorithm

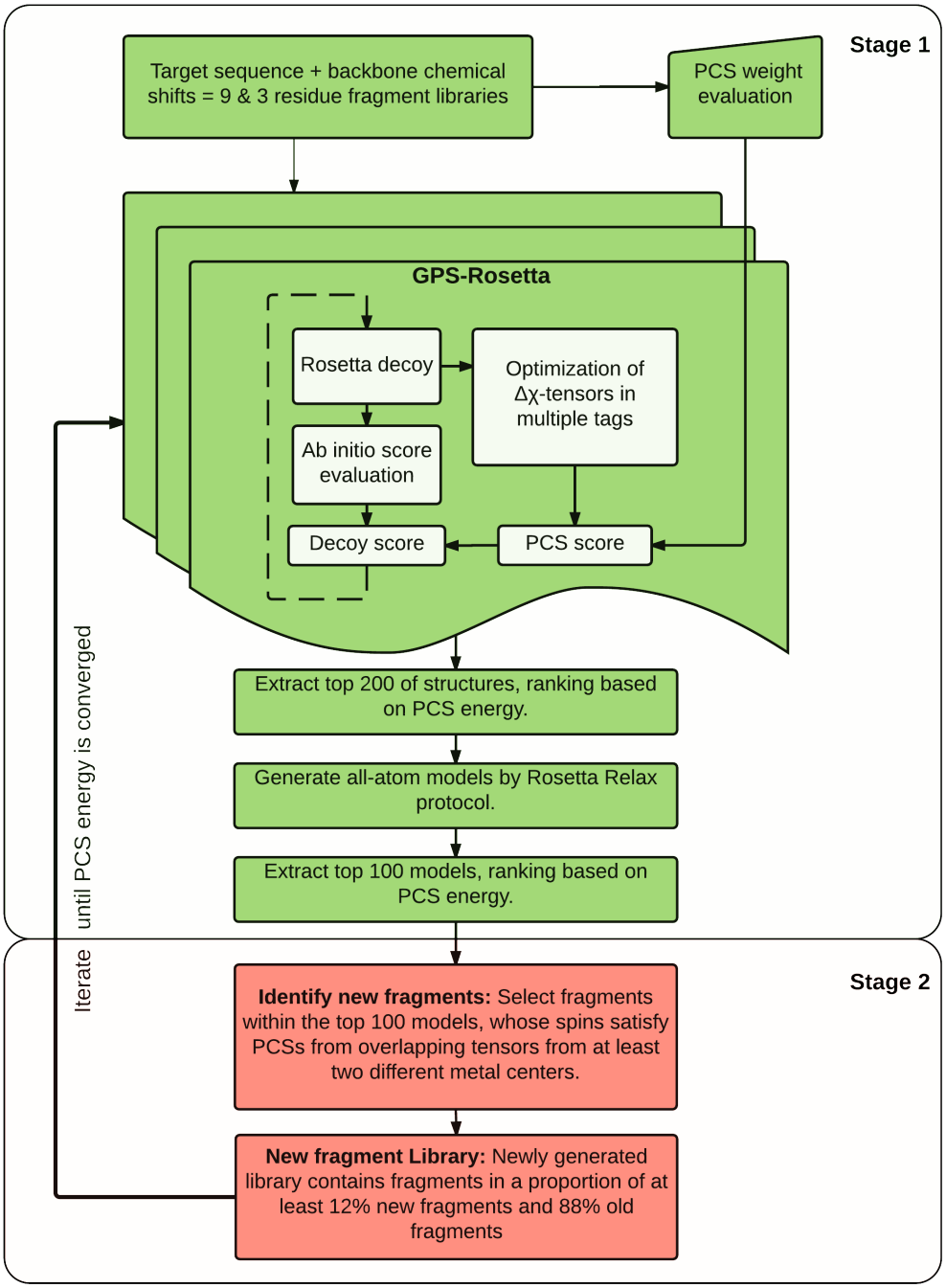

Figure 1. Flowchart of the iterative GPS-Rosetta protocol.

The iterative GPS-Rosetta protocol is divided into two stages. 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 a maximum number of iterations is reached.

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 [25,26]. 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 relative to the Rosetta's 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 by

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

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 score 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. The fit quality is scored as

$$s_k = Rc \sum_{q=1}^{m} \sqrt{\sum_{p=1}^{n_{pcs}} \left(PCS_{calc}^{pq} - PCS_{exp}^{pq} \right)^2} \qquad (3)$$

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 constant in units of $\frac{REU}{ppm}$ to convert PCS root-mean-square deviations to Rosetta energy units (REU). The total PCS energy (E_{PCS}) is given by:

$$E_{PCS} = \sum_{k=1}^{ntag} s_k \qquad (4)$$

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:

$$S_{total} = \sum_{k=1}^{ntag} s_k . w_k \qquad (5)$$

In the zeroth iteration, which uses the standard fragment libraries from the Robetta server [27], 3000 structures are generated. These structures are ranked according to their combined PCS energy (equation 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 equation 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 are 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.

It takes about 4000 CPU hours per iteration for a 200-residue protein. 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. The scripts to implement the algorithm are available for download from https://github.com/kalabharath/pcs_driven_iterative_resampling. The algorithm requires the Rosetta software suite, which is available for download from http://www.rosettacommons.org.

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 seven TM α-helical integral membrane protein containing 218 residues [24,28], 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 [29,38] 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 in the transmembrane region. 737 PCSs have been measured with Dy^{3+}, Tb^{3+} and Tm^{3+} 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 Tb^{3+} and Tm^{3+} at four different sites [21], using IDA-SH tags [30] ligated to the mutants C157S/S200C/K204D, C157S/A218C/A222D and C157S/Q241C/N245D, and the C1 tag [29] 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 [29,38] 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 [31]. 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 PRE 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, PCSs were randomly deleted 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

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 Figure 2. The scatter plots (Figure 2A and B) show how the combined Rosetta centroid and PCS energy of the final models improved the Cα RMSD relative to the crystal structure [32] both after calculation of the centroid decoys and after all-atom refinement.

The improvement in the local fragments by the PCS-based selection is particularly striking, showing a substantial improvement in the selection of native-like fragments over successive iterations (Figure 2C). 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 (Figure 2D). 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 RMSDs 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 (Figure 2D).

In the last four iterations, the combined Rosetta and PCS energies ranged between -390 REU and -434 REU (Figure 2B). 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 Cα RMSD for both centroid and all-atom structures (Figure 3A 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 Cα RMSD of 2.7 Å to the crystal structure and was chosen as the final representative structure of the calculation (Figure 2E). The back-calculated PCSs correlated closely with the experimental PCSs for this structure (Figure 3C-F), with a low quality factor *Q* [33] of 0.09 and only 2.6% of the PCSs deviating by more than the error bound of 0.05 ppm.

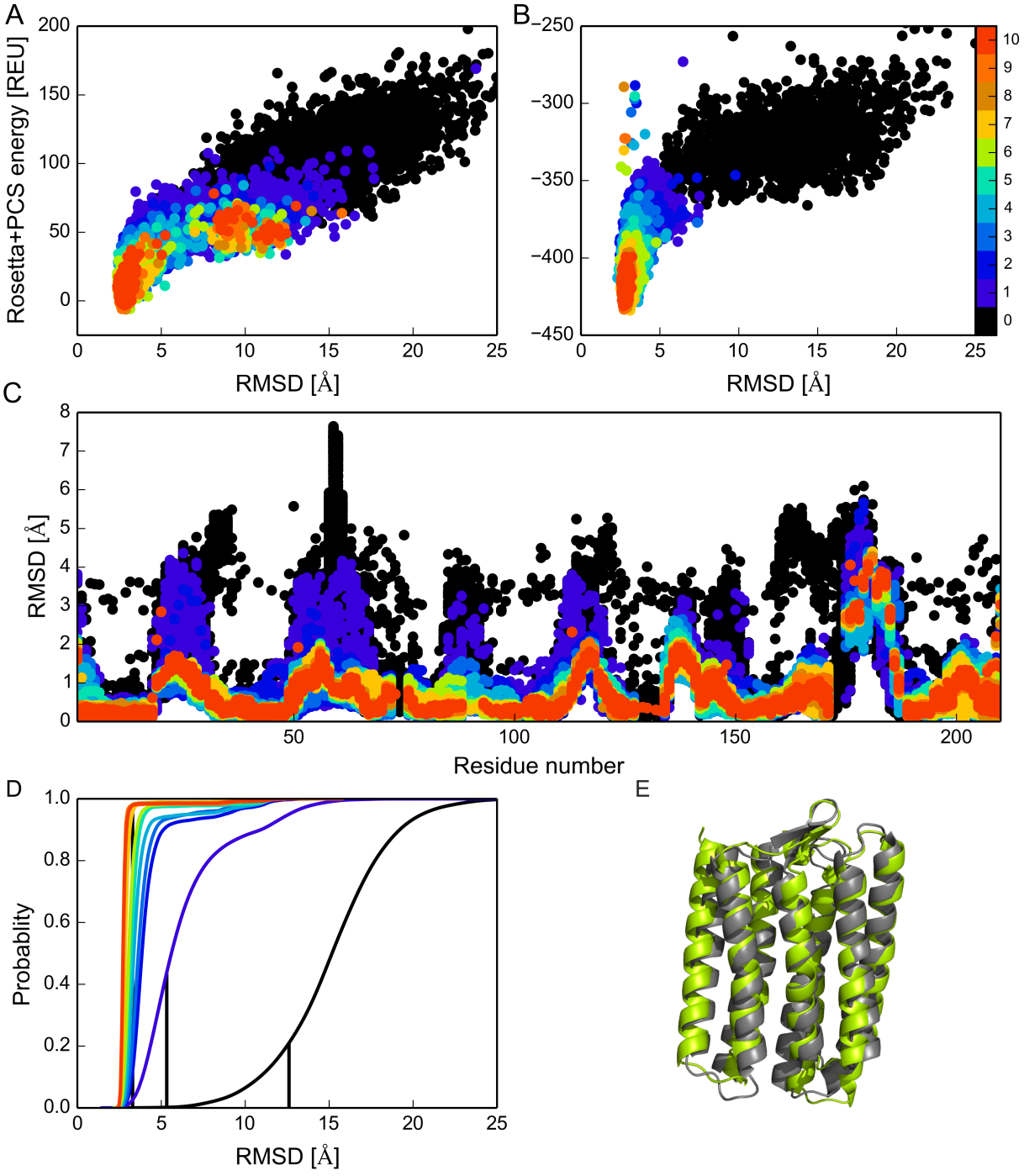

Figure 2. Results from PCS-driven iterative GPS-Rosetta applied to pSRII. (A) Scatter plot of structures sampled by GPS-Rosetta. The combined Rosetta centroid energy and PCS energy is plotted versus the Cα RMSD of the crystal structure (PDB ID 1H68 [reference]). The results from the different iterations are color-coded, with the zeroth iteration in black and the next ten iterations in blue to red as shown in the color bar on the right. The same color-coding 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 towards structures with lower Cα RMSD to the crystal structure. Vertical bars shown for the 0^{th}, 1^{st} and 7^{th} iterations identify the respective medians. (E) Superimposition of the structure with the lowest PCS energy (green) with the crystal structure (gray).

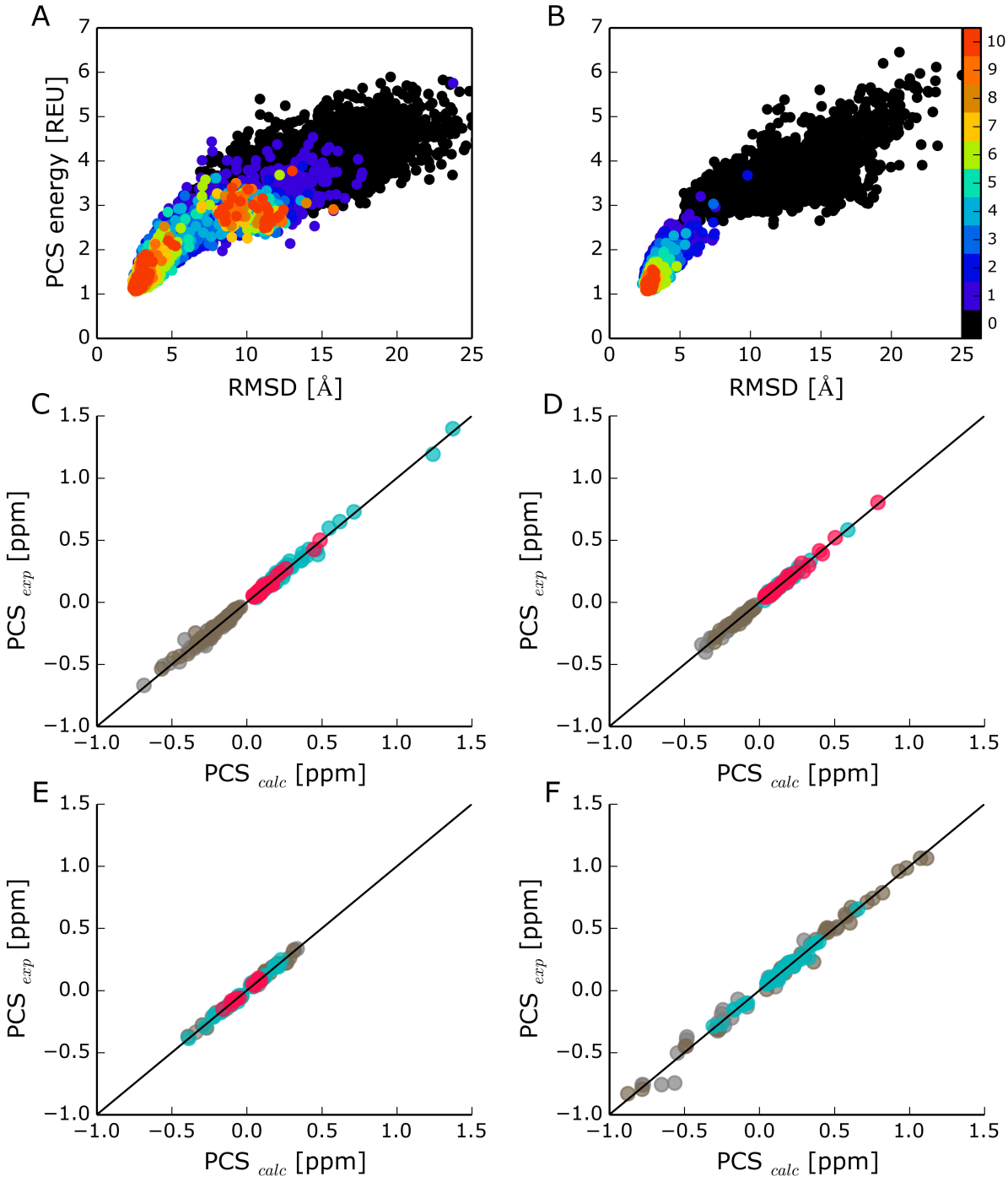

Figure 3. PCS assessment of the PCS-driven iterative GPS-Rosetta calculation applied to pSRII and (A) Scatter plot of centroid structures sampled by GPS-Rosetta as in Figure 2A, but showing the PCS energy only versus the Cα RMSD of the crystal structure. (B) Same as (A), but for the structures after 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 (Figure 2E). The data from the four different metals are represented in gray, cyan, brown,

and pink, respectively. (D)-(F) Same as (C), except for the mutants I121C, S154C and V169C, respectively.

Performance benchmark of iterative GPS-Rosetta algorithm

We benchmarked the performance of the iterative GPS-Rosetta algorithm on an additional set of seven proteins. The simulation setup for all proteins was identical to 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 (Figures 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 [33]. 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.

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

Targets	PDB ID	N_{res}[a]	Cα RMSD[b] (iteration)	Cα RMSD[c] (ordered residues)	*Q*-factor[d]	Cα RMSD (10^{th} iteration)	BMRB ID	Reference
A (pSRII)	1H68	218	2.7 Å (6)	2.4 Å (185)	0.09	2.7 Å	16678	[32]
B (ERp29-C)	2M66	106	3.0 Å (3)	2.2 Å (90)	0.12	3.4 Å	4920	[21]
C (OmpX)	2M06	148	3.3 Å (4)	2.5 Å (100)	0.10	3.3 Å	4936	[34]
D (Polyketide cyc-like protein)	2M47	157	3.5 Å (5)	2.1 Å (111)	0.09	3.9 Å	18989	unpublished

E (CAP protein)	1S0P	160	1.3 Å (10)	1.0 Å (136)	0.05	1.3 Å	5393	[35]
F (LEA protein)	1YYC	167	3.7 Å (6)	3.0 Å (112)	0.13	3.4 Å	6515	unpublished
G (OprH)	Auftrag static LHF	179	6.2 Å (10)	1.1 Å (92)	0.10	6.2 Å	17842	[36]
H (Human leukocyte function associated antigen-1)	1DGQ	188	3.5 Å (4)	3.1 Å (123)	0.11	3.5 Å	4553	[37]

[a] Number of amino acid residues; [b] the Cα 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 Cα RMSD was calculated as in the previous column, but including only ordered residues in the calculation; [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.

The fragment libraries rebuilt using PCSs had a marked effect on sampling. In all targets, every iteration sampled structures with lower RMSDs compared to the previous iteration as shown in Figure 4. 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 RMSDs 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.

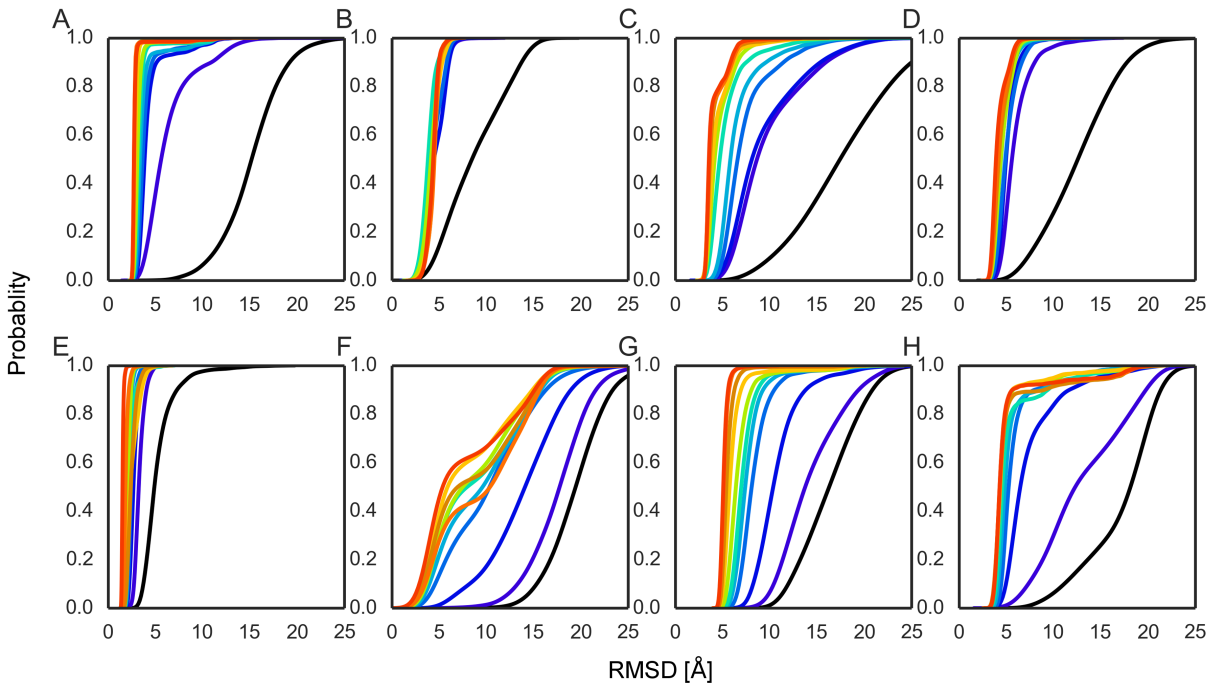

Figure 4: Cumulative probability density plots of all the eight proteins in the benchmark set (A-H). The targets are labelled as in Table 1.

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 similarly well as in the situation of four tags with four metals used in the benchmark set (Figures S8 and S9), but the PCS-based identification of improved fragments failed for target F (Figure S10). This failure was not alleviated when the PCS datasets were augmented by PCSs from four rather than two metals per center (Figure S11), which 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 metal centers was clearly crucial to increase the coverage and selection of a larger number of native-like fragments. As the algorithm requires PCSs from at least two different metal centers to identify an improved fragment, PCSs from four metal centers allows the algorithm to select from six different pair-wise combinations, whereas three metal centers allow only three combinations.

Discussion

The success of the iterative GPS-Rosetta approach lies in building the computational algorithm 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 restricts 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 [21–23,38] and to discriminate between different conformational states [39]. Here we extended this concept by taking 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 [40]. 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 present work, up to 25,000 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 of 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 towards 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 RMSDs of the fragments selected for the first iteration (panel C of Figures 2 and 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 [41], many of which are potential drug targets [42]. 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 [43]. 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 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 been previously 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 ^{13}C-methyl-labeling of the amino acids isoleucine, leucine, valine, alanine, methionine and threonine [28]. 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 Δχ-tensor parameters, sparse NOEs generated using ILVA (isoleucine, leucine, valine and alanine) labeled deuterated samples, backbone dihedral angles were predicted using TALOS [44] 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 [45,46] generated a structure with 2.6 Å RMSD to the reference structure [28]. Remarkably, using the PCS data from the same study, our iterative GPS-Rosetta protocol produced a quite similar result without using any other restraints and without making any assumptions about any of the Δχ-tensor parameters, instead optimizing them dynamically during fragment assembly. The resulting structure had a backbone RMSD of 5.0 Å to the reference structure [24].

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 traditional 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 (NCI), which is supported by the Australian Government.

References

1. Tai CH, Bai H, Taylor TJ, Lee B. Assessment of template-free modeling in CASP10 and ROLL. Proteins Struct Funct Bioinforma. 2014;82: 57–83. doi:10.1002/prot.24470

2. Xu D, Zhang Y. Ab initio protein structure assembly using continuous structure fragments and optimized knowledge-based force field. Proteins. 2012;80: 1715–1735. doi:10.1002/prot.24065

3. Roy A, Kucukural A, Zhang Y. I-TASSER: a unified platform for automated protein structure and function prediction. Nat Protoc. 2010;5: 725–738. doi:10.1038/nprot.2010.5

4. Bradley P, Baker D. Improved beta-protein structure prediction by multilevel optimization of nonlocal strand pairings and local backbone conformation. Proteins. 2006;65: 922–9. doi:10.1002/prot.21133

5. Blum B, Jordan MI, Baker D. Feature space resampling for protein conformational search. Proteins Struct Funct Bioinforma. 2010;78: 1583–1593. doi:10.1002/prot.22677

6. Qian B, Raman S, Das R, Bradley P, McCoy AJ, Read RJ, et al. High-resolution structure prediction and the crystallographic phase problem. Nature. Nature Publishing Group; 2007;450: 259–64. doi:10.1038/nature06249

7. Lange OF, Baker D. Resolution-adapted recombination of structural features significantly improves sampling in restraint-guided structure calculation. Proteins Struct Funct Bioinforma. 2012;80: 884–895. doi:10.1002/prot.23245

8. Fleishman SJ, Baker D. Role of the biomolecular energy gap in protein design, structure, and evolution. Cell. Elsevier; 2012;149: 262–273. doi:10.1016/j.cell.2012.03.016

9. van der Schot G, Zhang Z, Vernon R, Shen Y, Vranken WF, Baker D, et al. Improving 3D structure prediction from chemical shift data. J Biomol NMR. 2013;57: 27–35. doi:10.1007/s10858-013-9762-6

10. Lange OF, Rossi P, Sgourakis NG, Song Y, Lee H-W, Aramini JM, 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. 2012;109: 10873–8. doi:10.1073/pnas.1203013109

11. Raman S, Lange OF, Rossi P, Tyka M, Wang X, Aramini JM, et al. NMR structure determination for larger proteins using backbone-only data. Science. 2010;327: 1014–8. doi:10.1126/science.1183649

12. Rao T, Lubin JW, Armstrong GS, Tucey TM, Lundblad V, Wuttke DS. Structure of Est3 reveals a bimodal surface with differential roles in telomere replication. Proc Natl Acad Sci U S A. 2014;111:

214–8. doi:10.1073/pnas.1316453111

13. Sgourakis NG, Natarajan K, Ying J, Vogeli B, Boyd LF, Margulies DH, 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. Elsevier Ltd; 2014;22: 1263–1273. doi:10.1016/j.str.2014.05.018

14. Bertini I, Luchinat C, Parigi G, Pierattelli R. Perspectives in paramagnetic NMR of metalloproteins. Dalt Trans. 2008; 3782–3790. doi:10.1039/b719526e

15. Otting G. Protein NMR using paramagnetic ions. Annu Rev Biophys. 2010;39: 387–405. doi:10.1146/annurev.biophys.093008.131321

16. Liu W-M, Overhand M, Ubbink M. The application of paramagnetic lanthanoid ions in NMR spectroscopy on proteins. Coord Chem Rev. 2013; Available: http://www.sciencedirect.com/science/article/pii/S0010854513002464

17. Keizers PHJ, Ubbink M. Paramagnetic tagging for protein structure and dynamics analysis. Prog Nucl Magn Reson Spectrosc. Elsevier B.V.; 2011;58: 88–96. doi:10.1016/j.pnmrs.2010.08.001

18. Jaroniec CP. Structural studies of proteins by paramagnetic solid-state NMR spectroscopy. J Magn Reson. Elsevier Inc.; 2015;253: 50–59. doi:10.1016/j.jmr.2014.12.017

19. Bertini I, Luchinat C, Parigi G. Magnetic susceptibility in paramagnetic NMR. Prog Nucl Magn Reson Spectrosc. 2002;40: 249–273. doi:10.1016/S0079-6565(02)00002-X

20. Schmitz C, Vernon R, Otting G, Baker D, Huber T. Protein structure determination from pseudocontact shifts using ROSETTA. J Mol Biol. Elsevier B.V.; 2012;416: 668–677. doi:10.1016/j.jmb.2011.12.056

21. Yagi H, Pilla KB, Maleckis A, Graham B, Huber T, Otting G. Three-dimensional protein fold determination from backbone amide pseudocontact shifts generated by lanthanide tags at multiple sites. Structure. Elsevier; 2013;21: 883–890. doi:10.1016/j.str.2013.04.001

22. Li J, Pilla KB, Li Q, Zhang Z, Su X, Huber T, et al. Magic Angle Spinning NMR Structure Determination of Proteins from Pseudocontact Shifts. J Am Chem Soc. 2013;135: 8294–8303. doi:10.1021/ja4021149

23. Pilla KB, Leman JK, Otting G, Huber T. Capturing conformational States in proteins using sparse paramagnetic NMR data. PLoS One. 2015;10: e0127053. doi:10.1371/journal.pone.0127053

24. Crick DJ, Wang JX, Graham B, Swarbrick JD, Mott HR, Nietlispach D. Integral membrane protein

structure determination using pseudocontact shifts. J Biomol NMR. 2015; 1–9. doi:10.1007/s10858-015-9899-6

25. Rohl CA, Strauss CEM, Misura KMS, Baker D. Protein structure prediction using Rosetta. Methods Enzymol. 2004;383: 66–93. doi:10.1016/S0076-6879(04)83004-0

26. Bowers PM, Strauss CEM, Baker D. De novo protein structure determination using sparse NMR data. J Biomol NMR. Kluwer Academic Publishers; 2000;18: 311–318. doi:10.1023/A:1026744431105

27. Kim DE, Chivian D, Baker D. Protein structure prediction and analysis using the Robetta server. Nucleic Acids Res. 2004;32: W526–31. doi:10.1093/nar/gkh468

28. Gautier A, Mott HR, Bostock MJ, Kirkpatrick JP, Nietlispach D. Structure determination of the seven-helix transmembrane receptor sensory rhodopsin II by solution NMR spectroscopy. Nat Struct Mol Biol. Nature Publishing Group; 2010;17: 768–774. doi:10.1038/nsmb.1807

29. Graham B, Loh CT, Swarbrick JD, Ung P, Shin J, Yagi H, et al. DOTA-amide lanthanide tag for reliable generation of pseudocontact shifts in protein NMR spectra. Bioconjug Chem. American Chemical Society; 2011;22: 2118–2125. doi:10.1021/bc200353c

30. Swarbrick JD, Ung P, Chhabra S, Graham B. An iminodiacetic acid based lanthanide binding tag for paramagnetic exchange NMR spectroscopy. Angew Chem Int Ed Engl. 2011;50: 4403–4406. doi:10.1002/anie.201007221

31. Stanton-Cook MJ. pyParaTools v0.8-alpha. 2014; doi:10.5281/zenodo.10313

32. Royant a, Nollert P, Edman K, Neutze R, Landau EM, Pebay-Peyroula E, et al. X-ray structure of sensory rhodopsin II at 2.1-A resolution. Proc Natl Acad Sci U S A. 2001;98: 10131–10136. doi:10.1073/pnas.181203898

33. Cornilescu G, Marquardt JL, Ottiger M, Bax A. Validation of protein structure from anisotropic carbonyl chemical shifts in a dilute liquid crystalline phase. J Am Chem Soc. ACS Publications; 1998;120: 6836–6837. doi:10.1021/ja9812610

34. Hagn F, Etzkorn M, Raschle T, Wagner G. Optimized phospholipid bilayer nanodiscs facilitate high-resolution structure determination of membrane proteins. J Am Chem Soc. 2013;135: 1919–25. doi:10.1021/ja310901f

35. Ksiazek D, Brandstetter H, Israel L, Bourenkov GP, Katchalova G, Janssen K-P, et al. Structure of the n-terminal domain of the adenylyl cyclase-associated Protein (CAP) from Dictyostelium discoideum. Structure. 2003;11: 1171–1178. doi:10.1016/S0969-2126(03)00180-1

36. Edrington TC, Kintz E, Goldberg JB, Tamm LK. Structural basis for the interaction of lipopolysaccharide with outer membrane protein H (OprH) from Pseudomonas aeruginosa. J Biol Chem. 2011;286: 39211–23. doi:10.1074/jbc.M111.280933

37. Legge GB, Kriwacki RW, Chung J, Hommel U, Ramage P, Case DA, et al. NMR solution structure of the inserted domain of human leukocyte function associated antigen-1. J Mol Biol. 2000;295: 1251–64. doi:10.1006/jmbi.1999.3409

38. de la Cruz L, Nguyen THD, Ozawa K, Shin J, Graham B, Huber T, 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. 2011;133: 19205–19215. doi:10.1021/ja208435s

39. Skinner SP, Liu W-M, Hiruma Y, Timmer M, Blok A, Hass M a. S, et al. Delicate conformational balance of the redox enzyme cytochrome P450cam. Proc Natl Acad Sci. 2015; 201502351. doi:10.1073/pnas.1502351112

40. Dill KA, Ozkan SB, Weikl TR, Chodera JD, Voelz VA. The protein folding problem: when will it be solved? Curr Opin Struct Biol. 2007;17: 342–6. doi:10.1016/j.sbi.2007.06.001

41. Wallin E, von Heijne G. Genome-wide analysis of integral membrane proteins from eubacterial, archaean, and eukaryotic organisms. Protein Sci. 1998;7: 1029–1038. doi:10.1002/pro.5560070420

42. Yildirim M a, Goh K-I, Cusick ME, Barabási A-L, Vidal M. Drug-target network. Nat Biotechnol. 2007;25: 1119–1126. doi:10.1038/nbt1338

43. Kaptein R, Wagner G. NMR studies of membrane proteins. J Biomol NMR. 2015;61: 181–4. doi:10.1007/s10858-015-9918-7

44. Cornilescu G, Delaglio F, Bax A. Protein backbone angle restraints from searching a database for chemical shift and sequence homology. J Biomol NMR. Kluwer Academic Publishers; 13: 289–302. doi:10.1023/A:1008392405740

45. Schwieters CD, Kuszewski JJ, Tjandra N, Clore GM. The Xplor-NIH NMR molecular structure determination package. J Magn Reson. 2003;160: 65–73. doi:10.1016/S1090-7807(02)00014-9

46. Banci L, Bertini I, Cavallaro G, Giachetti A, Luchinat C, Parigi G. Paramagnetism-based restraints for Xplor-NIH. J Biomol NMR. 2004;28: 249–261. doi:10.1023/B:JNMR.0000013703.30623.f7