Homology modeling of larger proteins guided by chemical shifts

Yang Shen & Ad Bax

We describe an approach to the structure determination of large proteins that relies on experimental NMR chemical shifts, plus sparse nuclear Overhauser effect (NOE) data if available. Our alignment method, POMONA (protein alignments obtained by matching of NMR assignments), directly exploits pre-existing bioinformatics algorithms to match experimental chemical shifts to values predicted for the crystallographic database. Protein templates generated by POMONA are subsequently used as input for chemical shift–based Rosetta comparative modeling (CS-RosettaCM) to generate reliable full-atom models.

High-resolution protein structures, obtained by either X-ray crystallography or NMR spectroscopy, are available for only a small fraction of all known proteins. Computational methods are commonly used to model structures for the remainder. Current protein-structure prediction methods can be broadly separated into two classes: comparative modeling and *de novo* methods. Comparative modeling methods rely on detectable similarity between the query sequence and at least one protein of known structure and can be used to generate models for all proteins in a family using a single representative structure as the starting point[1,2]. *De novo* methods, which use only the amino acid sequence and no structural template, rely on an effective conformation-searching algorithm and good energy functions and can be used to build structural models from scratch. However, because of computational bottlenecks in the sampling of a conformational space that increase exponentially with the number of residues, this method remains restricted to small proteins[3].

NMR chemical shifts of proteins encode important structural information and are obtained at the early stage of any NMR structural study, even for large proteins[4]. It has long been recognized that integration of these data or other very limited, 'sparse' restraints into structural modeling can be highly beneficial[5]. This idea led to the development of powerful and popular *de novo* protein-structure prediction programs, including CHESHIRE[6], CS-Rosetta[7] and CS23D[8], which can generate good-quality, all-atom models for proteins with up to approximately 125 residues and a variety of folds. Supplementing the input chemical shift data with backbone residual dipolar couplings, sparse $^{1}H^{N}$-$^{1}H^{N}$ NOE data[9] or distance restraints extracted from remote homology models[10] can extend the size limit of the *de novo* structure generation approach, but the steeply increasing computational cost with increasing protein size poses serious challenges.

Here we introduce a more direct approach for integrating chemical shift and sparse NOE data into existing, powerful comparative modeling algorithms. We modified the Rosetta comparative modeling method, RosettaCM[11,12], to take advantage of the NMR data when filling in the missing parts and for energetically refining the final structures. Comparative modeling of a protein structure from a sequence principally consists of two steps: first, finding related templates from known structures that have some sequence similarity to the query sequence and optimally aligning the query sequence with the sequence of the templates, and second, generating full 3D models guided by information from the aligned templates.

Best alignment between two sequences is usually obtained through optimization of an alignment scoring function consisting of two components: a matrix of pairwise substitution scores for matching each residue in the database protein to every residue in the query sequence, and a gap penalty function. Once an optimized scoring function has been obtained, efficient dynamic programming is used to search for the optimal alignment between any pair of sequences. Many excellent comparative modeling methods are available, including the widely used MODELLER program[13] and I-TASSER[14].

Backbone torsion angles are encoded in NMR chemical shifts, and even though they are strictly local in character and often not unique, these chemical shifts contain far more information about structural homology than sequence alone. Much of the success of the popular chemical shift–based Rosetta (CS-Rosetta) method stems from the fact that chemical shifts facilitate the finding of structurally homologous peptide fragments in the protein structure database (PDB)[7,15].

The protocol we introduce here relies on a novel chemical shift–guided protein-alignment procedure, POMONA (protein alignments obtained by matching of NMR assignments), followed by adaptation of RosettaCM[12] to take advantage of the available chemical shifts. In the first step in the POMONA-based CS-RosettaCM structure-determination protocol (**Fig. 1a**), experimental $^{13}C^{\alpha}$, $^{13}C^{\beta}$, $^{13}C'$, ^{15}N, $^{1}H^{\alpha}$ and $^{1}H^{N}$ chemical shifts are analyzed to generate a ϕ/ψ probability map for each residue. This map, calculated using the neural network–based TALOS-N program[16], assigns a normalized probability to each 20° × 20° voxel of the Ramachandran map. POMONA uses these residue-specific Ramachandran probability maps to search the PDB for structures that are compatible with these ϕ/ψ probabilities,

Laboratory of Chemical Physics, National Institute of Diabetes and Digestive and Kidney Diseases, US National Institutes of Health, Bethesda, Maryland, USA. Correspondence should be addressed to Y.S. (shenyang@niddk.nih.gov) or A.B. (bax@nih.gov).

RECEIVED 27 OCTOBER 2014; ACCEPTED 25 MARCH 2015; PUBLISHED ONLINE 8 JUNE 2015; DOI:10.1038/NMETH.3437

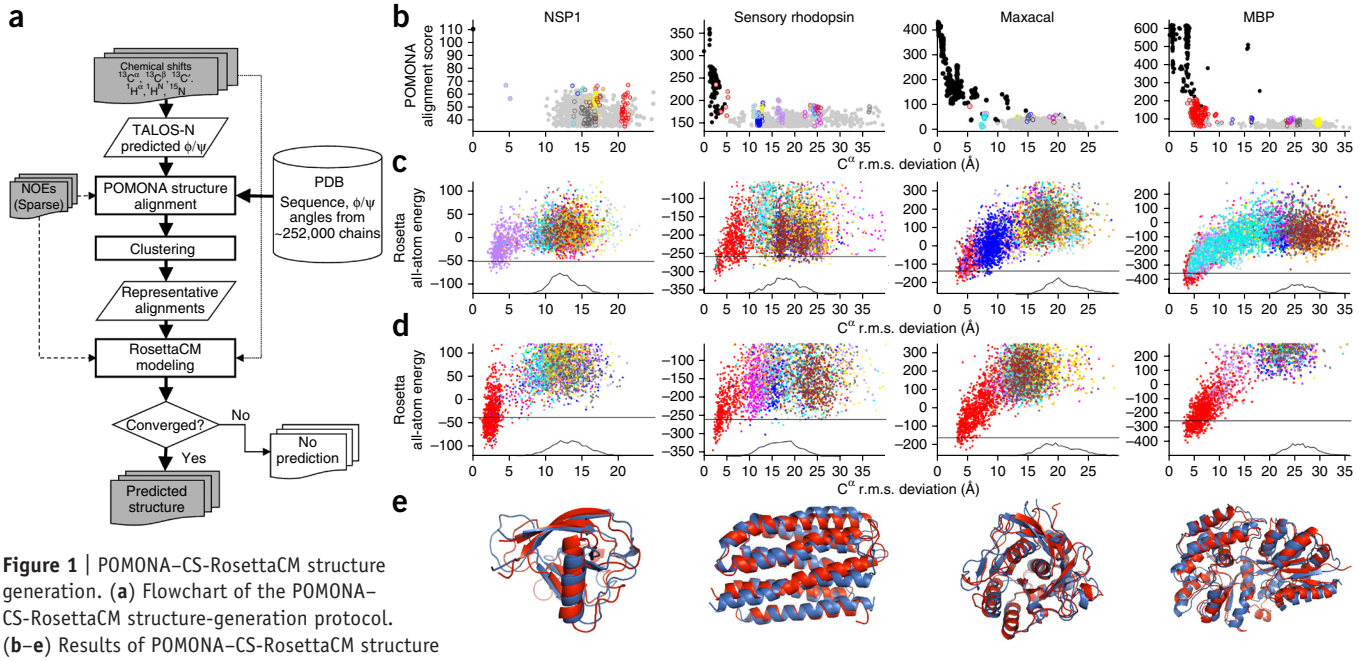

Figure 1 | POMONA–CS-RosettaCM structure generation. (**a**) Flowchart of the POMONA–CS-RosettaCM structure-generation protocol. (**b**–**e**) Results of POMONA–CS-RosettaCM structure generation for four representative test proteins: NSP1, sensory rhodopsin, Maxacal and maltose-binding protein (MBP). (**b**) For each of the test proteins, the POMONA alignment scores (*H′*; equation (10) in Online Methods) of the top 1,000 protein chains in the PDB are plotted versus the C^{α} r.m.s. deviation, calculated over the aligned residues between the query and the database protein. Gray and black dots correspond to sequence identities of <20% and ≥20%, respectively, between the query and database protein. After clustering analysis for the alignments with <20% sequence identity, alignments contained in the ten highest-scoring clusters were marked according to cluster number (red, purple, blue, magenta, light blue, yellow, cyan, orange, gray and brown open circles for clusters 1–10, respectively). Only the two highest-scoring alignments from each of these ten clusters were used as structural templates for CS-RosettaCM modeling. (**c**) ROSETTA all-atom energy, including the experimental chemical shift score, for the CS-RosettaCM models versus their C^{α} r.m.s. deviation relative to the experimental structure. Colors correspond to those of the starting template (**b**). For comparison, the horizontal line and the graph at the bottom of each plot represent the lowest Rosetta all-atom energy and the normalized number of structures, respectively, obtained by CS-Rosetta. (**d**) Same as **c** but for POMONA–CS-RosettaCM modeling including additional sparse ^{1}H-^{1}H NOE data. (**e**) Ribbon models of the lowest-energy CS-RosettaCM structure (red) (calculated without sparse NOEs) superimposed on the corresponding experimental structure (blue).

allowing for gaps and inserts in the residue sequence. After an automated clustering and selection procedure, the representative homologs identified by POMONA are used as structural templates for a modified comparative modeling protocol, based on the RosettaCM program[12], to generate all-atom structures. Further details are presented in the Online Methods.

To evaluate POMONA's accuracy and coverage, we relied on the widely used MaxSub score[17], which ranges from 1.0 for two aligned structures that have a C^{α} r.m.s. deviation of 0 Å for the full length of the query sequence to ~0.0 for sequences that lack detectable similarity. Typically, a MaxSub score greater than ~0.3 is indicative of notable structural similarity (**Fig. 2a** and **Supplementary Fig. 1**).

When one is evaluating the performance of POMONA in identifying suitable homologous structures in the PDB, a key question is, "How many suitable structures exist?" This can be answered with the program DALI, which is designed to identify structurally similar proteins, regardless of residue sequence[18]. A comparison of DALI alignments and POMONA-identified structural homologs was carried out for a set of 16 test proteins

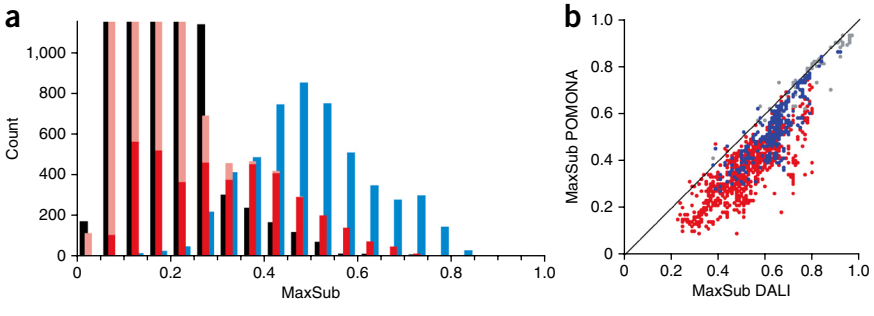

Figure 2 | Comparison of protein-structure alignments obtained by different methods for the 16 proteins listed in **Table 1**. (**a**) Histogram of protein-structure alignment quality, represented by a MaxSub score, for the top 1,000 alignments identified by POMONA (red bars), the sequence alignment method HHsearch (black bars) and the structure alignment method DALI (blue bars). Results are shown only for PDB proteins with <20% sequence identity to the target protein, and DALI and HHsearch results correspond to default thresholds of $Z \geq 2$ and probability ≥ 10%, respectively, used by these programs to identify homologs. The DALI histogram indicates the limit of how well any search program could possibly function. Positive POMONA alignments are taken from the top ten clusters (red bars) within the top 1,000 alignments (red + pink bars), as identified by the highest *H′* score (equation (10) in Online Methods). (**b**) Comparison of alignment quality obtained via DALI and POMONA methods. For each of the positive alignments identified by both DALI and POMONA, the MaxSub scores are compared, with color representing sequence identity to the query protein (gray, ≥30%; blue, 20%–30%; red, <20%) as observed in the DALI alignments. The diagonal line represents the approximate limit for the MaxSub score that could potentially be reached for any pair of proteins.

Table 1 | Performance of POMONA alignment and CS-RosettaCM structure generation for 16 test proteins

Protein name	Size[a]	PDB/BMRB number[b]	Fold	Homologs and alignments		CS-RosettaCM		CS-Rosetta
				DALI[c]	POMONA[d]	r.m.s. deviation$_{mean}$[e]	r.m.s. deviation$_{exp}$[e]	r.m.s. deviation$_{exp}$[e]
NSP1	113	2gdtA/7014	α/β	2/0/2 (0.50)	0.30/0.30	2.18 ± 0.63	3.30 ± 0.73[f]	12.1 ± 1.3
HR2876B	117	2ltmA[b]/18489	α/β	2/4/75 (0.47)	0.41/0.41	2.89 ± 0.72	4.21 ± 0.55	6.41 ± 2.76
YR313A	119	2ltlA[b]/18487	α/β	1/2/52 (0.45)	0.26/0.25	1.60 ± 0.27	3.67 ± 0.45	2.80 ± 0.68[g]
OR36	134	2lciA[b]/17613	α/β	4/5/799 (0.50)	0.36/0.34	2.19 ± 0.73	4.32 ± 0.56	3.05 ± 0.35
OR135	83	2ln3A[b]/18145	α/β	1/1/651 (0.70)	0.52/0.40	1.35 ± 0.49	1.88 ± 0.42	1.21 ± 0.13[f]
HR2876C	87	2m5oA[b]/19068	α/β	4/4/723 (0.57)	0.35/0.33	1.77 ± 0.27	2.24 ± 0.42	1.17 ± 0.20[f,g]
MTH1958	153	1tvgA/6344	β	5/14/147 (0.76)	0.53/0.51	1.30 ± 0.18	2.35 ± 0.17[f]	10.4 ± 4.9
SgR145	173	3merA/16806	α/β	3/43/896 (0.72)	0.66/0.52	2.30 ± 0.58	3.05 ± 0.74	8.2 ± 2.8
Fgf2	125	1basA/4091	β	262/23/449 (0.83)	0.74/0.65	1.06 ± 0.18	1.56 ± 0.19[f]	11.7 ± 1.5
tpx	167	2jszA[b]/15797	α/β	49/376/389 (0.70)	0.69/0.68	1.60 ± 0.23	2.32 ± 0.22[f]	17.7 ± 2.0
YwIE	150	1zggA/6460	α/β	17/66/308 (0.65)	0.69/0.69	1.19 ± 0.18	1.86 ± 0.23[f]	11.0 ± 3.7
FluA	184	1n0sA/5756	β/α	11/38/413 (0.63)	0.51/0.51	2.01 ± 0.49	3.46 ± 0.34[f]	8.5 ± 1.5
Mad2	196	1go4C	α/β	42/10/3 (0.44)	0.13/0.11	12.74 ± 4.45	19.81 ± 1.01	15.8 ± 2.6[g]
Sensory rhodopsin	222	2ksyA[b]/16678	α	23/153/149 (0.64)	0.62/0.62	2.32 ± 0.43	3.09 ± 0.51[f]	17.8 ± 3.5
Maxacal	269	1svnA	α/β	273/79/4 (0.51)	0.50/0.50	3.29 ± 0.57	4.51 ± 0.85[f]	19.4 ± 2.6
MBP	370	1dmbA	α/β	276/31/182 (0.52)	0.52/0.51	2.73 ± 0.50	4.24 ± 0.73[f]	26.3 ± 2.1

BMRB, Biological Magnetic Resonance Data Bank.
[a]Number of residues. [b]The PDB code for proteins with an NMR-derived structure as the reference. [c]Number of alignment hits with sequence identity of ≥30%, 30%–20% and <20%, respectively, and a minimum alignment length of at least 2/3 of the total number of target residues; the highest MaxSub value observed for the alignments with a sequence identity of <20% is listed in parentheses. [d]Highest MaxSub value observed among all top 1,000 POMONA alignments (sequence identity of <20%) and among the up to 20 templates used for subsequent CS-RosettaCM modeling. [e]C^{α} r.m.s. deviation value calculated for all nonflexible residues (as identified by a random coil index order parameter (S^2) ≥ 0.6 (ref. 20)). r.m.s. deviation$_{mean}$ is the C^{α} r.m.s. deviation between the ten lowest-energy models and their mean coordinates. r.m.s. deviation$_{exp}$ is the C^{α} r.m.s. deviation between the ten lowest-energy models (derived using database proteins with sequence identity <20%) and the experimental reference structure. [f]CS-RosettaCM and CS-Rosetta structures that met the acceptance criterion (Online Methods). To convert a calculated r.m.s. deviation value to its corresponding r.m.s. deviation value at 100 residues (r.m.s. deviation$_{100}$, used in our work to evaluate convergence (Online Methods)), r.m.s. deviation$_{100}$ = r.m.s. deviation/(1 + ln(N/100)), where N is the number of residues of the protein. [g]CS-Rosetta models with a lower Rosetta energy than obtained with the POMONA–CS-RosettaCM approach.

with available chemical shift information and representing a diverse set of folds (**Table 1**). This comparison showed that POMONA-identified structural homologs approached the maximum attainable alignment (or MaxSub score) provided by the DALI method (**Fig. 2b**), performing much better than sequence-based alignment by, for example, the powerful HHsearch method[19] (**Supplementary Fig. 2**).

The quality of POMONA alignments roughly correlated with the alignment score (**Fig. 1b** and **Supplementary Fig. 3**). However, there also was considerable scatter in this correlation, which meant that we could not simply use the top POMONA alignments as starting templates for CS-RosettaCM. Instead, we found it important to generate a diverse pool of structure templates by subjecting the top-scoring alignments to a cluster analysis and retaining only the two top-scoring alignments in each of the first ten clusters (Online Methods). For most of our 16 test proteins, the highest MaxSub score observed for this subset of up to 20 members was comparable to that obtained for the top 1,000 positive alignments (**Table 1**). For all but one of the 16 proteins, the best alignment in the selected representative alignments had a MaxSub value in the range of 0.25–0.69, making them useful structural templates for structure generation. Only for protein Mad2 did POMONA fail to find a suitable template. DALI found three suitable templates for Mad2 in the database, but all contained large gaps (>100 residues), preventing their identification by POMONA.

For four representative cases, we plotted the database proteins corresponding to the top 1,000 POMONA-derived alignment scores against their C^{α} r.m.s. deviation relative to the experimental reference structure (**Fig. 1b**). For comparison, POMONA hits for more homologous proteins (≥20% sequence identity) were included in the plot, but these were not used in our study, as they typically can be identified by standard homology search programs.

When the two highest-scoring members of each cluster were subjected to the CS-RosettaCM protocol, a clear correlation was seen between the lowest total all-atom energy reached for each cluster and the C^{α} r.m.s. deviation (**Fig. 1c**). Even though for all four proteins the highest POMONA alignment scores were comparable between the top clusters, the clusters that had the lowest C^{α} r.m.s. deviation relative to the native structure refined to lower total energy during CS-RosettaCM modeling. Correspondingly, the lowest-energy CS-RosettaCM models provided the best match to the query protein. However, because it is by no means guaranteed that a correct solution can be found, especially when there are no proteins with a similar fold in the database (e.g., as for Mad2, mentioned above), it is useful to compare the total energy with what can be achieved with the standard CS-Rosetta protocol. CS-Rosetta will typically fail for large proteins, and a requirement for accepting a CS-RosettaCM structure therefore is that the total energy, including the chemical shift scoring term, falls well below the lowest values obtained by CS-Rosetta. A second requirement for acceptance is that the ten lowest-energy structures have converged—that is, they cluster within a C^{α} r.m.s. deviation normalized for 100 residues of ≤2.5 Å from their average. We used both requirements to inspect all 16 proteins tested in our study (**Table 1**).

Immediately after backbone resonance assignment, it is usually straightforward to rapidly assign a limited number of unambiguous backbone $^{1}H^{N}$-$^{1}H^{N}$ NOEs. These sparse NOEs can be exploited both for guiding POMONA alignment and as restraints during CS-RosettaCM modeling. To evaluate their utility, we generated two sets of such artificial H^{N}-H^{N} NOE distance restraints by randomly selecting N/10 such distances from the total set that were ≤5 Å in the experimental structure and at least five residues apart in sequence, where N is the total number of residues in the protein. In practice, a somewhat larger number of such

NOEs is often obtained, particularly in work with perdeuterated proteins[4]. The inclusion of sparse NOEs enables POMONA to find improved alignments, resulting in better convergence and lower energies during the subsequent CS-RosettaCM modeling stage (**Fig. 1d** and **Supplementary Table 1**).

Whereas conventional CS-Rosetta approaches its convergence limits for proteins larger than about 100 residues, CS-RosettaCM remains robust in generating converged results, largely because it is inherently a comparative modeling method. Note, however, that the POMONA–CS-RosettaCM protocol is not aimed at reaching maximum convergence; rather, the clustering approach used by POMONA emphasizes diversity in the input templates to avoid false convergence to a wrong solution. As a result, the convergence rate for small structures can actually be higher for standard CS-Rosetta than for our new protocol.

For proteins larger than approximately 20 kDa, standard protein NMR structure determination typically remains quite labor intensive, even though chemical shift assignment and the collection of amide-amide NOEs are relatively straightforward. Considering that similar structures are already present in the PDB for ~90% of the newly deposited structures, we believe that the POMONA–CS-RosettaCM approach could dramatically reduce the workload required for protein-structure determination while extending the size of proteins that can readily be studied by NMR. The approach will fail, however, when no adequate structural template exists in the PDB, or when the only good potential templates have large alignment gaps.

Finding suitable templates is an efficient process that can be completed in a matter of hours, but subsequent CS-RosettaCM modeling is far more computationally intensive. Nevertheless, when suitable input templates are used, CS-RosettaCM does not suffer from the combinatorial explosion that restricts conventional Rosetta and CS-Rosetta applications. For large, multidomain proteins, it is important to note that the NMR chemical shifts do not contain information on relative domain orientation or position, and that this information stems strictly from the PDB template used for modeling. However, the measurement of residual dipolar couplings is often straightforward for larger systems and can be readily integrated into the modeling procedure to resolve such issues.

The POMONA software and server are at http://spin.niddk.nih.gov/bax/software/POMONA.

METHODS

Methods and any associated references are available in the online version of the paper.

Note: Any Supplementary Information and Source Data files are available in the online version of the paper.

ACKNOWLEDGMENTS

This work was funded by the Intramural Research Program of the NIDDK, US National Institutes of Health (NIH). We thank Y. Song, N. Sgourakis and D. Baker for help and advice on the use of RosettaCM. We also gratefully acknowledge use of the NIH high-performance computational Biowulf Linux cluster.

AUTHOR CONTRIBUTIONS

Y.S. and A.B. designed the methods and protocol and wrote the manuscript. Y.S. developed the code, optimized the parameterization of the protocol and analyzed the resulting data.

COMPETING FINANICAL INTERESTS

The authors declare no competing financial interests.

1. Martí-Renom, M.A. *et al. Annu. Rev. Biophys. Biomol. Struct.* **29**, 291–325 (2000).
2. Pieper, U. *et al. Nucleic Acids Res.* **37**, D347–D354 (2009).
3. Das, R. & Baker, D. *Annu. Rev. Biochem.* **77**, 363–382 (2008).
4. Tugarinov, V., Choy, W.Y., Orekhov, V.Y. & Kay, L.E. *Proc. Natl. Acad. Sci. USA* **102**, 622–627 (2005).
5. Bowers, P.M., Strauss, C.E.M. & Baker, D. *J. Biomol. NMR* **18**, 311–318 (2000).
6. Cavalli, A., Salvatella, X., Dobson, C.M. & Vendruscolo, M. *Proc. Natl. Acad. Sci. USA* **104**, 9615–9620 (2007).
7. Shen, Y. *et al. Proc. Natl. Acad. Sci. USA* **105**, 4685–4690 (2008).
8. Wishart, D.S. *et al. Nucleic Acids Res.* **36**, W496–W502 (2008).
9. Raman, S. *et al. Science* **327**, 1014–1018 (2010).
10. Thompson, J.M. *et al. Proc. Natl. Acad. Sci. USA* **109**, 9875–9880 (2012).
11. Misura, K.M.S., Chivian, D., Rohl, C.A., Kim, D.E. & Baker, D. *Proc. Natl. Acad. Sci. USA* **103**, 5361–5366 (2006).
12. Song, Y. *et al. Structure* **21**, 1735–1742 (2013).
13. Webb, B. & Sali, A. *Curr. Protoc. Bioinform.* **47**, 5.6.1–5.6.32 (2014).
14. Xu, D., Zhang, J., Roy, A. & Zhang, Y. *Proteins* **79**, 147–160 (2011).
15. Vernon, R., Shen, Y., Baker, D. & Lange, O.F. *J. Biomol. NMR* **57**, 117–127 (2013).
16. Shen, Y. & Bax, A. *J. Biomol. NMR* **56**, 227–241 (2013).
17. Siew, N., Elofsson, A., Rychiewski, L. & Fischer, D. *Bioinformatics* **16**, 776–785 (2000).
18. Holm, L. & Sander, C. *J. Mol. Biol.* **233**, 123–138 (1993).
19. Söding, J. *Bioinformatics* **21**, 951–960 (2005).
20. Berjanskii, M. & Wishart, D.S. *Nat. Protoc.* **1**, 683–688 (2006).

ONLINE METHODS

Measurement of local structure similarity. The optimal alignment between two protein sequences typically is based on a residue-substitution score for all residue pairs of the two sequences. Such substitution scores, which normally are derived from the amino acid similarity scores, are then used for guiding the alignment procedure to find a set of aligned residues along two sequences that have an optimal overall alignment score. Unlike in sequence-based alignment, POMONA aims to align residues of a query protein with known NMR chemical shifts to residues of a database protein with known structure. Structural information encoded in the NMR chemical shifts of the query protein, specifically the ϕ/ψ backbone torsion angles and the secondary structure predicted by TALOS-N[16], is much more definitive than the amino acid type alone in searches for structural similarity between query and database proteins. Therefore, these backbone torsion angles and the secondary structure derived from chemical shifts are used as the main terms when substitution scores are derived for the alignment procedure.

In POMONA, a substitution score $S(i,j)$ between residue i in the query protein and residue j in the database protein is defined as

$$S(i,j) = w_{\text{torsion}} \sum_{n}^{-1,0,1} \frac{D_{i+n,k(j+n)} - \langle D_{i+n} \rangle}{\sigma(D_{i+n})} + w_{\text{residue}} B(A_i, A_j) + w_{\text{ss}} \sum_{n}^{-1,0,1} P(\text{SS}_{i+n}, \text{SS}_{j+n}) \quad (1)$$

$S(i, j)$ contains three terms: (1) The ϕ/ψ fitness score, which has a weighting factor w_{torsion}, reflects how well the angles of query residue i match to the observed ϕ/ψ angles of database residue j. Here, $D_{i,k}$ (k = 1–324) is the TALOS-N–predicted density of voxel k in the 324-voxel ϕ/ψ density map of query residue i, and $k(j)$ is the index number in the 324-voxel Ramachandran map that corresponds to the ϕ/ψ angles of residue j of the database protein. One calculates the ϕ/ψ fitness score from $D_{i,k(j)}$ by first subtracting the average of the predicted densities $\langle D_i \rangle$ and then normalizing according to the s.d., $\sigma(D_i)$, of the predicted densities, which then represents the likelihood that the ϕ/ψ torsion angles of residues i match those of j. (2) The amino acid similarity score between residue i (of amino acid type A_i) and residue j (of type A_j), $B(A_i,A_j)$, is taken from the BLOSUM62 matrix[21]. (3) The third term is the secondary-structure similarity score between the TALOS-N–predicted three-state secondary structure SS_i (H, E and L, respectively) for residue i and the observed secondary structure SS_j (as assigned by the program DSSP)[22] for residue j.

$$P(\text{SS}_i, \text{SS}_j) = \begin{cases} \text{conf}(i) & \text{SS}_i = \text{SS}_j \\ -\text{conf}(i) & \text{SS}_i \neq \text{SS}_j \end{cases} \quad (2)$$

where conf(i) is the confidence of the TALOS-N–predicted secondary structure SS_i.

Note that terms 2 and 3 are the principal terms used in conventional, sequence-based homology search procedures. In our search, term 2 has a very low weight factor, and term 3 is derived from experimental chemical shifts, which have been shown to considerably increase the accuracy of predicted secondary structures[16].

As seen in equation (1), our local structure-similarity score between residues i and j includes terms for its two immediate neighbors, that is, between residues i – 1 and j – 1 and between residues i + 1 and j + 1. The weights w_{torsion} and w_{SS} of these terms have been optimized empirically, together with other parameters used by the POMONA alignment method, such that the calculated substitution scores $S(i,j)$ range from −2.0 to 3.0. The maximum contribution to $S(i,j)$ from residue-type similarity (term 2 in equation (1)) is less than approximately 10% when chemical shifts are available. For query residues that lack chemical shifts, only the sequence similarity and secondary-structure matching term in equation (1), with a comparable weight, are used to calculate an $S(i,j)$ score, which is then scaled to the same range of −2.0 to 3.0.

Gap penalty function used for protein alignment. An important element in evaluations of the alignment of two protein sequences is the concept of alignment gaps, or the presence of insertions or deletions in the sequence of either protein, which are counted as a penalty to the overall alignment score. For POMONA, a varied gap penalty (VGP) function[23] with a conformation-specific form is used as outlined below.

For opening a gap that extends from positions i' to i in the sequence of the query protein (referred to as the sequence block) and from j' to j in the sequence of the database protein of known structure (referred to as the structure block), as illustrated below for an example of aligned segments,

```
                 i'        i
Sequence block:  KT-------LTG
Structure block: EKAPKARIG-DL
                 j'        j
```

the varied gap penalty function $G(i, j, i', j')$ of Madhusudhan *et al.*[23] is used.

$$G(i,j,i',j') = \begin{cases} 0 & l = 0 \text{ and } l' = 0 \\ R \times u + (l + l') \times v & \text{otherwise} \end{cases} \quad (3)$$

$$l = i - i' - 1 \quad (4)$$

$$l' = j - j' - 1 \quad (5)$$

$$R = 1 + W_{\text{HS}} \times [\text{HS}(j,j') + \text{HS}(i,i')] + W_d \times D(j,j') \quad (6)$$

where l and l' are the lengths of the insertions in the sequence and the structure blocks, respectively, v is the gap-extension penalty, u is the gap-opening penalty, and R is the function that modulates the gap-opening penalty depending on the secondary structure at the position of the insertion in the sequence and structure blocks. R is at least 1 and can be larger to make the opening of gaps more difficult in the following circumstances: within elements of regular secondary structure (helices or strands), and between two spatially distant database residues. W denotes the weight of various properties in R. HS(i,i') and HS(j,j') are the consensus values for helical (or β-strand) content at position i in the sequence block and at position j in the structure block, respectively. The binary value of HS(j,j') is either 1 or 0 depending on whether

the conformation from positions j to j' is helical (or β-strand). HS(i,i') is a similar measure but is based on the TALOS-N–predicted secondary structure for query residues i' to i. $D(j,j')$ is the value derived from the distance of the two database residues spanning the gap.

$$D(j,j') = \max(0, d - d_0)^{\gamma} \quad (7)$$

where d is the distance between C^{α} atoms at positions j' and j in the structure block and d_0 is an empirical constant of 6.5 Å. For less than 6.5 Å, there is no increase in the gap-opening penalty. The exponent γ was optimized by trial and error, and best values for all parameters ($u = 3.0$, $v = 0.3$, $W_{HS} = 1.0$, $W_d = 2.0$, $d_0 = 6.5$ and $\gamma = 2.0$) were obtained by means of a grid search.

Protein-alignment algorithm. The problem of finding the optimal alignment of two amino acid sequences has been extensively studied and most commonly is solved by means of a dynamic programming algorithm[24,25]. POMONA essentially uses the standard Smith-Waterman dynamic programming algorithm[23] to find the best alignment between a query protein with ϕ/ψ angle information derived from chemical shifts and a database protein of known structure. Specifically, given a query protein and a database protein of sequence lengths M and N, a substitution scoring matrix S of dimensions $M \times N$ is constructed. Each element of this scoring matrix $S(i,j)$ (equation (1)) is derived from the local structural similarity between residue i in the query protein and residue j in the database protein. The aim is to align residues with matching local structure in the two proteins while optimizing the overall alignment score, which is a sum of the substitution scores of all aligned residue pairs (also referred to as equivalent residues) and gap penalties for residues lacking an equivalent residue in either sequence. The recursive dynamic programming equation used here for the local alignment of the two proteins is

$$H(i,j) = \mathrm{Max}_{M+1 \geq i' > i, N+1 \geq j' > j}[H(i',j') + G(i,j,i',j')] + S(i,j) \quad (8)$$

with the initial conditions for the recursion defined by $H(M+1, j) = 0$ and $H(i, N+1) = 0$, where M and N again are the sequence lengths of the query and the database protein, G is the VGP function (equation (3)), and $S(i, j)$ is the residue substitution score for residues i and j in the query and the database proteins, respectively (equation (1)). The dynamic programming maximum scoring matrix H is calculated for $i = M + 1$ to 1 and $j = N + 1$ to 1. For each position $[i, j]$ in H, all previously iterated positions $[i', j']$, with $i' = [i + 1{:}M]$ and $j' = [j + 1{:}N]$, are evaluated for a maximum value based on the previously calculated $H(i', j')$ value for position $[i', j']$, using a gap penalty $G(i, j, i', j')$ for opening a gap between positions $[i, j]$ and $[i', j']$. After its residue substitution score $S(i, j)$ has been added, this maximum value is assigned to the current position as score $H(i, j)$. After the calculation of all elements of the H matrix, the largest element, referred to as max(H), will correspond to the optimal alignment score. One can obtain the residue equivalence assignments by backtracking in matrix H, starting from the element with the max(H) score and ending with the first element of zero value[24]. Equivalent residues in this optimal alignment are further evaluated in terms of fitness between the experimental secondary chemical shifts (of the query residues) and those predicted by SPARTA+[26] (for the database residues) in terms of a χ^2 value.

$$\chi^2_{CS} = \sum_k \sum_{[i,j]} (\delta^{obs}_{k,i} - \delta^{pred}_{k,j})^2 / \sigma^2_{k,j} \quad (9)$$

where $\delta^{pred}_{k,j}$ is the backbone chemical shift predicted by SPARTA+ (k = $^{13}C^{\alpha}$, $^{13}C^{\beta}$, $^{13}C'$, ^{15}N, $^{1}H^{\alpha}$ and $^{1}H^{N}$) for a given database residue j, which is aligned to query residue i with experimental chemical shift $\delta^{obs}_{k,i}$, and $\sigma_{k,j}$ is the uncertainty of $\delta^{pred}_{k,j}$ reported by SPARTA+. This χ^2_{CS} value, after being scaled by a factor $c = 1/30$, is then added to the optimal alignment score as a penalty to derive a final alignment score for any given alignment of

$$H' = \max(H) - c \times \chi^2_{CS} \quad (10)$$

Structure alignment with additional NOE data. Some types of NOE data, in particular H^N-H^N NOEs, often can be obtained relatively easily and unambiguously once the backbone amide signals have been assigned, even for large perdeuterated proteins. Unfortunately, there is no straightforward method for directly integrating such sparse NOE distance information into the Smith-Waterman algorithm. However, the typically very sparse NOE data can be useful to aid the above-described chemical shift–guided POMONA protein-alignment scheme if one pre-filters possible solutions on the basis of these distance constraints and subsequently evaluates these possible matches using the above-described algorithm to generate optimally aligned sequences. The NOE-guided part corresponds to the general problem of finding the optimal alignment of protein-structure distance matrices (or protein contact maps)[18,27,28], as both the NOEs detected for the query protein and the actual distances measured for the database protein can be converted to contact maps.

Here we used the method of Wohlers *et al.*[28] to find the optimal overlap between two contact maps derived from the query and the database protein. For the query protein with a NOE list (NOE), a contact map X of size $M \times M$ is constructed.

$$X(i,i') = \begin{cases} 1 & \text{if}(i,i') \subset \text{NOE} \\ 0 & \text{otherwise} \end{cases} \quad (11)$$

where i and $i' = [1, \ldots, M]$ and M is the size of the query protein. For the database protein, an analogous contact map Y of size $N \times N$ is constructed.

$$Y(j,j') = \begin{cases} 1 & d(j,j') \leq 6.5\ \text{Å} \\ 0 & \text{otherwise} \end{cases} \quad (12)$$

where j and $j' = [1, \ldots, N]$, N is the size of the database protein, and $d(j, j')$ is the actual H^N-H^N distance between residues j and j' in the database protein. Contacting residues i and i' with $X(i, i') = 1$ and j and j' for $Y(j, j') = 1$ are stored as lists x and y, respectively. Optimal alignment then corresponds to finding a maximum set of matching $[i_k, j_k]$ pairs ($k = 1$ to L, $i_k \subset x$ and $j_k \subset y$, where L is the lesser of the two numbers of contacting residues, usually the size of list x) between the pairs of contacting residues

doi:10.1038/nmeth.3437

in the query and database proteins. The largest set of common contacts is based on the objective function

$$f(X,Y) = \max\left\{\sum_{1\le r\le L}\ \sum_{1\le s\le L, r\ne s} C([i_r,j_r],[i_s,j_s])\right\} \quad (13)$$

$$C([i_r,j_r],[i_s,j_s]) = \begin{cases} 1 & X(i_r,i_s)=1 \text{ and } Y(j_r,j_s)=1 \\ 0 & \text{otherwise} \end{cases} \quad (14)$$

After the optimal match between the query and the database contact maps has been found, the second step of the structure alignment, based on chemical shift data, is restricted to the regions identified by this optimal contact map. For the query and the database proteins with a set of optimally aligned contacting residues $[i_k, j_k]$, where $k = 1$ to L, $i_k \subset x$ and $j_k \subset y$, $i_k < i_{k+1}$ and $j_k < j_{k+1}$, the query and the database sequences are divided into $L - 1$ fragment pairs, each of which has a range from i_k to i_{k+1} and from j_k to j_{k+1}, respectively. The above-described POMONA protein-alignment scheme is then applied for each possible pair of query fragment (i_k, i_{k+1}) and database fragment (j_k, j_{k+1}) [$k = 1$ to $L - 1$], now using equation (1) as the scoring term. The final overall alignment is then obtained through combination of each of the 'subalignments', and the final alignment score is taken as the sum of the POMONA alignment scores from each of the subalignments, augmented by the penalizing, scaled chemical shift fitness score χ^2_{CS} (equation (9)).

Training and testing of POMONA. We obtained values for the parameters used by POMONA iteratively by evaluating the output results for a set of 16 test proteins of varying size and fold complexity (**Table 1**). POMONA was used to find optimal alignment between a test protein and each of approximately 252,000 protein chains in the PDB. POMONA initially retained the 1,000 PDB protein chains that exhibited the highest alignment scores. We performed parameter optimization of POMONA iteratively by monitoring the top 1,000 selected proteins in terms of (1) the ratio of the real structural homologs, as identified by the DALI structure alignment method[18] with the actual structure of the target protein, and (2) the accuracy of the POMONA-identified alignment to the target protein, expressed in terms of a coordinate r.m.s. deviation value calculated between the C^α atoms of the equivalent residues in the target and database protein.

Evaluation of POMONA structure alignment. We evaluated the accuracy and coverage achieved by POMONA by using the MaxSub score[17]. We calculated the MaxSub score for two aligned structures (i.e., the query and database proteins) by first identifying the maximum substructure for which the distances between equivalent residues of two structures after superposition were below a threshold value of 3.5 Å and then computing a normalized score of $\Sigma[1/(1 + (d_i/3.5)^2)]/N$, where d_i are the distances between equivalent C^α pairs of two structures in the maximum substructure (after best-fit superposition of the C^α pairs in the maximum substructure) and N is the total length of the query sequence. The spatial information of the aligned structures outside the maximum substructure was not taken into account. MaxSub scores range from 1.0, for perfect alignment, to near zero for sequences lacking structural similarity. Two aligned structures with a 0-Å C^α r.m.s. deviation for half of the query sequence length and two aligned structures with a C^α r.m.s. deviation of ~3.5 Å for the full length of the query sequence will have the same MaxSub value of 0.5, and a score greater than ~0.3 is usually indicative of meaningful structural similarity (**Fig. 2a** and **Supplementary Fig. 1**). A detailed evaluation of the performance of POMONA structure alignment is included in the **Supplementary Results**.

Clustering and selection of POMONA alignments. Among the top 1,000 alignments identified by POMONA for any given query protein, there will be many that are very similar to one another. Therefore, before using these proteins as input for the time-consuming RosettaCM comparative modeling, it is useful to separate this set into a much smaller number of distinct clusters (typically ten) and then use only the two best-scoring (see equation (10)) models in each cluster as RosettaCM input. Specifically, a hierarchical clustering procedure is used to group the top 1,000 database protein chains, using the normalized C^α r.m.s. deviation as a metric. The normalized C^α r.m.s. deviation between two database protein chains is calculated only over residues that are commonly aligned to a residue in the query protein (i.e., that do not correspond to inserts or gaps). Subsequently the C^α r.m.s. deviation is normalized to the r.m.s. deviation_{100} value[29]. A single-linkage algorithm is used for generating the clusters with a cutoff of C^α r.m.s. $\text{deviation}_{100} \le 4$ Å, and results are sorted by the highest alignment score observed in each cluster. The ten clusters with highest alignment scores are retained, and the top two alignments (or one, if there is only a single member in the cluster) are selected as representatives from each of the first ten clusters. Therefore, up to 20 representative alignments are selected from the first ten clusters, and these are used to prepare a pool of structural templates for the subsequent RosettaCM modeling procedure.

Structure generation using CS-RosettaCM. The recent RosettaCM protocol[12] offers a powerful comparative modeling module within the Rosetta software suite for generating accurate protein models. The inputs to RosettaCM comprise (1) sequence alignments between the query protein and database proteins that serve as structural templates and (2) standard Rosetta *de novo* modeling fragments, needed to model the unaligned regions and to explore deviations from the templates in the aligned regions. In our protocol, RosettaCM is used to build 3D protein models, starting from the up to 20 structural templates identified by POMONA.

The generation of complete, all-atom models involves three steps. First, RosettaCM assembles protein backbone topologies by recombining the aligned segments of the query protein and the database template in Cartesian space while building the unaligned regions *de novo* in torsion angle space. This process uses long fragments (corresponding to secondary-structure elements) derived from all template inputs and CS-Rosetta *de novo* fragments (with sizes of three and nine residues), respectively. In the standard RosettaCM implementation, these *de novo* fragments are selected on the basis of residue sequence, whereas in our work they were picked on the basis of the NMR chemical shifts, using the recently improved chemical shift–based Rosetta3 fragment picker[15], again excluding all proteins with ≥20% sequence identity from the library. In the second stage, all broken backbone segments are closed by means of a standard loop-closure method that

combines fragment superposition and structure minimization. The probabilistic distance restraints derived from the alignments, used in standard RosettaCM[30], are removed, but experimental NOE distance restraints, if available, are included during this stage. Third, the resulting backbone models are optimized using the final all-atom refinement step of standard CS-Rosetta[7], but using the most recent parameter set (talaris2013.wts) for scoring the energy.

Selection of all-atom models using energies and chemical shifts. Using the above protocol, for each query protein, CS-RosettaCM is parameterized to generate 500 all-atom models from each starting template, for a total of up to 10,000 models. Those models are further evaluated for fitness with respect to their experimental NMR chemical shifts using the same method developed for the standard CS-Rosetta protocol[7]. Specifically, for each all-atom model, a χ^2 value is calculated between the experimental chemical shifts and values predicted by SPARTA+[26], and this value is added to the Rosetta all-atom energy. This chemical shift re-scored Rosetta all-atom energy is used to evaluate and select the final models.

Criteria for convergence and model acceptance. The ten models with the lowest chemical shift re-scored Rosetta all-atom energies are retained for inspection of their convergence relative to the lowest-energy model and are accepted as the predicted structure only if (1) these models cluster within less than 2.5 Å, in terms of C^{α} r.m.s. deviation$_{100}$, from the model with the lowest energy, and (2) the average Rosetta energy of the ten lowest-energy models is at least two s.d. lower than that of the ten lowest-energy models obtained by standard CS-Rosetta (provided with the same inputs and the same all-atom energy scoring scheme).

Software availability. The POMONA software, including clustering scripts, all required databases and a complete example for ubiquitin, together with the scripts used for the RosettaCM comparative modeling and structure-selection procedure, can be freely downloaded from http://spin.niddk.nih.gov/bax/software/POMONA. A public Web server (http://spin.niddk.nih.gov/bax/nmrserver/pomona) is also provided, but only for performing the less time-consuming POMONA alignment method for a protein with experimental chemical shift data. Such a search procedure typically takes approximately 0.5 h on a 10-CPU desktop work station. By default, this server also generates all inputs and scripts required for running the RosettaCM comparative modeling structure generation. For this purpose, RosettaCM can be downloaded with the Rosetta Software Suite from http://www.rosettacommons.org/software.

21. Koonin, E.V. & Galperin, M.Y. (eds.). *Sequence - Evolution - Function: Computational Approaches in Comparative Genomics* (Kluwer Academic, 2003).
22. Kabsch, W. & Sander, C. *Biopolymers* **22**, 2577–2637 (1983).
23. Madhusudhan, M.S., Marti-Renom, M.A., Sanchez, R. & Sali, A. *Protein Eng. Des. Sel.* **19**, 129–133 (2006).
24. Needleman, S.B. & Wunsch, C.D. *J. Mol. Biol.* **48**, 443–453 (1970).
25. Smith, T.F. & Waterman, M.S. *J. Mol. Biol.* **147**, 195–197 (1981).
26. Shen, Y. & Bax, A. *J. Biomol. NMR* **48**, 13–22 (2010).
27. Caprara, A., Carr, R., Istrail, S., Lancia, G. & Walenz, B. *J. Comput. Biol.* **11**, 27–52 (2004).
28. Wohlers, I., Domingues, F.S. & Klau, G.W. *Bioinformatics* **26**, 2273–2280 (2010).
29. Carugo, O. & Pongor, S. *Protein Sci.* **10**, 1470–1473 (2001).
30. Thompson, J. & Baker, D. *Proteins* **79**, 2380–2388 (2011).

doi:10.1038/nmeth.3437