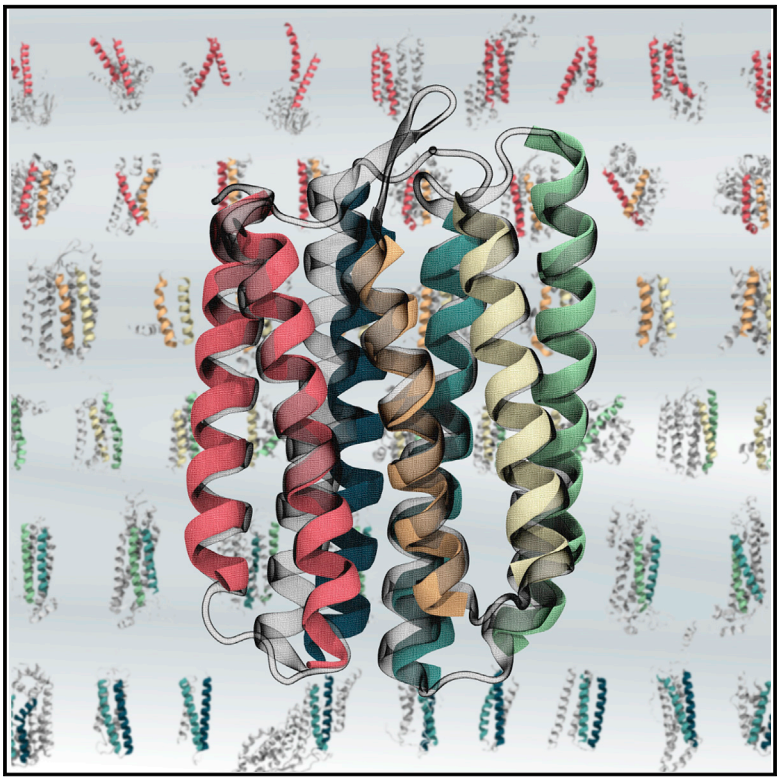
Resource

Structure

Protein Structure Determination by Assembling Super-Secondary Structure Motifs Using Pseudocontact Shifts

Graphical Abstract



Authors

Kala Bharath Pilla, Gottfried Otting, Thomas Huber

Correspondence

t.huber@anu.edu.au

In Brief

A new computational algorithm is introduced that assembles the 3D structure of a protein from its constituent super-secondary structural motifs with the help of pseudocontact shift (PCS) restraints for backbone amide protons, where the PCSs are produced from different metal centers.

Highlights

- Smotifs are used as building blocks for protein structure determination
- Three lanthanide tags resolve the position and orientation of Smotifs in 3D space
- The algorithm is dependent and driven only by sparse pseudocontact shift data
- The method is especially suitable for structure determination of large proteins

Pilla et al., 2017, Structure *25*, 559–568
March 7, 2017
http://dx.doi.org/10.1016/j.str.2017.01.011

CellPress

Structure
Resource

Protein Structure Determination by Assembling Super-Secondary Structure Motifs Using Pseudocontact Shifts

Kala Bharath Pilla,[1,2] Gottfried Otting,[1] and Thomas Huber[1,3,*]
[1]Research School of Chemistry, Australian National University, Canberra, ACT 2601, Australia
[2]Department of Chemistry, University of Calgary, Calgary, AB T2N 1N4, Canada
[3]Lead Contact
*Correspondence: t.huber@anu.edu.au
http://dx.doi.org/10.1016/j.str.2017.01.011

SUMMARY

Computational and nuclear magnetic resonance hybrid approaches provide efficient tools for 3D structure determination of small proteins, but currently available algorithms struggle to perform with larger proteins. Here we demonstrate a new computational algorithm that assembles the 3D structure of a protein from its constituent super-secondary structural motifs (Smotifs) with the help of pseudocontact shift (PCS) restraints for backbone amide protons, where the PCSs are produced from different metal centers. The algorithm, DINGO-PCS (3D assembly of Individual Smotifs to Near-native Geometry as Orchestrated by PCSs), employs the PCSs to recognize, orient, and assemble the constituent Smotifs of the target protein without any other experimental data or computational force fields. Using a universal Smotif database, the DINGO-PCS algorithm exhaustively enumerates any given Smotif. We benchmarked the program against ten different protein targets ranging from 100 to 220 residues with different topologies. For nine of these targets, the method was able to identify near-native Smotifs.

INTRODUCTION

Determining the three-dimensional (3D) structures of proteins by nuclear magnetic resonance (NMR) spectroscopy becomes increasingly challenging for proteins of increasing molecular weight, as the NMR spectra show more spectral overlap and not all signals can be resolved. Spectral overlap is particularly severe for ^{1}H-NMR resonances of amino acid side chains, hindering the unambiguous assignment of nuclear Overhauser effects (NOE) and making 3D structure determinations from NOEs difficult. As an alternative, structural restraints can be derived from pseudocontact shifts (PCSs) of the relatively well-resolved backbone amide resonances. PCSs can be measured in simple 2D ^{15}N-heteronuclear single quantum coherence spectra and can be observed for nuclear spins at large distances (up to about 40 Å) from the paramagnetic center, especially when strongly paramagnetic lanthanide ions are used (Otting, 2010). Computational algorithms have been established to resolve 3D protein structures using solely backbone amide PCSs (Schmitz et al., 2012; Pilla et al., 2016). Lanthanide-generated PCSs have also been used to elucidate ligand-induced conformational changes in proteins (Pilla et al., 2015; Saio et al., 2015) or to characterize the binding poses of ligand molecules (Guan et al., 2013; Chen et al., 2016).

PCSs are measured as the change in chemical shifts caused by the presence of a paramagnetic metal ion with an anisotropic component $\Delta\chi$ of the magnetic susceptibility tensor χ. The PCS of a nuclear spin is read from NMR spectra as the difference in chemical shift between paramagnetic and diamagnetic states, and is given by (Bertini et al., 2002)

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

where r, θ, φ, define the polar coordinates of the nuclear spin with respect to the center and principal axes of the $\Delta\chi$ tensor, where $\Delta\chi_{ax}$ and $\Delta\chi_{rh}$ are the axial and rhombic components of the $\Delta\chi$ tensor. If the location of the metal center and the size and orientation of the $\Delta\chi$ tensor with respect to the protein are known, Equation 1 can be used to convert the PCS of a nuclear spin into a restraint of the location of the spin in the $\Delta\chi$-tensor frame. Many methods have been devised for site specifically tagging proteins with paramagnetic lanthanide ions to obtain structural information in solution (Su and Otting, 2010; Keizers and Ubbink, 2011; Liu et al., 2014), in the solid state (Jaroniec, 2015), and inside cells (Pan et al., 2016).

The need to know the $\Delta\chi$ tensor parameters before PCSs can be used as structural restraints poses an intrinsic difficulty for 3D structure determination from PCSs alone. We overcame this problem by implementing PCSs as structural restraints in the structure prediction software Rosetta, where the PCSs help identify correctly folded decoys and guide folding of the polypeptide chain toward its native structure. For small proteins up to about 150 residues, this approach allowed us to determine the $\Delta\chi$ tensor and 3D structure simultaneously (Schmitz et al., 2012; Yagi et al., 2013a).

Limitations arise for larger proteins, requiring different approaches to 3D structure determination. Rosetta builds the 3D structure by assembling nine- and three-residue fragments from

a fragment library that is specifically generated for the target protein from homologous or orthologous protein structures. The fragment libraries are restricted in size and contain 200 fragments for any given nine- and three-residue fragment window. These restrictions in fragment library size are a compromise between minimizing conformational search space and maximizing the probability that the native protein structure can be accurately assembled from the fragments. Consequently, the Rosetta protocol has repeatedly been shown to perform exceptionally well for predicting the structures of small proteins (Moult et al., 2014), whereas producing near-native structures of larger proteins requires unique iterative search algorithms and additional experimental restraints (Raman et al., 2010; Pilla et al., 2016).

Structural motifs, characterized as a group of regular secondary structure elements connected by loops, such as zinc-/finger, helix-turn-helix, β meander motifs, and Greek key motifs, are commonly found in many protein families. The recurrence of these motifs is thought to reflect duplications, mutations, shuffling, and fusion of genes throughout the course of evolution (Lupas et al., 2001; Alva et al., 2015) and, therefore, they represent a more natural description of building blocks to assemble protein folds. The basic unit of a super-secondary structural motif (Smotif) is defined as a pair of regular secondary structure elements connected by a loop. By this definition, there are only four basic types of Smotifs, which can be referred to as α-α, β-β, α-β, and β-α, where α represents a helical element and β an extended polypeptide strand. Recently, Smotifs have been employed to build topology-independent structure classification tools for quantifiable identification of structural relationships between disparate topologies (Dybas and Fiser, 2016). Importantly, the total number of different Smotifs observed in all protein structures known to date has not increased since 2000 (last reported in 2010) (Fernandez-Fuentes et al., 2010), suggesting that our structural knowledge of Smotifs is close to complete. Furthermore, it has been shown that all known protein structures can be reconstructed with good accuracy from the finite set of Smotifs (Fernandez-Fuentes et al., 2010).

Given these properties, Smotifs lend themselves to use as basic building blocks for sampling native-like protein structures and replacing smaller fragment libraries. Two programs using Smotifs for structure prediction have been described earlier, called Smotifs in template-free modeling (SmotifTF) (Vallat et al., 2015) and chemical shift-guided Smotif assembly (SmotifCS) (Menon et al., 2013). Both approaches performed on a par with current state-of-the-art software such as I-TASSER (Roy et al., 2010), HHpred (Söding et al., 2005), and Rosetta (Rohl et al., 2004; Shen et al., 2009), but were reliant on Smotif libraries specifically generated for the chosen target, making the libraries non-universal. In the case of SmotifTF (Vallat et al., 2015), the Smotif libraries are generated from homologous protein structures, while the SmotifCS (Menon et al., 2013) approach selects the Smotifs that locally match the experimentally observed chemical shifts. A possible drawback of these approaches lies in the heuristic Monte Carlo sampling of protein structures, which limits their successful application to small proteins of around 110 residues. Here we present a new computational algorithm, DINGO-PCS (3D assembly of Individual Smotifs to Near-native Geometry as Orchestrated by PCSs), that utilizes PCSs of backbone amide protons as the only experimental data to identify, orient, and build the protein structure from its constituent Smotifs. PCSs of nuclear spins are expected to be available for paramagnetic metal centers at three different sites of the protein to pinpoint the location and orientation of the Smotif in a manner analogous to the use of satellites in the global positioning system (GPS). The analogy can be visualized by using Equation 1 to determine isosurfaces of constant PCS. The PCS isosurface comprises all coordinates where nuclear spins experience a particular PCS value, similar to the way in which the distance measurement to each satellite in the GPS system positions the user on a sphere that is centered on the satellite and has a radius corresponding to this distance. While an experimentally determined PCS value ties a nuclear spin to a location on an isosurface, a second PCS value for the same spin measured from a lanthanide attached at a different site restricts the location of the nuclear spin to lie on the line defined by the intersection of the respective PCS isosurfaces, and a third PCS value obtained from a sample with a lanthanide attached at yet another site leaves only two points as the possible positions of the nuclear spin. DINGO-PCS takes advantage of the unique information content carried by PCSs from multiple metal centers to select and assemble the constituent Smotifs of a target protein from the database of known Smotifs. The algorithm depends on PCSs as the only experimental data and works without any physical or knowledge-based energy function, which are usually required to validate the protein state. In contrast to Rosetta or other fragment assembly algorithms that generate target-specific fragment libraries by reference to the amino acid sequence, the Smotif libraries used by DINGO-PCS can be applied to any target protein. Therefore, the DINGO-PCS algorithm is independent of amino acid sequence or availability of 3D structures of homologs or orthologs, making it universally applicable to any protein topology including artificially designed proteins. To demonstrate this point, native Smotifs were excluded from the search in the context of this publication.

In the following we demonstrate the performance of the DINGO-PCS algorithm for ten different proteins ranging from 100 to 224 residues, including a large 218-residue, 7-transmembrane (7-TM) α-helical microbial membrane protein, the phototactic receptor sensory rhodopsin II (pSRII) from *Natronomonas pharaonis,* where experimental PCSs were measured from four different metal centers (Crick et al., 2015). In addition, we assess the performance of the DINGO-PCS algorithm with a range of proteins of different fold architecture, including membrane-bound, α-helical, β barrel, and α/β topologies, using simulated PCS data. Nine out of ten target structures were well reproduced by the algorithm.

RESULTS

DINGO-PCS Performance on the Integral Membrane Protein pSRII

The DINGO-PCS algorithm was first tested with the structure determination of the 7-TM α-helical protein pSRII from *Natronomonas pharaonis*. The final calculated structure was selected based on the Smotif assembly that best fits the experimental PCSs, using the score function in Equation 2 (see Figures 4 and 5 and the Experimental Procedures section for a detailed description of the algorithm and Smotif assembly). It consists

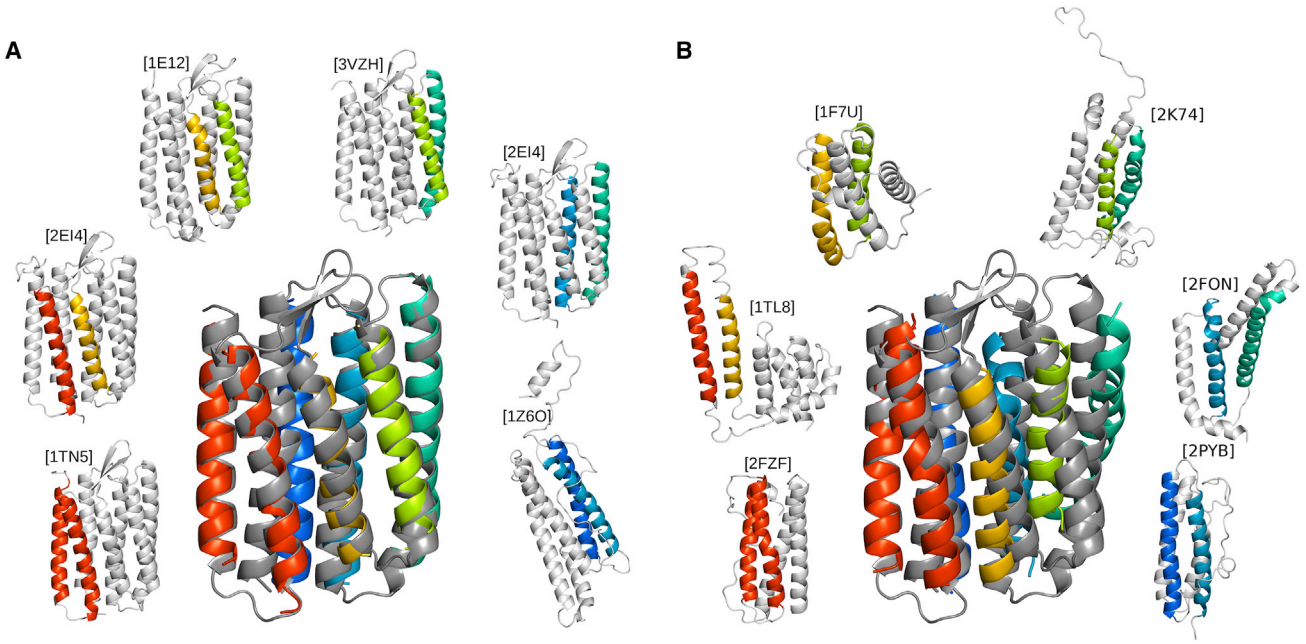


Figure 1. Smotif Assembly Performed by DINGO-PCS on the Integral Membrane Protein pSRII

(A) Cartoon representations of the final model superimposed onto the crystal structure (in gray) and of the parent proteins from which the Smotifs were derived. The respective parent proteins are labeled with their PDB identification code. The Smotifs of the parent proteins and in the final structure are highlighted by corresponding coloring.

(B) Same as (A), except that the Smotifs of the rhodopsin family were excluded from the library.

(C) Similarity between the sequences of the assembled Smotifs and the native sequence of the target. The model sequence was generated by concatenating the sequences from the individual Smotifs. Target A contains rhodopsin homolog sequences, whereas target A* excludes them. Similar amino acid types are highlighted in red and framed in blue. The sequence alignment was generated manually.

of six individual Smotifs (Figure 1A) and has a C^{α} root-mean-square deviation (RMSD) of 1.9 Å to the crystal structure (Royant et al., 2001), which drops to 1.0 Å without the C-terminal seventh helix. This quality is remarkable, considering that PCSs presented the only experimental data in the structure calculation and were available for only 66% of the residues and, furthermore, the native structure and any close homologs had been excluded from the Smotif library by design. The RMSD values mentioned throughout the text were calculated for the residues covered by Smotifs only, unless explicitly stated otherwise. Five out of the six assembled Smotifs stemmed from structural homologs of the sensory protein rhodopsin, specifically from bacteriorhodopsin structures of *Halobacterium salinarum* and its mutants (PDB: 1TN5, 2EI4, 1E12, 3VHZ). Even in an alignment of only the segments of regular secondary structure, the amino acid sequences of the Smotifs used to calculate the structure and the actual amino acid sequence of the target protein shared only 26% identity (Figure 1C).

The top six models identified by the scoring function of Equation 2 displayed C^{α} RMSDs ranging between 1.9 and 2.2 Å to the reference crystal structure (Royant et al., 2001). In all cases, the assembled structures derived most Smotifs from the CATH superfamily of rhodopsin 7-TM helix proteins (CATH ID 1.20.1070.10). This CATH class is populated with structural homologs of the target protein. To increase the stringency of the test, we removed all Smotifs from this superfamily for the assembly and recalculated the structure of pSRII with the same PCS data. The structure assembled from these data (Figure 1B) had a C^{α} RMSD of 4.9 Å to the reference crystal structure (Royant et al., 2001). In this calculation, none of the parent proteins were rhodopsins and the Smotifs originated from a very diverse set of proteins from archaea, prokaryotes, and eukaryotes. Specifically, the PDB IDs of the Smotifs selected for the final structure were, starting from the N terminus: PDB: 2FZF (a hypothetical protein from *Pyrococcus furiosus*), PDB: 1TL8 (a DNA topoisomerase from *Homo sapiens*), PDB: 1F7U (arginyl-tRNA synthetase from *Saccharomyces cerevisiae*), PDB: 2K74 (integral membrane enzyme DsbB from *Escherichia coli*), PDB: 2FON (acyl-CoA oxidase from *Lycopersicon esculentum*), and PDB: 2PYB (Napa protein from *Borrelia burgdorferi*). The sequence identity of the final assembled structure to the residues in the respective secondary structure elements was only

Table 1. Performance Benchmark of the DINGO-PCS Algorithm

Targets	PDB:	N_{res}[a]	N_{res} in Smotifs	C^{α} RMSD[b] (Å)	Q-Factor[c]	BMRB:
A (pSRII)	1H68	218	181	1.9 Å	0.12	16678
A* (pSRII)	1H68	218	178	4.9 Å	0.20	16678
B (ERp29-C)	2M66	106	67	3.3 Å	0.14	4920
C (OmpX)	2M06	148	86	2.1 Å	0.15	18796
D (polyketide cyc-like protein)	2M47	157	92	5.6 Å	0.21	18989
E (peptidyl-tRNA hydrolase)	2Z2I	179	98	3.3 Å	0.18	7055
F (human leukocyte function-associated antigen-1)	1DGQ	188	91	5.0 Å	0.18	4553
G (Talin, C-terminal actin binding site)	2JSW	189	137	4.5Å	0.32	15411
H (Pactolus domain-1)	2IUE	212	104	5.3 Å	0.31	7313
I (STARD6)	2MOU	220	103	3.5 Å	0.21	19952
J (adhesion protein delta-Bd37)	2LUD	224	NA	Fail	NA	18517

*Excludes structural homologs.
[a]Number of amino acid residues.
[b]The C^{α} root-mean-square deviation (RMSD) was calculated between the best Smotif assembly calculated by DINGO-PCS, which was identified as the structure best fulfilling the PCS data, and the Smotif residues in the corresponding reference structure.
[c]The Q-factor was calculated as the RMSD between experimental and back-calculated PCSs divided by the RMS of the experimental PCSs.

4.5% (Figure 1C). This result demonstrates that DINGO-PCS can identify the topologically correct Smotifs even in the absence of structural homologs.

DINGO-PCS Performance Benchmark

We benchmarked the performance of the DINGO-PCS algorithm using a set of nine additional proteins. Only PCSs that could realistically be measured experimentally were used to assemble the Smotifs and no other data. The quality of Smotif assemblies varied between different the different benchmark proteins, with the C^{α} RMSDs ranging between 1.5 and 5.6 Å with respect to the respective NMR or X-ray reference structure (Table 1). Figure 2 depicts superpositions of the Smotif assemblies of the targets with their 3D reference structures. The Q-factor metric offers an alternative way to assess structural quality (Bax, 2003), which penalizes less heavily for outlying residues than the RMSD metric and, in the absence of a reference 3D structure, allows ranking the Smotif assemblies that best define the observed PCSs. The Q-factors of the different target proteins ranged from 0.12 to 0.32 (Table 1), indicating very good agreement of the PCS data with the structural models (Cornilescu et al., 1998).

All targets except targets D and J were successfully assembled with RMSD values below 5.0 Å to their corresponding reference structures. The relatively large RMSD value observed for target D (5.6 Å) arose primarily from an incorrect orientation of the N-terminal β sheet (Figure 2D). For target J, the Smotif assembly failed beyond the first pair of secondary structure elements (Figure 2J). While the first Smotif comprising residues 104–121 and 132–149 had a low RMSD of 2.5 Å to the native structure, no hit was found in the library for the next Smotif (residues 77–91 and 104–121) within the specified PCS and RMSD thresholds (see the Experimental Procedures). An explicit search in the library for the presence of native-like Smotifs for target J revealed no entry within 3.0 Å RMSD of the native structure. Therefore, progression in the assembly of this target was derailed by the absence of this key Smotif in the database.

We also tested the DINGO-PCS algorithm with reduced PCS datasets. For example, targets A, C, G, and H were tested by restricting the PCS data to three metal centers and two paramagnetic metals per center, corresponding to a 62% reduction in PCS data. For targets A, C, and G, DINGO-PCS was found to perform equally well when compared with Smotif assemblies using PCS data from four metal centers (Figure S1 and Table S1), but target H failed to complete the Smotif assembly as the DINGO-PCS algorithm correctly assembled only six out of ten Smotifs.

All-Atom 3D Structures

The Smotif assemblies produced by DINGO-PCS are devoid of coordinates of side-chain atoms and loop regions. However, starting from the assembled Smotif structure, the missing atoms can easily be filled in either by a comparative modeling algorithm or by an ab initio fragment assembly algorithm (see the Experimental Procedures). Figures 3A and 3B show the result obtained by using Rosetta's comparative modeling (RosettaCM) protocol applied to targets A and B. The top five models for targets A and B, ranked using Equation 3, showed C^{α} RMSD values over all residues ranging between 2.6 and 3.4 Å to the reference structure (Royant et al., 2001) for target A and 2.6–3.8 Å to the reference structure (Yagi et al., 2013a) for target B.

Smotif assemblies with C^{α} RMSDs greater than 4.0 Å are not suitable templates for comparative modeling. Nonetheless, such Smotif assemblies can be used to improve the sampling in ab initio fragment assembly algorithms such as iterative GPS-Rosetta (Pilla et al., 2016). This is illustrated by target A*, where the Smotif assembly produced an RMSD of 4.9 Å from the reference crystal structure (Royant et al., 2001). Two different types of restraints were derived from the Smotif assembly of target A*. First, fragment libraries were populated with fragments from the Smotifs of the assembled structure by translating the coordinates into the format of Rosetta's 9- and 3-residue fragments to replace the first 20 entries in the standard fragment libraries. Next, a distance constraint map was generated,

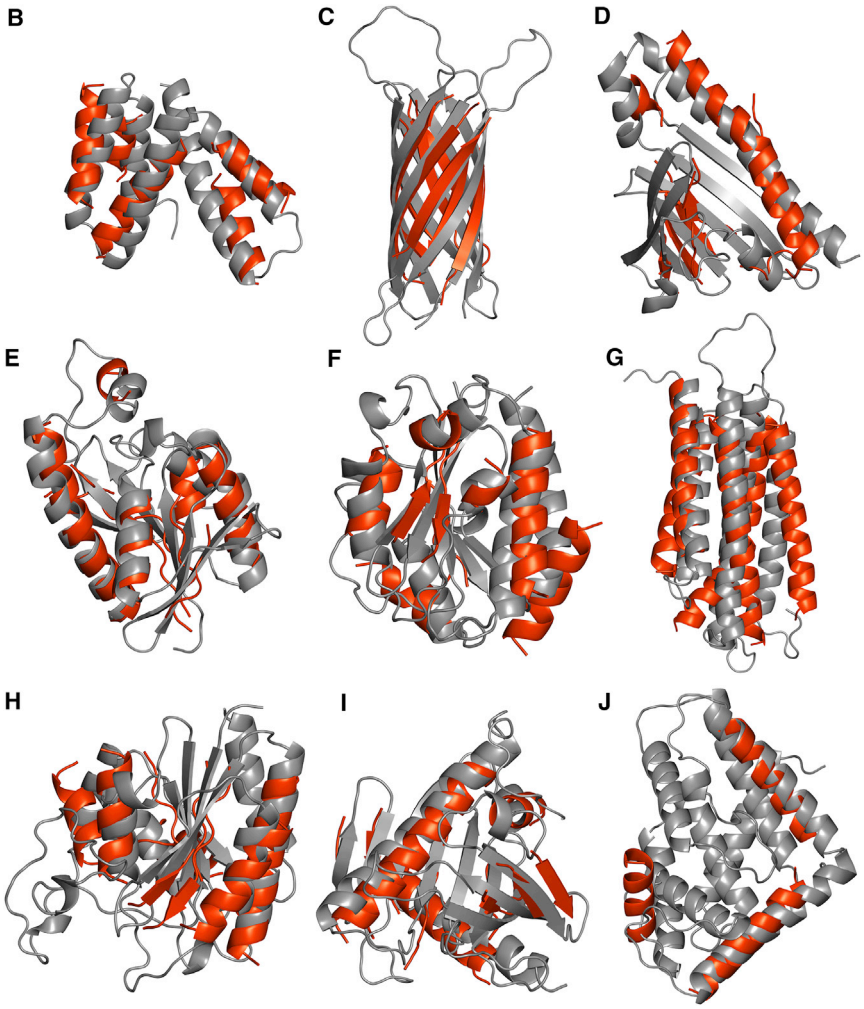


Figure 2. Superpositions of the Backbone Structures of the Best Smotif Assemblies Calculated with DINGO-PCS, Shown in Red, onto the Corresponding Reference Structures, Shown in Gray

The best Smotif assemblies were identified as the assemblies best fulfilling the PCS data. The targets are labeled as in Table 1 (See also Figure S1 and Table S1).

which identified the pairs of amino acid residues located within 3.5–7.5 Å in the Smotif assembled structure. The Smotif-enhanced fragments and the constraint map were combined with the PCS data from multiple metal centers as input for the iterative GPS-Rosetta algorithm (Pilla et al., 2016) to obtain all-atom models. The models were ranked using Equation 3. The C^{α} RMSDs (over all residues) of the top five models to the reference structure (Royant et al., 2001) were within 2.5–3.3 Å (Figure 3C). The complete result of the PCS-driven iterative resampling GPS-Rosetta algorithm is shown in Figure S2. The incorporation of Smotif-derived fragments and the constraint map drastically improved the sampling of good quality structures. To demonstrate this, we generated two sets of 3,000 structures each following the GPS-Rosetta algorithm, first using the Rosetta standard nine- and three-residue non-homologous fragment libraries, and second using the combined Smotif-enhanced fragment library and Smotif-derived restraints. The probability plots depicting the change in the quality of structures sampled by the introduction of Smotif-derived restraints are shown in Figure 3D. The median C^{α} RMSD (over all residues) of the distribution (Figure 3D) shifted from 13 Å without Smotif-derived restraints to 6 Å with Smotif-derived restraints.

DISCUSSION

Historically, structure prediction algorithms were developed to solve the structures of small proteins, which can be assembled from short fragments. In the case of Rosetta, the fragments are at most nine residues long. Structure predictions of small proteins have been very successful. Large proteins, however, resist the current computational approaches as the increase in protein size is coupled with a very large expansion of conformational space, which becomes correspondingly difficult to explore. Various iterative methods such as identification and resampling of structural features during fragment assembly in Rosetta (RASREC) (Lange and Baker, 2012) and combination of short molecular dynamics simulation with Rosetta sampling (Lindert et al., 2013) have been proposed to combat the sampling problem. These different approaches still require an enormous amount of computational time (in the order of 10^5 CPU hr), yet yield good structures for only 70% of the targets (van der Schot et al., 2013). Our present work shows that the sampling problem can be overcome by replacing short fragment libraries by a saturated Smotif library and using DINGO-PCS to assemble the correct Smotifs with the help of overlapping PCS datasets from multiple metal centers.

Identification of correct Smotifs is crucial and incorporation of any false positives, especially during early stages of assembly, quickly propagates the error, leading to premature termination of the whole assembly. The identification of correct Smotifs depends on two major factors. Firstly, the secondary structure prediction of the target sequence must be accurate. This can be improved with the inclusion of backbone chemical shift information, but the accuracy is still limited to 89% (Shen and Bax, 2013). Adding an uncertainty to the termini of the secondary structure elements in a Smotif only partially overcomes the inaccuracies, since short secondary structure elements of less than five residues are omitted from the library, as they do not carry enough PCS data for identification as correct Smotifs. Second, although targets with short secondary structure elements (five to ten residues) carry enough PCSs to estimate the $\Delta\chi$ tensors accurately, sufficient coverage with sizable PCSs from different

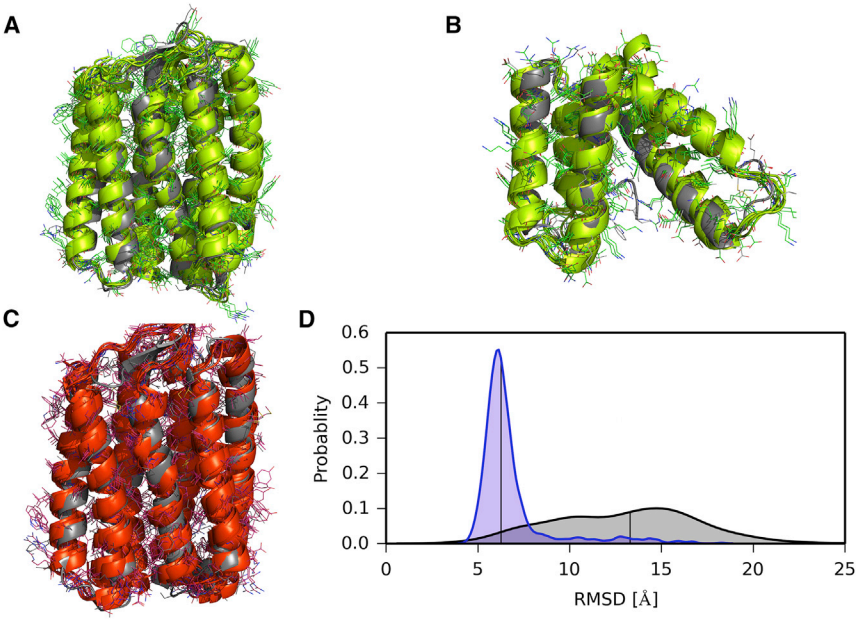


Figure 3. All-Atom 3D Models Generated from Smotif Assemblies

(A) Superposition of the five best RosettaCM models (selected for best fit of the PCS data), shown in green, onto the reference crystal structure, shown in gray, for target A.

(B) Same as (A) but for target B.

(C) Superposition of the five best iterative GPS-Rosetta models, shown in red, onto the reference crystal structure, shown in gray, for target A*.

(D) Probability plots illustrating the conformational sampling bias created by incorporating Smotif-derived restraints from target A*. Results from GPS-Rosetta sampling without and with Smotif-derived restraints are shown in gray and blue, respectively. The vertical lines identify the respective medians (See also Figure S2).

metal centers is required for fitting $\Delta\chi$ tensors of sufficient accuracy to fall within the specified thresholds.

The DINGO-PCS algorithm aims to position and orient Smotifs uniquely in 3D space, and, by analogy with the GPS principle, PCS data from three metal centers may suffice. We tested this hypothesis on targets A, C, G, and H. Indeed, the Smotif assemblies were of similar or even marginally better quality when compared with assemblies performed with PCS data from four metal centers and four metal ions per center (Figure S1 and Table S1). Only for target H, did reduction of the number of metal sites from four to three allow identification of only six out of ten Smotifs, and the assembled structure had an RMSD of 8.1 Å to the reference structure. In this case, increasing the PCS datasets from two metals per centers to four metals per center (12 PCS datasets) improved the structural quality to 3.5 Å RMSD, but without increasing the number of assembled Smotifs. As a final test, we also ran the Smotif assembly using PCS data from four metal centers and two metals per center (eight PCS datasets). This led to a small improvement in the structure (3.2 Å RMSD) but still failed to complete the assembly with only six out of ten Smotifs completed. These results indicate that PCS constraints from multiple metal ions at the same metal position do not add substantially new information. Target H is similar in size to targets A, C, and G, but spans over 11 secondary structure elements (10 Smotifs), where 8 of the 11 secondary structure elements are short, ranging from 6 to 11 residues. As short Smotifs require a larger number of sizable PCSs for accurate positioning and orientation, this may explain the difficulties of assembling target H. Indeed, targets A, C, and G have longer Smotifs, which provide the required coverage and PCS magnitudes to derive sufficient information from three metal centers.

PCSs by themselves present exceptionally useful restraints for Smotif identification and assembly, more than other types of NMR data such as sparse residual dipolar couplings (RDCs), paramagnetic relaxation enhancements (PREs) or NOEs. Using RDCs, the lack of distance information in the alignment tensor would make the selection of Smotifs ambiguous. PRE and NOE effects are both relatively short range and therefore would require large datasets to identify the best Smotifs, which can be difficult to obtain. In contrast, PCSs of backbone amide protons are sufficient for DINGO-PCS. Such PCS datasets can easily be established from sensitive NMR experiments that can be recorded even for proteins with poor solubility, adding to the usefulness of the method.

Although the DINGO-PCS algorithm exhaustively enumerates the Smotif library, it is computationally inexpensive. For example, it took a mere 50 CPU hr or 30 min on a 128-processor cluster to assemble all Smotifs for target A. This is in stark contrast to GPS-Rosetta, which took 28,000 CPU hr to achieve the same result (Pilla et al., 2016). Notably, however, DINGO-PCS takes longer to enumerate the Smotifs for targets with smaller Smotifs (fewer than ten residues per secondary structure element). For target B, it took 3,000 CPU hr to assemble the Smotifs, but GPS-Rosetta still took 12,000 CPU hr for this target (Pilla et al., 2016). The more the Smotif libraries are populated with a large number of small Smotifs, the more CPU time is required for exhaustive enumeration of the library. Therefore, DINGO-PCS is better suited for structure calculations of proteins with longer secondary structure elements. The DINGO-PCS software package can be downloaded free from https://github.com/kalabharath/DINGO-PCS and the precompiled universal Smotif libraries are freely available from http://comp-bio.anu.edu.au/huber/Smotifs/.

Conclusion

In conclusion, we established a new method to calculate structures of large proteins using PCSs as the only experimental data. The DINGO-PCS algorithm relies on multiple PCS datasets from site-specifically attached metal tags to identify and assemble Smotifs. We benchmarked the method on a set of ten different proteins with 100–220 residues. For nine out of ten targets, the DINGO-PCS algorithm was consistently successful in assembling all of the constituent Smotifs of the targets. Low *Q*-factor values ranging from 0.12 to 0.32 present an objective criterion to judge the quality of the final

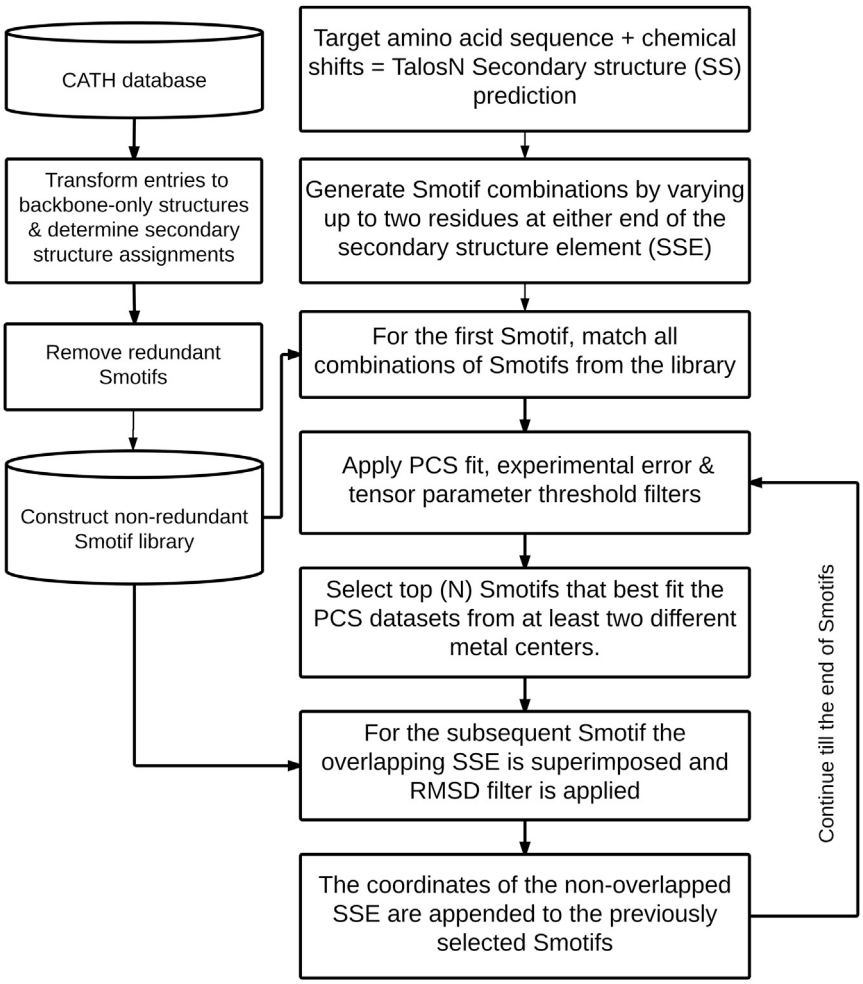


Figure 4. Flowchart Describing the Various Steps Involved in the DINGO-PCS Algorithm

structures with respect to the input PCS data and indicate very good agreement.

EXPERIMENTAL PROCEDURES

Secondary Structure Assignment of the Target Proteins

Accurate secondary structure assignment of the target proteins is the essential first step, as incorrect assignment can alter the number of Smotif definitions. The secondary structure assignments of the target proteins were obtained using the TALOS-N server (Shen and Bax, 2013). TALOS-N uses backbone chemical shifts to predict torsion angles and results in 89% correct secondary structure assignments. To further decrease possible mis-assignments in the secondary structure prediction, the length of every discrete secondary structure element was varied up to two residues at either end.

3D Assembly of Individual Smotifs to Near-Native Geometry as Orchestrated by DINGO-PCS

Figure 4 shows a flowchart of the DINGO-PCS algorithm. The secondary structure assignment of the target protein is used to delineate the Smotifs of the target. The Smotifs are ranked by the amount of PCS data they explain. A search sequence is generated by starting from the Smotif associated with the largest number of PCSs. The next overlapping Smotif is chosen based on the number of PCS data available for it, resulting in either N- or C-terminal extension of the initial Smotif. The Smotif assembly is performed in two stages, (1) identifying the initial Smotif and (2) extending the initial and, subsequently, following Smotifs.

Stage 1: Identifying the Initial Smotif

To allow for errors of secondary structure assignment in the termini, each secondary structure element is used in 11 different permutations (1, 0), (2, 0), (−1, 0), (−2, 0), (0, 1), (0, 2), (0, −1), (0, −2), (−1, −1), (1, 1), and (0, 0), where the first and second numbers, respectively, give the addition/truncation of residues at the N and C termini. For the initial Smotif with two secondary structure elements, this leads to a total of 121 combinations to be tested. All Smotif entries in the combination are exhaustively searched and $\Delta\chi$ tensors are fitted to identify the potentially correct Smotifs. Each $\Delta\chi$ tensor fit requires determining eight tensor parameters, namely the coordinates of the paramagnetic center (x,y,z coordinates), axial and rhombic components of the $\Delta\chi$ tensor, and three Euler's angles (α,β,γ) that define the orientation of the tensor frame relative to the frame of the protein or Smotif. If the metal center is known, the $\Delta\chi$ tensor fit becomes a linear least square-fitting problem that is very fast to compute. To fit the $\Delta\chi$ tensor, DINGO-PCS restricts the locations of the metals by constructing a series of 40 concentric spherical shells with a step size of 1.0 Å and a maximum radius of 40 Å (illustrated in Figure 5A). The metal positions are confined to the spherical shells. The innermost shell is represented by 200 equidistant points and the number of points increases linearly by 200 with the number of shells. These concentric shells are centered on the center of mass of the assembled Smotif.

All entries from the 121 possible secondary structure combinations are exhaustively enumerated, and the $\Delta\chi$ tensors are independently fitted for each metal center. An Smotif is considered a potential hit if it passes through the following three filters. Filter 1: the back-calculated and input PCSs must agree within an error threshold of 0.05 ppm, as calculated by the following score of fit quality:

$$PCS\ Fit\ (F) = \sum^{metal\ centers} \left(\frac{\sum^{N} [PCS_{observed} - PCS_{experimental}]^2}{\sqrt[3]{N} \times (N - K)} \right), \quad \text{(Equation 2)}$$

where N is the total number of PCSs available and K is the total number of fitted parameters per metal center. Filter 2: the magnitudes of the axial and rhombic components of all $\Delta\chi$ tensors are within $\pm 150 \times 10^{-32}$ m^3. These boundaries aid in the filtering of incorrect Smotifs, especially when the PCSs are small or few PCSs are available, which can be fitted by unrealistic $\Delta\chi_{ax}$ and $\Delta\chi_{rh}$ parameters. The fixed tensor magnitudes also help to exclude non-ideal tensor magnitudes being included. Filter 3: the $\Delta\chi$ tensor fits from at least two metal centers pass the filters 1 and 2.

All the hits are ranked according to their total PCS fit score, F, and the top 100 hits (or a user-defined number of hits) are taken to the next stage. These hits are screened for sequence redundancy and only Smotifs of non-redundant amino acid sequence are taken further.

Stage 2: Extending the Smotif Assembly

The next Smotif is the Smotif that (1) can extend the initial Smotif at either end and (2) is characterized by the largest number of PCSs available for the overlapping secondary structure segment from any combination of two metal tags. If Smotifs at either end of the initial Smotif are characterized by the same

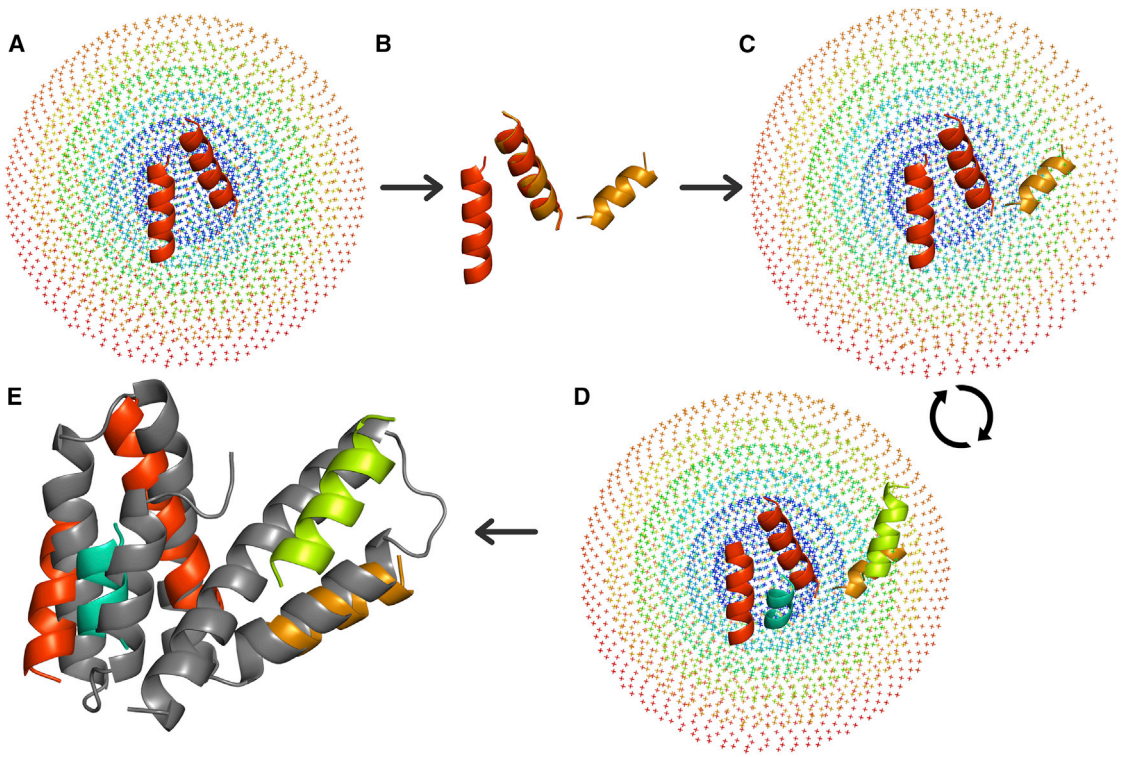


Figure 5. Schematic Illustration of Smotif Assembly by the DINGO-PCS Algorithm, Depicted for Target B

(A) The first Smotif is selected by positioning paramagnetic metal centers on concentric shells centered at the center of mass of the Smotif, and scanning for the Smotif and metal positions that best fulfill the PCS data.

(B) The next Smotif must first pass an RMSD filter. The remaining Smotifs are trialed by appending the coordinates of the non-overlapping secondary structure element to the previous Smotif assembly.

(C) The assembly is evaluated for the best $\Delta\chi$ tensor fit as in (A).

(D) Steps (B) and (C) are repeated for the all Smotifs in the target.

(E) Comparison of the assembly that best fits the PCS data to the native protein.

number of PCSs, the direction of chain growth is decided by the number of PCSs available for the next Smotif. The four-step approach described underneath is followed to assemble the sequence of Smotifs.

Step 1: In this step, the length of the secondary structure element shared between the previous Smotif and the current Smotif is fixed, while the other secondary structure element in the current Smotif is varied by 11 possible combinations.

Step 2: By definition, Smotifs come in overlapping pairs, and this enables us to filter out non-overlapping Smotifs without evaluating them for PCS fitting. No RMSD greater than 2.0 Å is accepted for the overlapping secondary structure element. The Newton-Raphson quaternion-based RMSD calculation algorithm (Liu et al., 2010), which is capable of calculating 7,000 RMSDs/s, is utilized in this step for rapid filtering of Smotifs. The surviving Smotifs are taken to the next step.

Step 3: In this step, the newly identified Smotif is translated to the coordinate frame of the previous Smotif and the coordinates of the common secondary structure element in the current Smotif are discarded. The coordinates and PCSs of the non-overlapping secondary structure element are appended to the previous Smotif. The assembly is filtered for any backbone clashes with other secondary structure elements. Those that pass this filter are further propagated to the next step.

Step 4: In this final step, a new set of concentric spherical grid points is generated as described in the previous section (Stage 1) to restrict the metal position and fit new $\Delta\chi$-tensor parameters that also account for the newly added PCSs. The three filters described for Stage 1 are reapplied, and new hits that satisfy all three filters are propagated for the next round of selection of Smotifs. The process is repeated until all Smotifs of the target protein have been assembled. The schematic representation of Smotif assembly is illustrated in Figure 5. DINGO-PCS is a genetic algorithm, where a population of Smotifs is evaluated and propagated in each stage. If no Smotifs survive the filters in any given stage, the assembly is terminated and marked as incomplete.

The completed Smotif assemblies are ranked based on their quality of PCS fit (Equation 2), and the best-fitting Smotif assembly is reported as the final model. The PCS quality can vary dramatically from the first-ranked Smotif assembly to the second-ranked assembly. Therefore only the best-ranked model is reported.

Even if PCSs are available from four different metal centers, they are not utilized all at once. In the first few steps, while less than 25% of the Smotifs assembly of the target has been completed, the algorithm accepts assemblies that satisfy the PCS data from at least two metal tags. For 25%–75% completion of the assembly, datasets from at least three tags need to be satisfied, with higher ranking given to Smotifs that satisfy PCS data from four tags. From 75% of the assembly onward, the data from all four tags are expected to be satisfied.

PCS Data

Experimental PCS Data

At present, there are only two proteins for which PCS datasets from four different metal centers have been published. These are pSRII, which is a 7-TM α-helical integral membrane protein containing 218 residues (Gautier et al., 2010; Crick et al., 2015) and the C-terminal domain of the ER protein 29 (ERp29-C), which contains 106 residues (Yagi et al., 2013a).

pSRII is a good example for the DINGO-PCS algorithm. The PCSs for this protein were obtained using C2 lanthanide tags (Graham et al., 2011) ligated

to the four different single-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 reported with Dy^{3+}, Yb^{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 (Crick et al., 2015).

For ERp29-C, 212 PCSs have been reported for Tb^{3+} and Tm^{3+} at four different sites (Yagi et al., 2013a), using iDASH tags (Swarbrick et al., 2011; Yagi et al., 2013b) ligated to the mutants C157S/S200C/K204D, C157S/A218C/A222D, and C157S/Q241C/N245D, and the C1 tag (Graham et al., 2011) ligated to the wild-type protein.

Simulated PCS Data

As no experimental PCS data are available for the other benchmark proteins, datasets were generated by mimicking real experimental conditions, computationally grafting the C2 tag (Graham et al., 2011) 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 the 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 only for Dy^{3+}, Tb^{3+}, Tm^{3+}, and Yb^{3+}, using the $\Delta\chi_{ax}$ and $\Delta\chi_{rh}$ values determined by fitting the PCS data measured for the L56C mutant of pSRII (Crick et al., 2015) to the pSRII crystal structure (Royant et al., 2001). PCS data were generated only for the backbone amide protons using PyParaTools (Stanton-Cook et al., 2011). No PCSs were attributed to spins within 12 Å from the metal centers to account for the loss of signal due to PREs. A random error of ±0.04 ppm, which is twice the SD obtained in the $\Delta\chi$ tensor fits for pSRII, was added to all PCS data. PCSs larger than ±1.4 ppm were deleted, as they were not observed in the experimental data for pSRII. To mimic the sparseness observed in experimental datasets, PCSs were randomly deleted from each of the datasets until the total coverage was no more than 60%. In total, the four metal centers, each carrying four different lanthanide metals, resulted in 16 datasets.

Generating Smotif Libraries

A custom Smotif library was designed and generated using the CATH 3D structural database (Sillitoe et al., 2015). Specifically, we used the CATH database 4.0. Within the database, the S100 dataset, which exclusively contains 3D structures without sequence redundancy, was utilized to generate the Smotif libraries. The dataset contained 63,864 domain files in PDB format. The program STRIDE (Frishman and Argos, 1995) was used to define the secondary structure elements for all CATH domains. The secondary structure elements were represented by one of two letters, E for β strands and H for helices, including α helices, 3_{10} helices, and the Π helix. For building the Smotif database, the 3D coordinate files were further simplified by removing all side-chain atoms, while backbone amide hydrogens were added to all domains using the pdb2gmx program from the Gromacs software package (Van Der Spoel et al., 2005). Proline residues in the Smotifs were modified by adding a pseudo-hydrogen to the backbone amide, to avoid the loss of data, if the corresponding residue in the target protein is a non-proline residue. The custom Smotif database was then constructed by ensuring that each secondary structure element in a Smotif is at least five residues in length and that the Smotif is free of loop residues. The omission of loops between secondary structure elements ensures that a large number of entries can be screened for any given Smotif definition. The Smotifs were binned into individual files based on the length of their respective secondary structure elements. For example, all Smotifs with 2 α-helical secondary structure elements consisting of 20 and 30 residues were collected in a file labeled as "hh_20_30.db". Each bin was screened again for the presence of any structurally redundant entries, where redundancy was defined by an RMSD below 0.07 Å between any given pair of Smotifs. Only Smotifs above this cutoff value were retained in the database. The final Smotif library consisted of 2,707 binned files with a total of 435,889 Smotif entries.

Generating All-Atom Models

As the Smotif assemblies contain only coordinates of backbone atoms and no loops, they can easily be transformed into all-atom models. We applied two different methods, using (1) the assembled Smotifs as structural templates for comparative modeling and (2) translating the assembled Smotifs into short fragments in Rosetta format and completing the assembly by the iterative GPS-Rosetta protocol (Pilla et al., 2016).

Comparative Modeling Using Rosetta

Using the assembled Smotifs as structural templates, 1,000 models were built using the RosettaCM protocol (Song et al., 2013). These models were further refined using the Rosetta Relax protocol (Conway et al., 2014), generating 10,000 models. The top five models were selected based on their best fit to PCS data, ranked using Rosetta's inbuilt PCS energy scoring function (Schmitz et al., 2012):

$$PCS\ Energy\ (\mathrm{E}) = \sum^{metal\ centers} \left(\sum^{metals} \sqrt{\sum^{N} \left(PCS_{observed} - PCS_{experimental}\right)^2} \right), \quad \text{(Equation 3)}$$

where N is the total number of available PCSs observed for the lanthanide metal ion.

PCS-Driven Iterative Resampling Using GPS-Rosetta

The assembled Smotif coordinates were translated into 9- and 3-residue fragment libraries in the format used by Rosetta's ab initio structure determination protocol. To maintain diversity in the library, only 20 of the fragments in the native library were replaced with Smotif-derived fragments. Furthermore, additional distance restraints were implemented on residue pairs separated by 3.5–7.5 Å and located in different secondary structure elements. The Smotif-derived fragment libraries along with PCS data were used to run the PCS-driven iterative resampling protocol of GPS-Rosetta for a total of ten iterations (Pilla et al., 2016), generating a total of 24,000 structures. The pairwise distance restraints between residues in different secondary structure elements were only used for the first five iterations and turned off for the following five iterations. The final models were selected from the structures generated in the tenth iteration based on their best fit to PCS data, scored using Equation 3.

SUPPLEMENTAL INFORMATION

Supplemental Information includes two figures and one table and can be found with this article online at http://dx.doi.org/10.1016/j.str.2017.01.011.

AUTHOR CONTRIBUTIONS

Conceptionalization, K.B.P., G.O., and T.H.; Methodology, Software and Validation, K.B.P.; Writing – Review and Editing, K.B.P., G.O., and T.H.; Funding Acquisition, G.O. and T.H.; Supervision, T.H.

ACKNOWLEDGMENTS

Financial support to T.H. and G.O. by the Australian Research Council (DP150100383) 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.

Received: November 24, 2016
Revised: January 17, 2017
Accepted: January 29, 2017
Published: February 16, 2017

REFERENCES

Alva, V., Söding, J., and Lupas, A.N. (2015). A vocabulary of ancient peptides at the origin of folded proteins. Elife *4*, e09410.

Bax, A. (2003). Weak alignment offers new NMR opportunities to study protein structure and dynamics. Protein Sci. *12*, 1–16.

Bertini, I., Luchinat, C., and Parigi, G. (2002). Magnetic susceptibility in paramagnetic NMR. Prog. Nucl. Magn. Reson. Spectrosc. *40*, 249–273.

Chen, W.-N., Nitsche, C., Pilla, K.B., Graham, B., Huber, T., Klein, C.D., and Otting, G. (2016). Sensitive NMR approach for determining the binding mode of tightly binding ligand molecules to protein targets. J. Am. Chem. Soc. *138*, 4539–4546.

Conway, P., Tyka, M.D., DiMaio, F., Konerding, D.E., and Baker, D. (2014). Relaxation of backbone bond geometry improves protein energy landscape modeling. Protein Sci. *23*, 47–55.

Cornilescu, G., Marquardt, J.L., Ottiger, M., and Bax, A. (1998). Validation of protein structure from anisotropic carbonyl chemical shifts in a dilute liquid crystalline phase. J. Am. Chem. Soc. *120*, 6836–6837.

Crick, D.J., Wang, J.X., Graham, B., Swarbrick, J.D., Mott, H.R., and Nietlispach, D. (2015). Integral membrane protein structure determination using pseudocontact shifts. J. Biomol. NMR *61*, 197–207.

Dybas, J.M., and Fiser, A. (2016). Development of a motif-based topology-independent structure comparison method to identify evolutionarily related folds. Proteins *84*, 1859–1874.

Fernandez-Fuentes, N., Dybas, J.M., and Fiser, A. (2010). Structural characteristics of novel protein folds. PLoS Comput. Biol. *6*, e1000750.

Frishman, D., and Argos, P. (1995). Knowledge-based protein secondary structure assignment. Proteins *23*, 566–579.

Gautier, A., Mott, H.R., Bostock, M.J., Kirkpatrick, J.P., and Nietlispach, D. (2010). Structure determination of the seven-helix transmembrane receptor sensory rhodopsin II by solution NMR spectroscopy. Nat. Struct. Mol. Biol. *17*, 768–774.

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

Guan, J.-Y., Keizers, P.H.J., Liu, W.-M., Loehr, F., Skinner, S.P., Heeneman, E.A., Schwalbe, H., Ubbink, M., Siegal, G.D., Löhr, F., et al. (2013). Small molecule binding sites on proteins established by paramagnetic NMR spectroscopy. J. Am. Chem. Soc. *135*, 5859–5868.

Jaroniec, C.P. (2015). Structural studies of proteins by paramagnetic solid-state NMR spectroscopy. J. Magn. Reson. *253*, 50–59.

Keizers, P.H.J., and Ubbink, M. (2011). Paramagnetic tagging for protein structure and dynamics analysis. Prog. Nucl. Magn. Reson. Spectrosc. *58*, 88–96.

Lange, O.F., and Baker, D. (2012). Resolution-adapted recombination of structural features significantly improves sampling in restraint-guided structure calculation. Proteins *80*, 884–895.

Lindert, S., Meiler, J., and McCammon, J.A. (2013). Iterative molecular dynamics-Rosetta protein structure refinement protocol to improve model quality. J. Chem. Theor. Comput. *9*, 3843–3847.

Liu, P., Agrafiotis, D.K., and Theobald, D.L. (2010). Rapid communication fast determination of the optimal rotational matrix for macromolecular superpositions. J. Comput. Chem. *31*, 1561–1563.

Liu, W.M., Overhand, M., and Ubbink, M. (2014). The application of paramagnetic lanthanoid ions in NMR spectroscopy on proteins. Coord. Chem. Rev. *273–274*, 2–12.

Lupas, a N., Ponting, C.P., and Russell, R.B. (2001). On the evolution of protein folds: are similar motifs in different protein folds the result of convergence, insertion, or relics of an ancient peptide world? J. Struct. Biol. *134*, 191–203.

Menon, V., Vallat, B.K., Dybas, J.M., and Fiser, A. (2013). Modeling proteins using a super-secondary structure library and NMR chemical shift information. Structure *21*, 891–899.

Moult, J., Fidelis, K., Kryshtafovych, A., Schwede, T., and Tramontano, A. (2014). Critical assessment of methods of protein structure prediction (CASP) - round x. Proteins *82*, 1–6.

Otting, G. (2010). Protein NMR using paramagnetic ions. Annu. Rev. Biophys. *39*, 387–405.

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

Pilla, K.B., Leman, J.K., Otting, G., and Huber, T. (2015). Capturing conformational states in proteins using sparse paramagnetic NMR data. PLoS One *10*, e0127053.

Pilla, K.B., Otting, G., and Huber, T. (2016). Pseudocontact shift-driven iterative resampling for 3D structure determinations of large proteins. J. Mol. Biol. *428*, 522–532.

Raman, S., Lange, O.F., Rossi, P., Tyka, M., Wang, X., Aramini, J.M., Liu, G., Ramelot, T.A., Eletsky, A., Szyperski, T., et al. (2010). NMR structure determination for larger proteins using backbone-only data. Science *327*, 1014–1018.

Rohl, C.A., Strauss, C.E.M., Misura, K.M.S., and Baker, D. (2004). Protein structure prediction using Rosetta. Methods Enzymol. *383*, 66–93.

Roy, A., Kucukural, A., and Zhang, Y. (2010). I-TASSER: a unified platform for automated protein structure and function prediction. Nat. Protoc. *5*, 725–738.

Royant, A., Nollert, P., Edman, K., Neutze, R., Landau, E.M., Pebay-Peyroula, E., and Navarro, J. (2001). X-ray structure of sensory rhodopsin II at 2.1 Å resolution. Proc. Natl. Acad. Sci. USA *98*, 10131–10136.

Saio, T., Ogura, K., Kumeta, H., Kobashigawa, Y., Shimizu, K., Yokochi, M., Kodama, K., Yamaguchi, H., Tsujishita, H., and Inagaki, F. (2015). Ligand-driven conformational changes of MurD visualized by paramagnetic NMR. Sci. Rep. *5*, 16685.

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

Shen, Y., and Bax, A. (2013). Protein backbone and sidechain torsion angles predicted from NMR chemical shifts using artificial neural networks. J. Biomol. NMR *56*, 227–241.

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

Sillitoe, I., Lewis, T.E., Cuff, A., Das, S., Ashford, P., Dawson, N.L., Furnham, N., Laskowski, R.A., Lee, D., Lees, J.G., et al. (2015). CATH: comprehensive structural and functional annotations for genome sequences. Nucleic Acids Res. *43*, D376–D381.

Söding, J., Biegert, A., and Lupas, A.N. (2005). The HHpred interactive server for protein homology detection and structure prediction. Nucleic Acids Res. *33*, W244–W248.

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

Stanton-Cook, M., Su, X.-C., Otting, G., and Huber, T. (2011). pyParaTools v0.8-alpha. https://github.com/mscook/pyParaTools.

Su, X.-C., and Otting, G. (2010). Paramagnetic labelling of proteins and oligonucleotides for NMR. J. Biomol. NMR *46*, 101–112.

Swarbrick, J.D., Ung, P., Chhabra, S., and Graham, B. (2011). An iminodiacetic acid based lanthanide binding tag for paramagnetic exchange NMR spectroscopy. Angew. Chem. Int. Ed. Engl. *50*, 4403–4406.

van der Schot, G., Zhang, Z., Vernon, R., Shen, Y., Vranken, W.F., Baker, D., Bonvin, A.M.J.J., and Lange, O.F. (2013). Improving 3D structure prediction from chemical shift data. J. Biomol. NMR *57*, 27–35.

Van Der Spoel, D., Lindahl, E., Hess, B., Groenhof, G., Mark, A.E., and Berendsen, H.J.C. (2005). GROMACS: fast, flexible, and free. J. Comput. Chem. *26*, 1701–1718.

Vallat, B., Madrid-Aliste, C., and Fiser, A. (2015). Modularity of protein folds as a tool for template-free modeling of structures. PLoS Comput. Biol. *11*, e1004419.

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

Yagi, H., Maleckis, A., and Otting, G. (2013b). A systematic study of labelling an α-helix in a protein with a lanthanide using IDA-SH or NTA-SH tags. J. Biomol. NMR *55*, 157–166.