Published in final edited form as:

Proteins. 2014 April ; 82(4): 587–595. doi:10.1002/prot.24427.

BCL::Fold – Protein topology determination from limited NMR restraints

Brian E. Weiner¹, Nathan Alexander¹, Louesa R. Akin¹, Nils Woetzel¹, Mert Karakas¹, and Jens Meiler^{1,*}

¹Department of Chemistry, Center for Structural Biology, Vanderbilt University, Nashville TN, 37232, USA

Abstract

When experimental protein NMR data is too sparse to apply traditional structure determination techniques, *de novo* protein structure prediction methods can be leveraged. Here we describe the incorporation of NMR restraints into the protein structure prediction algorithm BCL::Fold. The method assembles discreet secondary structure elements using a Monte Carlo sampling algorithm with a consensus knowledge-based energy function. New components were introduced into the energy function to accommodate chemical shift, nuclear Overhauser effect, and residual dipolar coupling data. In particular, since side chains are not explicitly modeled during the minimization process, a knowledge based potential was created to relate experimental side chain proton-proton distances to C_β-C_β distances. In a benchmark test of 67 proteins of known structure with the incorporation of sparse NMR restraints, the correct topology was sampled in 65 cases, with an average best model RMSD100 of 3.4 ± 1.3 Å versus 6.0 ± 2.0 Å produced with the *de novo* method. Additionally, the correct topology is present in the best scoring 1% of models in 61 cases. The benchmark set includes both soluble and membrane proteins with up to 565 residues, indicating the method is robust and applicable to large and membrane proteins that are less likely to produce rich NMR datasets.

Keywords

protein structure prediction; topology prediction; sparse data; NOE; RDC

Introduction

Traditional structure determination via NMR spectroscopy requires a rich dataset with a preference for distance restraints between amino acids that are far apart in sequence which serve to define the protein topology. In cases of sparse or primarily local restraints, identification of the correct topology becomes more difficult as several incorrect topologies may also satisfy the restraints. Additionally, knowledge of the topology is often required to assign otherwise ambiguous nuclear Overhauser effect (NOE) cross peaks that can then be used as additional distance restraints to further refine the structure. Recently, spectroscopists have begun taking advantage of advances in protein NMR such as perdeuteration, selective labeling, and TROSY to study large proteins that were previously considered outside the realm of protein NMR. Nonetheless, the data collected on these large proteins are often sparse and of reduced quality, making structure determination challenging. Thus

*Corresponding author: Jens Meiler, Vanderbilt University, 465 21st Ave. S., Nashville, TN 37232, USA, jens.meiler@vanderbilt.edu; Ph: 615-936-5662; Fax: 615-936-2211 .

computational tools designed to predict protein topology from sparse data could facilitate the structure determination process.

Incorporating sparse NMR data into computational protein structure prediction algorithms has been shown to be extremely successful¹⁻⁴. Rosetta, for example, was able to correctly fold proteins up to 25 kDa using backbone-only NMR data⁵. For larger proteins, the algorithm was unable to sample native-like topologies, which indicates that conformational sampling is still the computational bottleneck, even with the inclusion of experimental restraints. Incorporation of sparse side chain distance restraints from deuterated samples increased the feasible upper limit to 40 kDa⁶.

Like many protein structure-prediction methods, Rosetta uses a simplified side chain approximation during the model building stages, so handling of any available side chain-side chain NOE restraints is not directly modeled. In these cases, an arbitrary amount is typically added to the distance restraint in order to represent the restraint as a backbone-backbone distance. This approach however reduces the information content of each restraint. The problem of relating experimentally determined distances to distances measurable during the minimization process is not unique to NMR data. In site-directed spin labeling electron paramagnetic resonance (SDSL-EPR) experiments, distances are reported between two spin labels covalently attached at specific sites on the protein model. A knowledge-based potential has been developed and successfully used to evaluate the probability of observing the C_{β} - C_{β} distance given the spin label-spin label distance⁷. We take a similar approach with side chain-side chain proton distances from NOE data to evaluate C_{β} - C_{β} distances in the model with the hypothesis that this method will produce more native-like models.

A protein structure prediction method, BCL::Fold, was recently introduced with the goal of efficiently sampling larger and more complex topologies than those accessible to other *de novo* protein structure prediction algorithms⁸. Like most algorithms, BCL::Fold begins with protein secondary structure prediction. The predicted secondary structure elements (SSEs) are then collected into a pool, with loops and side chains being discarded. A Monte Carlo algorithm assembles the SSE building blocks into a viable topology, guided by a consensus knowledge-based energy function. The final model is generated via subsequent loop building and side-chain replacement. Both the assembly and scoring stages are flexible, making the incorporation of experimental restraints possible. This has already been successfully demonstrated with cryo-electron microscopy data⁹.

Here we describe the incorporation of three types of NMR restraints – chemical shifts (CSs), NOEs, and residual dipolar couplings (RDCs) – into the BCL::Fold algorithm. A novel NOE knowledge-based potential was developed in order to evaluate C_{β} - C_{β} distances observed in the model based on experimental side chain-side chain restraints. The method was benchmarked using 23 structures with experimental restraints and an additional 44 proteins with simulated restraints. The incorporation of restraints enhanced native-like sampling and facilitated the selection of low RMSD models. BCL::Fold is therefore a viable method for rapid identification of protein topology from sparse NMR restraints.

Materials and Methods

INPUT FILES

Chemical shift data is read in indirectly as a TALOS+¹⁰ secondary structure prediction file (*SS.tab). Both RDC and NOE data are read in directly using the NMR-STAR 3.1 format¹¹ as supported by the BMRB. RDC data can be normalized to N-H values or have signs adjusted to account for the negative gyromagnetic ratio of nitrogen via command-line flags, but was not necessary for the selected benchmark proteins.

SELECTION OF BENCHMARK PROTEINS

67 total proteins were selected from three groups: 1) 6 large proteins from the BCL::Fold benchmark, 2) 38 membrane proteins from the BCL::MP-Fold benchmark, and 3) 23 small, soluble proteins containing experimental NMR data. The experimental benchmark set contains proteins that have both NOE and RDC data available on the BMRB¹², aside from 1CFE¹³, 1ULO¹⁴, and 2EE4, which have no RDC data. The benchmark proteins with experimental data contain no ligands, have less than 30% sequence similarity, range in length from 58 to 224 residues, and are soluble, single chains. Additionally, the proteins were selected to have a diverse set of alpha, beta, and alpha/beta topologies with > 50% SSE content.

MODIFICATION TO THE ALGORITHM

The NMR restraint scores are added to the BCL::Fold method as part of the restraint protocol. Refer to the supplementary information for required command line flags and modifications to the stage and score weight set files. Iterative folding rounds were also introduced to better leverage experimental restraint information. After generating 1000 models, the top 10 models were selected by restraint score and used as start models to generate a new set of 1000 models. For the six large, soluble proteins, this process was repeated once more. In the subsequent analysis, only the models produced by the last iteration are considered.

BENCHMARK

1000 models were generated with and without the incorporation of NMR restraints for each protein in the benchmark set. All CS and RDC data for residues in SSEs were used when available. When CS data was not available for SSE pool generation, it was simulated using SPARTA+¹⁵. In order to simulate sparse NOE data, random subsets of the experimental restraints were selected where both atoms were in SSEs and at least five residues apart. Here we exclude short and medium range distance restraints in order to focus on the long range distance restraints that serve to constrain the topology. Experimental selective labeling strategies also enrich for long range distance restraints since there is an increased chance neighboring atoms are not labeled; instead there is a predominance of side chain methyl groups that engage in long range van der Waals contacts in the protein core. For each protein, ten random subsets were selected, and the subset size was equal to the number of residues in SSEs. These datasets were further reduced (down to 0.1 restraints/residue) and expanded (up to 2.0 restraints/residue) in order to evaluate the effect of restraint density on topology prediction accuracy. To generate the complete 1000 models, 100 models were constructed for each NOE restraint subset. Example command lines for running BCL::Fold can be found in the Supporting Information.

AVAILABILITY

BCL::Fold is implemented as part of the BioChemical Library, a suite of software currently under development in the Meiler laboratory (www.meilerlab.org). BCL software, including BCL::Fold, is freely available for academic use.

Results and Discussion

RESTRAINT SCORE FUNCTIONS

Three scoring functions were introduced into BCL::Fold in order to accommodate evaluation of NMR restraints. RDCs are evaluated using the traditional Q-value measure¹⁶. To evaluate NOE distance restraints, a knowledge-based score, NOE-KB, and an atom distance penalty score, NOE-pen, are used in conjunction. CS's are evaluated indirectly

using the previously described secondary structure prediction agreement score¹⁷ via the program TALOS+¹⁰.

To evaluate RDC restraints, the optimal tensor is determined using the Saupe order matrix approach¹⁸⁻²⁰ after each minimization step. This gives a calculated theoretical RDC value for each supplied experimental value. The Q-value is then calculated,

$Q = \sqrt{\sum_{ij} (D_{\text{exp}}^{ij} - D_{\text{theor}}^{ij})^2 / \sum_{ij} (D_{\text{exp}}^{ij})^2}$, where D^{ij} is the dipolar coupling between nuclei i and j ¹⁶. The unweighted score is given by, $RDC = Q - 1$, so that a perfect agreement gives a score of -1 .

Since BCL::Fold assembles SSEs lacking side chain atoms, a method was needed to relate distance restraints between side chain protons to useable backbone-to-backbone distances. The PISCES databnk²¹ was used to cull a list of 4379 proteins with less than 25% sequence identity and better than 2.0 Å resolution. Proton atoms were added using the program Reduce²². Statistics were then collected in order to relate each H-H distance to the corresponding C_β - C_β distance. A separate histogram was created for the total number of bonds the protons were away from the C_β . For example, a $H_{\beta3}$ - $H_{\delta2}$ pair totals four bonds away from C_β 's. Separate histograms were generated for restraints to H_α or amide H since the coordinates of these atoms can be determined directly from BCL::Fold models. The C_β - C_β distance minus the H-H distance was computed and placed in a corresponding 0.5 Å bin. This process was repeated for each H-H pair at least 5 residues apart in sequence but no more than 6.0 Å apart in space for each of the proteins in the dataset. Each histogram was then converted to a cubic spline such that distances in the most common bin receive a score near -1 and distances not observed receive a score of zero (Figure 1A-C). The unweighted NOE-KB score is set as the mean individual restraint scores.

The NOE-pen score is simply a trigonometric transition between the maximal score, zero, and the ideal score, -1 . The width of the transition is set to 25 Å. The curve is generated such that it reaches a value of -1 at a distance of 2 Å greater than the smallest observed distance for the given atom types (Figure 1D). This score was introduced to evaluate moderately to severely violated distance restraints; the NOE-KB score has a rather narrow minimum, and thus cannot adequately discriminate these violations.

The standard BCL::Fold KB energy potentials scale linearly with respect to protein size. For consistency, each restraint score is therefore multiplied by the number of residues in the protein model to achieve the same property. An additional consideration for restraint scores is how to handle scaling of the score with the number of restraints. We chose to have the score scale logarithmically with the number of restraints. This allows for the score to change with additional restraints, but not overwhelmingly so. Finally, each score was given a relative weight of 5.0. With this scaling the experimental data contribute approximately 50% to the total score of the model while the KB potentials contribute the remainder of the score. The final restraint energy is given by the following equation:

$$E_{\text{rest}} = N \left(w_{RDC} (Q - 1) \log (M_{RDC} + 1) + \left(w_{KB} \bar{s}_{KB} + w_{pen} \bar{s}_{pen} \right) \log (M_{NOE} + 1) \right),$$

where M is the number of restraints, N is the number of amino acids in the target, w is the weight (the default case being 5.0), and $\bar{s}$ is the average NOE score.

SELECTION OF A DIVERSE BENCHMARK SET

A benchmark set of proteins of known structure was collected to test for the ability of the NMR scores to enhance native-like sampling during BCL::Fold minimizations. The set contains 67 total proteins, broken into three groups. 23 proteins are small, soluble proteins, with structures determined by NMR and with CS, NOE, and/or RDC data available on the BMRB. An additional six are large (> 220 residues) proteins from the original BCL::Fold method benchmark test⁸. The final 38 proteins are membrane proteins from the BCL::MP-Fold benchmark test²³. Membrane proteins are on the frontier of protein NMR, and are therefore more likely to produce sparse, rather than complete, datasets.

The small soluble proteins have complete datasets, so random subsets of NOE restraints were selected for a total of one long-range restraint per residue in SSEs to create sparse data. NMR restraints were simulated for the large soluble proteins and the membrane proteins. Again one restraint per residue was selected as the initial restraint density. For the membrane proteins, side chain NOE restraints (1 restraint/residue) were limited to isoleucine, leucine, and valine residues to mimic the increasingly popular strategy of specific isotopic labeling of methyl groups²⁴.

NOE KNOWLEDGE-BASED FUNCTION ENRICHES FOR NATIVE-LIKE MODELS

Each small, soluble native protein in the benchmark set was scored with the NOE-KB score and the NOE-pen score for agreement with all available long range experimental NOEs. With an ideal score of -1.00 , the mean NOE-KB score was -0.84 ± 0.07 BCL energy units (BCLEUs), and the mean NOE-pen score was -1.00 ± 0.00 BCLEUs. The NOE-KB score is not exactly -1.00 BCLEUs due to experimental error and the fact that the score represents a rather wide distribution of observed distances, with only the most commonly occurring receiving scores near -1.00 BCLEUs.

In order to test the ability of NOE scores to select for native-like models, we created a set of decoy models. For each protein, 10,000 decoys were generated by de novo protein structure prediction without restraints using BCL::Fold. These decoys were then also scored with the two NOE scores. We define any model with less than $8.0 \text{ \AA}$ RMSD100²⁵ to the native as “native-like” or a “good” model. RMSD100 is the C_{α} RMSD normalized to a protein length of 100 residues. This measure is useful when evaluating proteins of varying sizes, such as those used in this benchmark. Using the $8.0 \text{ \AA}$ cutoff, the enrichment was calculated for those proteins which produced at least 0.1% “good” models¹⁷. Ranking the models by the sum of the NOE scores produces an average enrichment of 5.5 ± 1.6 out of a maximal 10.0. In contrast, using a quadratic energy function analogous to the bounded energy potential in Rosetta¹ produces an average enrichment of 4.9 ± 1.4 ($p = 0.02$). This demonstrates that the NOE-KB and NOE-pen scoring functions improve the identification of native-like models when compared to the traditional score.

NATIVE-LIKE SAMPLING IS ENHANCED WITH NMR RESTRAINT SCORES

For each protein in the benchmark set, 1000 models were generated using the de novo BCL::Fold method. An additional 1000 models were also constructed using the available NMR restraints in combination with the implemented scoring functions. Over all proteins, the average C_{α} RMSD100 of the best model to the native structure was $3.4 \pm 1.3 \text{ \AA}$ with restraints and $6.0 \pm 2.0 \text{ \AA}$ without (Table I, Figures 2,3). When a structure with an RMSD100 of less than $8.0 \text{ \AA}$ is considered to be the correct topology, the inclusion of restraints allows for sampling of the correct topology in 65 of 67 cases (97%) compared to 54 of 67 cases (81%) when no restraints are incorporated. With a cutoff of $6.0 \text{ \AA}$, the correct topology is sampled in 64 cases (96%) with restraints and in 41 cases (61%) without. With a cutoff of $4.0 \text{ \AA}$, the correct topology is sampled in 54 cases (81%) with restraints and in 9

cases (13%) without. When looking at the top 5% of models produced from the first round, the best dataset contributes 18% of the top models on average (vs 10% expected with a random distribution), with the worst contributing 3% (Table S1). We conclude that while there is a dataset bias, even the ‘worst’ dataset is capable of producing highly accurate models – possible additional sampling is needed.

Of the small, soluble proteins, 2KYY showed the largest improvement upon the incorporation of restraints, with a best model RMSD100 decrease of 5.8 Å. The protein is a mixed α/β fold with 153 residues. The de novo method assembles a sheet, but the strand order is incorrect and the helices are not properly placed on either side of the major sheet. In contrast, the NMR method is able to build the sheet with the proper ordering and the helices are appropriately placed. Of the proteins with simulated NMR data, 1VIN²⁶ showed the largest improvement upon the incorporation of restraints, with a best model RMSD100 decrease of 7.5 Å. This protein contains thirteen helices and 252 residues, placing it on the upper edge of de novo BCL::Fold’s predictive capabilities; the native topology is sampled however, even without restraints⁸. Here restraints serve to improve accuracy by promoting sampling of those models with the correct topology. After the first round of iterative folding, the best model produced has an RMSD100 of 4.7 Å. The subsequent iterations then are typically starting their minimizations with the correct topology, making production of an accurate model much more likely.

BCL::FOLD COMPARES FAVORABLY WITH THE ROSETTA METHOD

Rosetta is a well established protein structure prediction method with a proven track record of producing quality models with limited experimental data. The structures of the soluble proteins in the benchmark were also predicted using the same sparse datasets using the AbinitioRelax application in Rosetta. Chemical shift data were used to generate fragments, and both NOE and RDC data were used during the minimizations. Side chain NOE restraints were converted to C_β restraints by adding 1.0 Å to the restraint distance per bond from the side chain proton to the C_β . 1000 models were generated per target, and the top 5% of models selected by RMSD100 to the native were retained for comparison with BCL models. The mean RMSD100 of the top Rosetta models was 4.9 ± 1.8 Å compared to 3.9 ± 1.4 Å for BCL::Fold (Table S2, $p = 0.003$). While BCL::Fold appears to sample topologies slightly better than Rosetta in our experiment, it should be noted that Rosetta is still the method of choice for loop building and side chain replacement once the topology has been constructed.

FEW NOE RESTRAINTS ARE REQUIRED FOR THE SAMPLING IMPROVEMENT

The previously described benchmark test used one NOE restraint per residue in SSEs. As a next step, additional restraint densities (0.1, 0.2, 0.5, and 2.0 restraints/residue) were tested for those proteins containing experimental data (Figure 4). After iterative folding, the top 5% of models by RMSD100 were analyzed from each group. The model quality improves up to 0.5 restraints/residue, but further increasing the number of restraints to 1.0 restraints/residue shows no effective additional improvement (the mean RMSD100 decrease is 0.2 ± 0.9 Å, $p = 0.31$). Analyzed separately, however, sampling for the larger proteins (> 125 residues) does improve overall from 0.5 to 1.0 restraints/residue. For proteins less than 125 residues, the average improvement in the top 5% of models selected by RMSD100 sampled is 0.0 ± 0.6 Å. For proteins with more than 125 residues, the improvement is 0.6 ± 1.3 Å.

RESTRAINT SCORES FACILITATE MODEL SELECTION

The selection of the best model(s) out of the thousands generated is a difficult problem, especially when using low-resolution energy functions, as is the case with BCL::Fold. Table I highlights this problem by listing the RMSD100 of the lowest energy model. When no restraints are considered, the average RMSD100 is 10.6 ± 2.3 Å. However when NMR

restraints are used, the average RMSD100 of the model with the lowest score is 5.4 ± 2.6 Å. Perhaps more strikingly, when the top 1% of models are selected by score, the native topology is contained within this subset in 27 out of 67 cases (40%) without restraints versus 61 out of 67 cases (91%) when using sparse NMR data.

BUILDING FULL ATOM MODELS

In order to explore the feasibility of constructing full atom models from BCL::Fold-generated topologies, we used the protein 1VIN as a test case. For this 252 residue helical protein, BCL::Fold produced models with an RMSD100 down to 1.8 Å compared to the native when sparse restraints were considered. The 50 lowest scoring models of the 1000 generated during the BCL::Fold benchmark test were retained for loop building using the Rosetta CCD loop building protocol. Side chains were then added using the Rosetta FastRelax protocol to generate 1000 complete, full atom models. Of the 20 best scoring final models, the mean backbone C α RMSD100 was 2.4 ± 0.2 Å RMSD100 to the native SSE residues and 4.5 ± 0.4 Å over all residues.

POTENTIAL APPLICATIONS

One potential use of sparse restraints with BCL::Fold is to assist in the identification of ambiguous NOE assignments. For proteins that are suitable for traditional NMR structure determination methods, this would speed up the process by allowing for more confident NOE assignments during the structure determination process. Additionally, the BCL::Score program can be used to identify any violated restraints in the given model, which can lead to subsequent NOE re-assignments or model refinement.

Perhaps the most exciting application for BCL::Fold lies with membrane proteins. Membrane proteins constitute roughly 50% of all known drug targets, yet only 2% of the deposited PDB structures²⁷. BCL::Fold can sample the native topology in all but 2 of the 38 membrane proteins in the benchmark when combined with sparse NMR data. This includes predicted models of less than 4.0 Å RMSD100 to the native for five proteins larger than 400 residues (with up to 15 transmembrane helices).

Conclusions

The *de novo* protein structure prediction method, BCL::Fold, has been updated to incorporate sparse experimental NMR data. Scoring functions were introduced to evaluate CS, NOE, and RDC data. In particular, a NOE knowledge-based potential was developed to relate experimental side chain proton-proton distance restraints to C β -C β distances that are measurable during the BCL::Fold minimization.

The benchmark test using a robust dataset demonstrated that sparse NMR data can be combined with BCL::Fold to produce native topologies in 97% of the cases. Using 1.0 NOE distance restraint per residue produces a mean improvement of 2.6 Å RMSD100 versus the *de novo* method. Reducing the number of restraints to 0.1 per residue still produces a mean improvement of 1.1 Å RMSD100 versus the *de novo* method. BCL::Fold, therefore, has the potential to provide experimentalists with feasible models that satisfy available NMR data to be used to generate further structure-based hypotheses.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

The authors thank the Vanderbilt University Center for Structural Biology computational support team for hardware and software maintenance. We also thank the Vanderbilt University Advanced Computing Center for Research and Education for computer cluster access and support. Work in the Meiler laboratory is supported through NIH (R01 GM080403, R01 MH090192, R01 GM099842) and NSF (Career 0742762).

References

1. Rohl CA. Protein structure estimation from minimal restraints using Rosetta. *Methods Enzymol.* 2005; 394:244–260. [PubMed: 15808223]
2. Li W, Zhang Y, Kihara D, Huang YJ, Zheng D, Montelione GT, Kolinski A, Skolnick J. TOUCHSTONE: protein structure prediction with sparse NMR data. *Proteins.* 2003; 53(2):290–306. [PubMed: 14517980]
3. Latek D, Kolinski A. CABS-NMR--De novo tool for rapid global fold determination from chemical shifts, residual dipolar couplings and sparse methyl-methyl NOEs. *J Comput Chem.* 2011; 32(3): 536–544. [PubMed: 20806263]
4. Zheng D, Huang YJ, Moseley HN, Xiao R, Aramini J, Swapna GV, Montelione GT. Automated protein fold determination using a minimal NMR constraint strategy. *Protein Sci.* 2003; 12(6):1232–1246. [PubMed: 12761394]
5. Raman S, Lange OF, Rossi P, Tyka M, Wang X, Aramini J, Liu G, Ramelot TA, Eletsky A, Szyperski T, Kennedy MA, Prestegard J, Montelione GT, Baker D. NMR structure determination for larger proteins using backbone-only data. *Science.* 2010; 327(5968):1014–1018. [PubMed: 20133520]
6. Lange OF, Rossi P, Sgourakis NG, Song Y, Lee HW, Aramini JM, Ertekin A, Xiao R, Acton TB, Montelione GT, Baker D. Determination of solution structures of proteins up to 40 kDa using CS-Rosetta with sparse NMR data from deuterated samples. *Proc Natl Acad Sci U S A.* 2012; 109(27): 10873–10878. [PubMed: 22733734]
7. Hirst SJ, Alexander N, McHaourab HS, Meiler J. RosettaEPR: an integrated tool for protein structure determination from sparse EPR data. *J Struct Biol.* 2011; 173(3):506–514. [PubMed: 21029778]
8. Karakas M, Woetzel N, Staritzbichler R, Alexander N, Weiner BE, Meiler J. BCL::Fold--de novo prediction of complex and large protein topologies by assembly of secondary structure elements. *PLoS One.* 2012; 7(11):e49240. [PubMed: 23173050]
9. Lindert S, Staritzbichler R, Wotzel N, Karakas M, Stewart PL, Meiler J. EM-fold: De novo folding of alpha-helical proteins guided by intermediate-resolution electron microscopy density maps. *Structure.* 2009; 17(7):990–1003. [PubMed: 19604479]
10. Shen Y, Delaglio F, Cornilescu G, Bax A. TALOS+: a hybrid method for predicting protein backbone torsion angles from NMR chemical shifts. *J Biomol NMR.* 2009; 44(4):213–223. [PubMed: 19548092]
11. Hall SR, Cook APF. Star Dictionary Definition Language - Initial Specification. *Journal of Chemical Information and Computer Sciences.* 1995; 35(5):819–825.
12. Ulrich EL, Akutsu H, Doreleijers JF, Harano Y, Ioannidis YE, Lin J, Livny M, Mading S, Maziuk D, Miller Z, Nakatani E, Schulte CF, Tolmie DE, Kent Wenger R, Yao H, Markley JL. BioMagResBank. *Nucleic Acids Res.* 2008; 36:D402–408. (Database issue). [PubMed: 17984079]
13. Fernandez C, Szyperski T, Bruyere T, Ramage P, Mosinger E, Wuthrich K. NMR solution structure of the pathogenesis-related protein P14a. *J Mol Biol.* 1997; 266(3):576–593. [PubMed: 9067611]
14. Johnson PE, Joshi MD, Tomme P, Kilburn DG, McIntosh LP. Structure of the N-terminal cellulose-binding domain of *Cellulomonas fimi* CenC determined by nuclear magnetic resonance spectroscopy. *Biochemistry.* 1996; 35(45):14381–14394. [PubMed: 8916925]
15. Shen Y, Bax A. Protein backbone chemical shifts predicted from searching a database for torsion angle and sequence homology. *J Biomol NMR.* 2007; 38(4):289–302. [PubMed: 17610132]

16. Cornilescu G, Marquardt JL, Ottiger M, Bax A. Validation of protein structure from anisotropic carbonyl chemical shifts in a dilute liquid crystalline phase. *Journal of the American Chemical Society*. 1998; 120(27):6836–6837.
17. Woetzel N, Karakas M, Staritzbichler R, Muller R, Weiner BE, Meiler J. BCL::Score--knowledge based energy potentials for ranking protein models represented by idealized secondary structure elements. *PLoS One*. 2012; 7(11):e49242. [PubMed: 23173051]
18. Losonczi JA, Andrec M, Fischer MW, Prestegard JH. Order matrix analysis of residual dipolar couplings using singular value decomposition. *J Magn Reson*. 1999; 138(2):334–342. [PubMed: 10341140]
19. Saupe A. Recent Results in Field of Liquid Crystals. *Angewandte Chemie-International Edition*. 1968; 7(2):97.
20. Meiler J, Peti W, Griesinger C. DipoCoup: A versatile program for 3D-structure homology comparison based on residual dipolar couplings and pseudocontact shifts. *J Biomol NMR*. 2000; 17(4):283–294. [PubMed: 11014592]
21. Wang G, Dunbrack RL Jr. PISCES: recent improvements to a PDB sequence culling server. *Nucleic Acids Res*. 2005; 33:W94–98. (Web Server issue). [PubMed: 15980589]
22. Word JM, Lovell SC, Richardson JS, Richardson DC. Asparagine and glutamine: using hydrogen atom contacts in the choice of side-chain amide orientation. *J Mol Biol*. 1999; 285(4):1735–1747. [PubMed: 9917408]
23. Weiner BE, Woetzel N, Karakas M, Alexander N, Meiler J. BCL::MP-Fold: Folding Membrane Proteins through Assembly of Transmembrane Helices. *Structure*. 2013
24. Rosen MK, Gardner KH, Willis RC, Parris WE, Pawson T, Kay LE. Selective methyl group protonation of perdeuterated proteins. *J Mol Biol*. 1996; 263(5):627–636. [PubMed: 8947563]
25. Carugo O, Pongor S. A normalized root-mean-square distance for comparing protein three-dimensional structures. *Protein Sci*. 2001; 10(7):1470–1473. [PubMed: 11420449]
26. Brown NR, Noble ME, Endicott JA, Garman EF, Wakatsuki S, Mitchell E, Rasmussen B, Hunt T, Johnson LN. The crystal structure of cyclin A. *Structure*. 1995; 3(11):1235–1247. [PubMed: 8591034]
27. Bakheet TM, Doig AJ. Properties and identification of human protein drug targets. *Bioinformatics*. 2009; 25(4):451–457. [PubMed: 19164304]

Figure 1. NOE knowledge based potentials. The energy potential for each cumulative bond distance is plotted versus the measured C_{β} - C_{β} distance subtracted from the experimental H-H distance. The bond distance is the number of bonds between the measured proton and the C_{β} atom of the same residue. For example, an NOE between $H_{\beta 3}$ and $H_{\beta 2}$ would have a cumulative bond distance of four. (A) Potentials for side chain-side chain NOEs. (B) Potentials H_{α} -side chain NOEs. (C) Potentials for backbone amide H-side chain NOEs. (D) The NOE-KB and NOE-pen potentials are plotted for a cumulative bond distance of 5.

Figure 2.

NMR restraints improve native-like sampling. (A) The mean RMSD100 values of the best 10 models sampled with and without restraints are plotted. Soluble proteins are represented by circles and membrane proteins by squares. Proteins are colored according to size: < 150 residues (green), 150 and < 250 residues (yellow), 250 and < 400 residues (orange), and 400 residues (red). The dashed line at 8.0 Å indicates the cutoff for the correct topology, and the dashed line at 4.0 Å indicates a feasible target for continuing with full atom refinement. The error bars are ± 1 S.D. (B) Of the top 10 models by score, the RMSD100 value of the best model is plotted for folding with and without restraints. Marker shapes and colors are the same as in panel A.

Figure 3.

Gallery of select benchmark results. Left column – Distribution of RMSD100 to native SSE values for models produced by the de novo method (red) and the restraint-based method (green). Right column – Superimposition of the best model produced by the restraint method (rainbow) with the native protein (gray). Refer to the supporting information for the complete gallery of benchmark results.

Figure 4.

Sampling efficiency depends upon restraint density. The size of the random subset of NOEs selected for folding was adjusted relative to the total number of residues in native SSEs. Each of the 23 proteins with experimental data was folded at varying restraint densities (0.0, 0.1, 0.2, 0.5, 1.0, and 2.0 restraints/residue). The distribution of the mean RMSD100 for the top 5% (selected by RMSD100) of models for each benchmark protein are shown. The boxes contain values within one standard deviation of the mean (of mean RMSD100 values) and the lines represent the minimum and maximum values observed from the 23 proteins for that restraint density. *Improvement over previous restraint density ($p < 0.01$).

Figure 5.

Core side chain conformations can be accurately predicted. Native protein model 1VIN is shown in gray, with side chain atoms displayed for His63, Leu64, Tyr68, and Phe97. The corresponding side chains from the best scoring Rosetta model after full-atom refinement are shown in black.

Table 1

Benchmark statistics and results.

PDB	Statistics				Restraints	RMSD100 (Å)				Score (Å)				
	AA	Type	Helices	Strands		rco	Best		Top 5%		Best		Top 5%	
							NMR	dn	NMR	dn	NMR	dn	NMR	dn
1Q2N	58	A	3	0	0.41	exp	4.0	2.8	4.7	4.0	5.2	8.4	5.5	10.4
2KIQ	62	A	4	0	0.37	exp	3.0	4.9	4.3	6.7	5.1	13.8	5.3	12.5
2L9R	69	A	3	0	0.27	exp	3.0	4.1	3.5	5.2	9.5	13.3	5.2	10.8
1WCL	76	A	5	0	0.27	exp	2.9	5.5	3.2	7.1	3.9	10.7	4.4	11.0
2L7K	76	A	4	0	0.45	exp	3.3	5.9	3.9	7.3	7.4	10.4	7.2	11.0
1OP1	82	A	3	0	0.31	exp	2.8	3.4	3.2	4.2	5.1	12.9	6.9	10.7
2AMW	83	A	3	0	0.38	exp	4.1	4.2	4.5	6.2	6.7	7.1	7.0	10.2
2KYW	87	B	0	7	0.37	exp	4.8	7.0	5.3	8.5	5.4	10.7	6.7	11.7
2BG9	91	A (MP)	3	0	0.41	sim	2.3	2.8	2.5	3.4	2.8	9.9	2.8	6.7
1W09	92	A	3	0	0.44	exp	1.9	3.5	2.0	4.5	4.4	10.1	2.8	11.3
1NKZ	93	A (MP)	3	0		sim	5.8	4.3	6.8	4.6	16.2	11.2	12.0	8.3
2KCT	94	B	0	6	0.27	exp	4.0	8.6	4.6	9.3	10.0	12.0	9.1	12.3
2H45	95	B	0	6	0.32	exp	4.1	4.1	6.1	5.7	10.2	13.3	8.2	9.7
2L35	95	A (MP)	3	0		sim	2.6	3.1	2.8	3.7	3.5	17.2	3.5	9.7
2KLC	101	A/B	1	5	0.28	exp	3.6	4.6	4.4	7.2	7.2	11.8	5.4	11.6
2KSF	107	A (MP)	4	0	0.34	sim	2.9	3.9	3.1	4.5	3.6	5.1	3.3	5.6
2JV3	110	A	6	0	0.28	exp	2.5	5.1	3.2	7.1	5.7	8.9	4.9	10.0
2A7O	112	A	3	0	0.34	exp	1.7	2.3	2.1	4.2	4.1	11.6	3.7	11.0
2KCK	112	A	6	0	0.18	exp	3.0	5.7	3.8	7.8	6.3	12.6	5.3	10.0
1J4N	116	A (MP)	4	0	0.40	sim	2.6	4.9	3.2	5.9	4.6	9.6	4.9	9.0
2KD1	118	A	5	0	0.25	exp	2.6	4.5	2.8	5.5	5.0	9.8	4.6	9.2
3SYO	122	A (MP)	4	0	0.33	sim	4.9	5.2	5.4	6.3	7.6	9.7	8.6	10.0
1PY7	123	A (MP)	4	0	0.28	sim	2.4	3.9	2.7	4.7	3.1	5.4	3.2	6.4
2PNO	130	A (MP)	4	0	0.29	sim	1.8	5.0	2.3	6.7	2.8	5.4	3.1	8.6
1CFE	135	A/B	4	4	0.35	exp	2.8	5.7	3.2	8.3	3.9	12.2	4.3	10.8
2L3W	143	A	7	0	0.32	exp	2.8	6.2	3.3	8.1	3.4	9.6	5.3	10.3

PDB	Statistics				Restrains	RMSD100 (Å)				Score (Å)					
	AA		Type	Helices		Strands	rco	Best		Top 5%		Best		Top 5%	
								NMR	dn	NMR	dn	NMR	dn	NMR	dn
2BL2	145	A (MP)	6	0	0.37	sim	2.2	2.9	2.5	3.8	3.2	6.7	3.6	7.3	
1CMZ	152	A	9	0	0.26	exp	4.4	7.7	5.0	9.6	5.7	12.2	5.8	12.6	
1ULO	152	B	0	10	0.34	exp	4.1	6.9	4.6	8.7	5.4	12.4	6.1	11.3	
2KYY	153	A/B	3	6	0.31	exp	3.2	9.0	3.6	9.8	4.8	11.5	4.3	12.0	
2K73	164	A (MP)	6	2	0.33	sim	3.3	4.7	4.1	5.9	9.0	10.1	6.8	9.1	
1RHZ	166	A (MP)	6	0	0.33	sim	3.8	6.7	4.3	8.0	5.7	9.9	5.4	10.4	
1IWG	168	A (MP)	7	0	0.31	sim	2.4	4.3	2.9	5.6	3.2	8.5	3.6	8.3	
3P5N	179	A (MP)	8	0	0.24	sim	2.6	5.8	3.3	7.4	4.4	8.3	4.5	9.8	
2IC8	182	A (MP)	8	0	0.25	sim	2.9	6.0	3.8	7.2	4.3	9.5	5.2	9.3	
2YVX	188	A (MP)	5	0	0.34	sim	3.3	5.1	4.1	6.9	5.5	9.2	5.5	9.4	
1PV6	189	A (MP)	11	0	0.42	sim	2.6	5.7	2.8	6.8	3.4	10.6	4.1	9.4	
1OCC	191	A (MP)	5	0	0.33	sim	2.2	4.6	2.5	5.9	3.2	8.5	3.7	8.0	
2NR9	192	A (MP)	8	0	0.24	sim	3.5	5.7	4.1	7.2	4.7	8.7	5.0	9.5	
4A2N	192	A (MP)	6	2	0.31	sim	3.7	4.3	4.0	6.2	4.0	8.1	4.7	8.8	
1RW5	199	A	5	0	0.38	exp	1.6	4.7	1.8	7.9	2.3	11.5	3.0	11.1	
1KPL	203	A (MP)	8	0	0.31	sim	3.0	8.7	3.4	10.5	6.6	14.4	4.9	12.5	
2EE4	209	A	12	0	0.23	exp	2.8	7.5	3.5	9.4	3.6	12.8	4.6	11.4	
2ZW3	216	A (MP)	8	3	0.35	sim	2.6	4.0	3.2	5.1	5.3	9.2	5.8	8.1	
2BS2	217	A (MP)	8	0	0.27	sim	3.4	5.4	3.9	6.9	5.1	11.0	4.8	9.2	
1LOV	221	A (MP)	9	0	0.27	sim	3.3	5.2	3.9	7.2	8.2	9.0	7.5	9.4	
1UAI	223	B	0	16	0.25	sim	5.8	7.9	6.7	9.1	8.2	11.0	8.2	10.8	
2KSY	223	A (MP)	9	2	0.26	sim	2.1	5.1	2.6	6.3	3.4	9.3	3.2	8.6	
1PY6	227	A (MP)	7	2	0.27	sim	2.1	4.8	2.5	5.9	2.4	6.1	3.3	8.4	
1VIN	252	A	13	0	0.12	sim	1.8	9.3	2.3	10.1	2.9	12.3	2.7	11.9	
3KCU	252	A (MP)	14	0	0.29	sim	3.5	7.3	4.0	8.5	3.8	11.2	4.8	10.5	
1XQO	253	A	14	0	0.23	sim	6.6	8.8	7.6	10.1	9.7	12.6	9.3	12.2	
1FX8	254	A (MP)	12	0	0.28	sim	4.0	6.4	4.7	7.6	5.5	9.3	5.7	9.8	
2OF3	266	A	15	0	0.13	sim	3.4	9.6	3.9	11.2	4.7	13.5	4.8	13.6	
1U19	278	A (MP)	10	2	0.24	sim	3.0	5.3	3.9	6.6	3.8	8.9	4.2	8.8	

PDB	Statistics				Restrains	RMSD100 (Å)				Score (Å)				
	AA	Type	Helices	Strands		rco	Best		Top 5%		Best		Top 5%	
							NMR	dn	NMR	dn	NMR	dn	NMR	dn
2ZCO	284	A	15	0	0.17	sim	2.3	8.9	2.7	10.2	2.7	13.0	3.1	12.3
2R0S	285	A	14	0	0.20	sim	3.1	9.1	3.4	10.0	4.8	11.2	4.0	11.9
1OKC	292	A (MP)	11	0	0.25	sim	4.4	7.1	4.9	8.2	5.6	9.9	8.1	10.3
3KJ6	311	A (MP)	15	0	0.28	sim	3.5	5.9	4.8	7.4	3.5	10.5	5.5	10.0
3B60	319	A (MP)	11	0	0.27	sim	4.7	9.5	5.6	10.8	7.3	12.4	7.4	13.2
3HD6	403	A (MP)	15	2	0.23	sim	3.5	7.2	4.1	8.2	4.5	11.0	4.6	10.3
3GIA	433	A (MP)	18	0	0.34	sim	3.0	9.6	3.6	10.7	6.6	13.4	7.3	12.6
3OOR	449	A (MP)	18	0	0.15	sim	2.9	6.9	3.6	8.2	2.9	10.2	4.1	10.3
2XUT	488	A (MP)	24	0	0.22	sim	8.8	7.7	9.6	9.0	12.1	10.2	11.6	11.4
3HFX	493	A (MP)	18	0	0.36	sim	3.2	8.9	3.7	9.7	4.1	13.1	4.6	11.4
1YEW	528	A (MP)	20	3		sim	8.2	9.7	9.6	11.5	10.4	14.1	11.8	13.3
2XQ2	565	A (MP)	28	0	0.29	sim	3.5	8.2	4.0	10.1	5.4	12.2	5.7	12.1
Mean	199		8	1	0.30		3.4	6.0	4.0	7.3	5.4	10.6	5.5	10.3
SD	119		6	3	0.07		1.3	2.0	1.5	2.1	2.6	2.3	2.1	1.7

Protein types are "A" for alpha-helical and "B" for beta-strands. "MP" denotes a membrane protein. The NMR restraints used were from published experimental data ("exp") or simulated computationally ("sim"). The best models were selected by either RMSD100 ("RMSD100" columns) or score ("Score" columns). RMSD100 values are displayed for both the best model and the mean of top 5% of models. The models generated with NMR restraints ("NMR") and without ("dn").