Journal of Biomolecular NMR, **18:** 311–318, 2000.
KLUWER/ESCOM
© 2000 *Kluwer Academic Publishers. Printed in the Netherlands.*

De novo protein structure determination using sparse NMR data

Peter M. Bowers[a,*], Charlie E.M. Strauss[b,*] & David Baker[a,**]
[a]*Department of Biochemistry, University of Washington School of Medicine, Seattle, WA 98195, U.S.A.*
[b]*Biosciences Division, Los Alamos National Laboratory, Los Alamos, NM 87545, U.S.A.*

Received 12 July 2000; Accepted 18 September 2000

Key words: de novo, limited constraints, NOE, protein structure determination, Rosetta

Abstract

We describe a method for generating moderate to high-resolution protein structures using limited NMR data combined with the ab initio protein structure prediction method Rosetta. Peptide fragments are selected from proteins of known structure based on sequence similarity and consistency with chemical shift and NOE data. Models are built from these fragments by minimizing an energy function that favors hydrophobic burial, strand pairing, and satisfaction of NOE constraints. Models generated using this procedure with ∼1 NOE constraint per residue are in some cases closer to the corresponding X-ray structures than the published NMR solution structures. The method requires only the sparse constraints available during initial stages of NMR structure determination, and thus holds promise for increasing the speed with which protein solution structures can be determined.

Introduction

The accelerated pace of genome sequencing has resulted in an abundance of protein sequence data with few corresponding protein structures, sparking interest in the development of rapid structure determination methods and ambitious proposals for genome-scale structure determination. Traditional experimental structure determination methods are time- and labor-intensive processes. NMR solution structure determination requires concentrated samples, extensive isotope labeling, assignment of backbone and side-chain resonances, and iterative assignment of NOE spectra. Upwards of 15 constraints per residue are typically required to generate high-resolution solution structures. Using less information, low-resolution global folds can be obtained for large proteins that can greatly simplify resonance and constraint assignments (Venters et al., 1995; Rosen et al., 1996; Gardner et al., 1997; Battiste and Wagner, 2000).

Recent methods have sought to improve structure prediction and determination by confining the conformational search space (Figure 1). The Rosetta ab initio protein structure prediction method assembles protein structures with buried hydrophobic cores and paired beta strands from fragments of known protein structures with sequences similar to the target protein (Simons et al., 1997, 1999b). Bax and co-workers used fragment libraries with several hundred dipolar coupling constants to calculate a high-resolution structure of human ubiquitin (Delaglio et al., 2000). Simplified protein representations have been combined with knowledge-based potentials and selected short- and long-range distance restraints to produce >3 Å models for proteins as large as 180 residues (Skolnick et al., 1997; Kolinski and Skolnick, 1998; Debe et al., 1999; Standley et al., 1999). Despite significant recent progress, methods developed thus far for rapid structure determination are unable to consistently generate and identify high-resolution protein structures (<2 Å) utilizing small, easily obtained experimental data sets.

Here, we combine the strengths of the methods described in the previous paragraph by incorporating into the Rosetta method data that is typically acquired early in the NMR structure determination process. Using this approach, we are able to consistently generate and identify moderate to high-resolution models for pro-

*C.E.M.S. and P.M.B. contributed equally to this work.
**To whom correspondence should be addressed. E-mail: dabaker@u.washington.edu

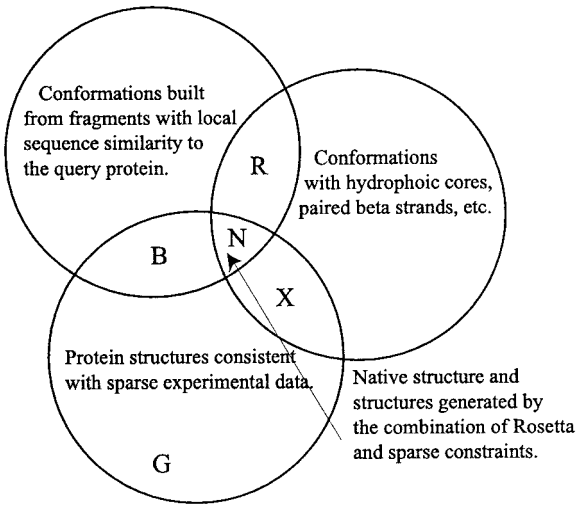

Figure 1. Diagram highlighting various experimental approaches for rapid protein structure determination. Methods include sparse experimental constraints (G) (Venters et al., 1995; Rosen et al., 1996; Gardner et al., 1997; Battiste and Wagner, 2000), experimental constraints combined with fragment libraries (B) (Delaglio et al., 2000), fragment libraries combined with protein scoring potentials that favor protein-like structures (R) (Simons et al., 1997, 1999a, b), and experimental constraints combined with scoring functions that favor protein-like structures (X) (Skolnick et al., 1997; Kolinski and Skolnick, 1998; Debe et al., 1999; Standley et al., 1999). The combined Rosetta/constraint method utilizes all three components: fragment libraries from the PDB with structural similarity to the query protein, a scoring potential that favors conformations with protein-like features, and sparse experimental distance constraints that can discriminate the native fold (N).

teins as large as 150 residues using $\leq$1 NOE constraint per residue.

Materials and methods

Generation of fragment libraries

Three and nine residue fragments taken from the Protein Data Base were given position-specific rankings as follows. Each fragment residue was scored according to its agreement with a multiple sequence alignment and a derived probability for the $\phi\psi$ angles given the chemical shift assignments for the ^{15}N, ^{13}Cα, ^{13}C$'$, ^{13}Cβ, and ^{1}Hα nuclei (see following paragraph). The score for each peptide fragment was the product of the component residue scores. Additionally, fragments with gross violation ($>$2 Å) of short range NOE upper bound distance constraints were discarded. To ameliorate regions where our estimation of the $\phi\psi$ was either ambiguous or mistaken, we augmented the library with fragments chosen by agreement with the multiple sequence alignment and the sequence-based predicted secondary structure (Simons et al., 1997, 1999b). The final library consisted of the 1000 top ranked fragments per residue in the query protein; in generating any given structure, roughly 25% percent of these are sampled.

For each residue, we used the TALOS (Cornilescu et al., 1999) algorithm to select a set of likely $\phi\psi$ pairs and corresponding quality scores based on the chemical shift and sequence information. We converted this discrete output into an estimator of the angular probability distribution about the mean ϕ and ψ angles. A simple, unimodal distribution function was selected to optimally utilize the TALOS prediction while not underestimating its error or overtraining on its database. If there was a large scatter in the angles of the top 10 predictions ($\sigma_{\phi SD}$), or if their TALOS-derived quality scores varied greatly (Z_{score}), then a large error bar was assigned to the angular prediction. Empirically, we found these errors bars to be well fit by $\phi_{err} = 3.4 * (\sigma_{\phi SD} * Z_{score})^{0.67}$, $\psi_{err} = 3.2 * (\sigma_{\psi SD} * Z_{score})^{0.81}$. Moreover, the observed distribution ($\Delta\phi$,$\Delta\psi$) between the TALOS-derived mean angle pair and the correct value in the database had infrequent but large deviations beyond these error bars and was represented well as two independent Poisson distributions (rather than a gaussian): Prob($\Delta\phi$,$\Delta\psi$) $\propto$ exp($-\Delta\phi/\phi_{err}$) exp($-\Delta\psi/\psi_{err}$).

Experimental NOE constraints (HN-HN, HN-Hα, and Hα-Hα) and chemical shift assignments were taken from the PDB and Biomagnetic Resonance Bank (BMRB) or were generously provided by individual NMR laboratories. Constraint sets were then paired *randomly* until the number of constraints was equal to the number of residues. For Rosetta fragment selection and structure calculation, upper bound distance constraints were used without further alteration.

Alternatively, artificial backbone upper bound NOE constraints were generated for all query proteins using a conservative protocol. All ^{1}H-^{1}H distances less than 5.0 Å in the query protein were identified. This set was stochastically paired to be consistent with the empirical distance dependent frequency with which NOEs of neighboring atoms are *unobserved* due to solvent exchange, resonance degeneracy, or rapid relaxation processes (Doreleijers et al., 1999). This data set was further paired *randomly* until the number of constraints was equal to the number of residues. Distance constraints were then grouped into bins of $<$3.0 Å, $<$4.5 Å, and $<$5.5 Å. Constraints falling within 0.2 Å of a bin boundary were conser-

vatively moved to the next higher bin. Artificial data sets contained 0.7–1.0 constraints/residue while experimental data sets ranged in size from 0.14 to 1.0 constraints/residue.

Generation of folded proteins

Starting from an extended chain, model proteins were generated by substituting backbone angle sets from the fragment library using a Monte Carlo/simulated annealing protocol (Simons et al., 1997, 1999b). The scoring function is identical to that used in the ab initio application of Rosetta (except as noted below), and its components and derivation have been described in detail previously (Simons et al., 1997, 1999b). Very briefly, the function was derived from a Bayesian treatment of the residue distributions in known protein structures and includes a residue environment term which represents primarily solvation effects (notably hydrophobic burial), a specific pair interaction term (primarily electrostatics), and terms favoring strand pairing, overall compactness, and other features of native protein structures. A NOE score, calculated as the sum of upper bound distance violations, was added to the previously described potential function. To reduce the extent to which the conformational search was constrained by the presence of long range NOEs, the NOE scoring term was cycled on and off in the early phase of each simulation. This proved necessary because efficient fragment insertion in torsion-space tended to be hindered by long range NOE constraints. Compact structures with <1 Å error per constraint were further optimized with a scoring function that incorporated a full atom Lennard-Jones (LJ) attractive and repulsive term. Upon each fragment insertion, a Monte Carlo/simulated annealing search of a Dunbrack rotamer library was used to identify a low energy set of side-chain conformers for that backbone structure (Kuhlman and Baker, 2000), from which LJ energies were then calculated.

Ten structures for each protein data set were chosen using a non-parametric selection scheme that first removes the structures with poor composite Rosetta scores (ab initio, NOE distance constraints, and LJ terms), followed by those structures with poor individual score components (LJ attractive, radius of gyration, NOE constraint score). The remaining structures were then ranked by their composite Rosetta score.

Global-fold determination

Protein global folds were generated for each NOE constraint set using the program X-PLOR and standard NMR substructure embedding and distance geometry/simulated annealing protocols (dg_sub_em bed.inp / dgsa.inp / refine.inp) (Brünger, 1992). Upper bound distance constraints were combined with lower distance bounds of 1.8 Å for all X-PLOR runs. Dihedral angle constraints derived from TALOS $\phi\psi$ predictions were also used without further alteration. Boundaries for the angle constraints were set as three times the standard deviation of the top 10 TALOS predicted angles. Square well potentials were used for both NOE and dihedral constraints. Ten structures were generated using the real or synthetic NOE constraint sets as well as angle constraints, having no NOE violations >0.5 Å, dihedral-angle constraint violations $>5°$, deviations from ideal bond length $>0.03°$, or deviations from ideal bond geometry >3 Å. Final vdW radii were set to 0.8 of their full CHARMM value.

Results and discussion

The Rosetta method is composed of two distinct components: selection of a library of protein fragments likely to resemble the query protein at each residue position and assembly of the fragments to produce models with features similar to those of known protein structures (Simons et al., 1997, 1999a, b). As illustrated in the CASP3 protein structure prediction experiment, Rosetta can in some cases produce quite reasonable low resolution structures for proteins of up to $\sim$100 residues in length from sequence information alone (Simons et al., 1997, 1999a, b). We conjectured that the addition of sparse NMR experimental data to improve selection of a fragment basis set and to better identify structures consistent with the native fold would enable Rosetta to generate and identify high-resolution structures (Figure 1, N). With the goal of developing a technique suitable to a high throughput experimental approach, we selected experimental constraints derived from NMR data that could be collected with minimal effort: backbone HN-HN NOE distance constraints and chemical shifts. The fragment selection strategy combined backbone chemical shift information and local NOE constraints with profiles from multiple sequence alignments and predicted secondary structure information (Simons et al., 1997). Torsion angles (ϕ, ψ) were predicted by matching the sequence and chemical shifts of the query protein

to the sequence and chemical shifts of a set of proteins with known structures (Cornilescu et al., 1999), and the predictions were then used in the fragment selection process. The combination of fragment libraries derived from different sources of information (chemical shifts, amino acid sequence profiles, sequence based secondary structure predictions) insures that Rosetta is resistant to large errors in any one prediction method. As our goal is to improve de novo structure determination, no fragments were taken from proteins with homology to the protein being folded. The fragment insertion/simulated annealing process was guided by the addition of an NOE constraint score to the scoring function used in ab initio calculations (Simons et al., 1997, 1999a, b).

Backbone NOE data sets were abstracted from the literature and distances were used without further alteration. These data sets, however, varied in coverage from 0.16 to 1 constraint per residue, making the results for different proteins difficult to compare. To facilitate comparison, we also calculated models for each protein using synthetic backbone NOE data sets with a coverage of $\sim$1 constraint per residue. Artificial NOE constraint sets were generated to be consistent with the observed completeness of experimental NOE data in NMR structure determinations as well as to reflect the natural abundance of local and non-local NOE (Doreleijers et al., 1999).

To test the Rosetta/NMR constraint approach, we selected nine proteins ranging in size from 52 to 152 residues with varied secondary structure and topology (Table 1). For each protein, a thousand structures were generated using real or artificial NOE constraint sets. The Rosetta algorithm ranks these structures without knowledge of the native structure, and we summarize these rankings with two performance metrics. In Table 1, we report the root mean square deviation (RMSD) to native of the top ranked Rosetta structure (*best score*) and the lowest RMSD to native observed in the 10 best scoring structures (*best RMSD*). The first metric measures the combined performance of the search algorithm and the discrimination function, while the second assesses the search algorithm undiminished by an imperfect discriminator.

For comparison, 10 distance geometry (DG) structures were generated using standard X-PLOR NMR structure determination protocols and potentials with the same experimental NMR constraint data used in the Rosetta calculations. Because the DG structures satisfied the distance constraints, the composite X-PLOR score was unable to distinguish the model with the best RMSD; we report the RMSD range spanned by the 10 DG models in Table 1.

The combined Rosetta and NOE constraint approach (referred to as Rosetta below) is substantially better than DG (Table 1) for calculating accurate protein structures. For proteins of less than 125 residues, Rosetta generates and identifies protein structures with RMSD values of 1–3 Å to the reference structure for either real or artificial constraint data (Table 1 and Figure 2). Structures generated using real or artificial constraint sets of comparable size had comparable RMSDs from the native structures (Table 1), suggesting that the synthetic data sets are representative of typical experimental data sets of the same size. Figures 2 and 3 show ribbon diagrams for protein structures generated by full experimental X-ray and NMR methods, Rosetta using sparse constraints, and DG using sparse constraints. The illustrations demonstrate that comparing RMSDs of structures generated by DG and Rosetta does not fully capture the improvements in local secondary structure and overall topology in the Rosetta structures. The results for specific proteins are summarized in Table 1 and are discussed below.

Proteins 1poh, 1ubq and 1acf contain mixed $\alpha\beta$ topologies with a moderate proportion of long-range backbone NOE constraints, providing a good test case for both the Rosetta and DG procedures. In addition, the 1poh, 1ubq and 1acf sequences have both published NMR solution and X-ray crystal structures, allowing us to benchmark the accuracy of the Rosetta and full NMR solution structures relative to the high-resolution X-ray crystal structures (Figure 2). All of the top 10 1poh and 1ubq structures were nearly identical with accuracies spanning a narrow range of 1.09–1.58 Å RMSD to the corresponding X-ray structures. Likewise, the published complete NMR solution structure determinations for 1poh, utilizing 10–20 constraints per residue, had an RMSD of 0.98 Å from the corresponding X-ray structure. The published 1ubq solution structure (1d3z) achieved a small 0.33 Å deviation from the reference X-ray structure (vs. 1.09 Å for Rosetta), but was calculated with $>$20 constraints per residue including several hundred dipolar coupling constraints. The best score and best RMSD Rosetta structures for 1acf, a larger 125-residue protein, varied slightly more, having RMSDs of 3.09 Å and 2.08 Å, respectively. This still compares favorably with the published solution structure which had an RMSD of 2.36 Å to the X-ray reference. Our results indicate that protein structures with accuracy (as measured by backbone RMSD match to the corresponding X-ray

Table 1. RMS deviation (Å) from reference structure

Protein[a]	#Res	Constraint[b]	Best RMSD Rosetta[c]	Best score Rosetta[d]	Distance geometry[e]	X-ray vs. NMR[f]
1bw5	52	Art. 52/0	1.34	2.60	>13.0	
1ubq$^{(\sim 2000)}$	76	Real 33/22	1.37	1.53	3.3–7.7	0.33[i]
1ubq$^{(\sim 2000)}$	76	Art. 39/38	1.09	1.52	2.8–6.5	0.33[i]
1poh$^{(\sim 1500)}$	85	Art. 49/37	1.31	1.58	4.7–5.6	0.98[h]
1imq$^{(1172)}$	86	Real 84/5	2.52	3.01	8.3–13.1	
1imq$^{(1172)}$	86	Art. 51/7	1.96	1.96	5.7–13.0	
1ck2$^{(1418)}$	107	Art. 67/41	3.06	3.06	5.4–10.6	
1acf	125	Art. 77/49	2.09	3.08	6.0–13.5	2.36[f]
1cfe$^{(1701)}$	135	Real 65/41	5.72	5.72	10.8–14.7	
1cfe$^{(1701)}$	135	Art. 83/53	3.48	3.48	8.1–15.3	
1ulo$^{(1928)}$	152	Real 16/35	6.97	6.97	10.8–13.6	
1ulo$^{(1928)}$	152	Art. 19/141	3.90	3.90	5.7–11.1	
1cmz	152	Real 22/0	5.79	9.47	>15.0	
1cmz	152	Art. 105/1	7.98	12.1	>15.0	

[a]pdb code for the reference X-ray or NMR structure used to calculate RMSD values. Proteins used include insulin gene enhancer protein Isl-1 (1bw5), human ubiquitin (1ubq), histidine-containing phosphocarrier protein (1poh), colicin E9 immunity protein Im9 (1imq), ribosomal protein L30 (1ck2), actin-binding protein profilin (1acf), pathogenesis-related protein P14a (1cfe), N-terminal cellulose-binding domain (1ulo), and Gα interacting protein (1cmz). In parentheses, the number of constraints used in the full experimental NMR solution structure determination.
[b]Number of real or artificial constraints used: number of short $|(i-j)| \leq 5$ and long $|(i-j)| > 5$ range NOE constraints used in the Rosetta and X-Plor structure determinations.
[c]Best RMSD (C′,Cα, N nuclei in regular secondary structure) Rosetta structure amongst the top 10 scoring structures.
[d]The best scoring Rosetta structure.
[e]Best and worst of 10 X-Plor structures satisfying all NOE constraints, ideal bond lengths, ideal covalent geometry, and vdW contacts.
[f]Comparison of NMR solution and X-ray crystal structures.
[g,h,i]NMR solution structures of 2prf, 1hdn, 1d3z.

structure) roughly equivalent to full NMR solution structure determinations can be generated using data sets with 1 constraint per residue in combination with the Rosetta fragment libraries and scoring potential.

The native structures of 1bw5 (52 residues), 1imq (86 residues), and 1cmz (152 residues) are all α-helical bundles and represent a more difficult test case for our combined approach. Because the distance between amide protons in adjacent packed α-helices is almost always >7 Å, few long-range backbone–backbone NOE constraints are observed in the experimental or synthetic constraint sets for these proteins. Not surprisingly, the real and artificial constraint DG structures for 1bw5, 1imq, and 1cmz have large RMSD values to their reference structures and lack both native topology and secondary structure. Despite the dearth of long-range constraints in the 1bw5 or 1imq data sets, high-resolution structures were generated and identified in each case using the Rosetta protocol, with best RMSD values of 1.34 Å (1bw5) and 1.96 Å (1imq) (Table 1). For 1cmz, a much larger protein, Rosetta was able to generate a 5.8 Å structure using the small experimental data set, containing only 22 local and 0 long-range NOE constraints (0.14 constraints/residue). While among the poorest in the set of proteins studied here, it is still substantially better than the corresponding distance geometry structure (Figure 3 and Table 1). The synthetic data set, which contained a single long range constraint, performed worse than the smaller experimental data set containing no long range constraints, presumably due to a combination of the more conservative upper bounds used in the artificial data set and sensitivity to which key constraints are included or excluded.

Rosetta, in combination with sparse constraints, was also able to dramatically improve the accuracy of structures calculated for the larger and more complex proteins compared to those calculated using global-

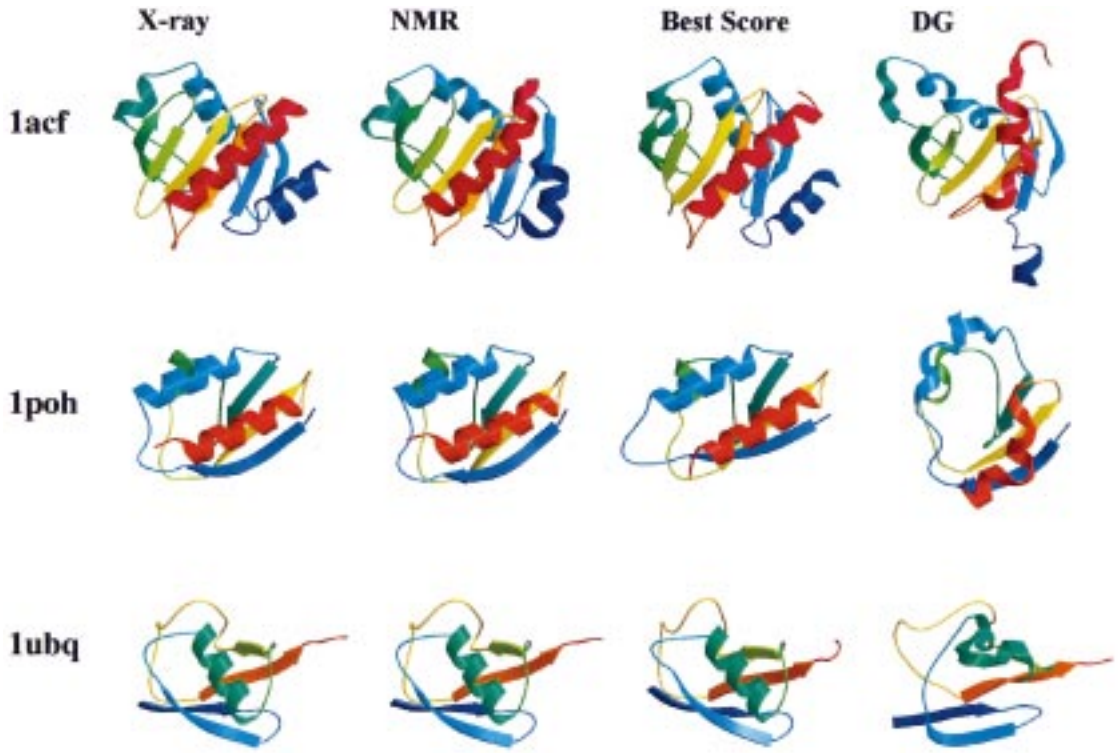

Figure 2. Comparison of Rosetta and DG results for αβ proteins. From top to bottom: proteins 1acf, 1poh, and 1ubq. From left to right: the reference X-ray crystal structure, the reference NMR solution structure, the best scoring Rosetta/limited constraint structure, and the best RMSD distance geometry structure (of 10 structures). Structures generated by Rosetta often reproduce the subtle structural details of experimentally determined protein structures, such as sharply defined secondary structure, sheet rolls, and helix bends. These features are particularly evident when comparing the Rosetta and DG structures for 1acf. The main variation in tertiary structure between the 1poh X-ray and Rosetta models occurs in only two of the loops. In fact, the published NMR structure (1hdn) itself differs from the X-ray (1poh) structure in this area as well. The Rosetta and DG structures were determined using artificial (∼1 constraint/residue) NOE data sets. The coloring scheme traces the protein backbone from the N (violet) to the C (red) terminus. Figures were created using the program MOLSCRIPT and rendered by Raster3D (Kraulis, 1991; Merritt and Murphy, 1994).

fold techniques. For these proteins, Rosetta does not efficiently generate compact structures satisfying the constraints, when starting from an extended chain conformation. As the structures produced by distance geometry satisfy nearly all of the constraints, we found it more efficient to use the Rosetta algorithm with constraints to refine the DG models.

1ulo, a 152-residue β-sheet sandwich with large numbers of non-local contacts and complex topology, is a difficult test case for the Rosetta method because of the large fraction of non-local interactions. The best RMSD DG structure generated using simulated (1 constraint per residue) and real (0.33 constraint per residue) data sets had RMSDs of 5.7 Å and 10.6 Å, respectively. Rosetta refined these to 3.9 and 7.0 Å (Figure 3). Similar results were observed for 1cfe, a 135-residue protein with mixed αβ topology. The artificial and real constraint DG structures had best RMSD structures of 10.6 Å and 8.1 Å resolution, respectively. These were refined by Rosetta to 5.7 Å (0.7 constraints/residue) and 3.5 Å RMSD (1.0 constraints/residue), relative to the NMR solution structure (Figure 3). In both cases, the combined Rosetta method dramatically improved the DG structures, but high-resolution structures were not obtained.

While the results described above are quite promising, it is likely that still better results could be obtained by improving the basic methodology and by using additional NMR data or information from homologous structures. Rosetta best scoring and best RMSD structures *do not* completely satisfy the long range NOE constraints, with average distance violations ranging from 0.1–0.8 Å per constraint. It is likely that the conformational search strategy could be considerably improved: experimental constraint violations could be used to direct fragment insertion (a stochas-

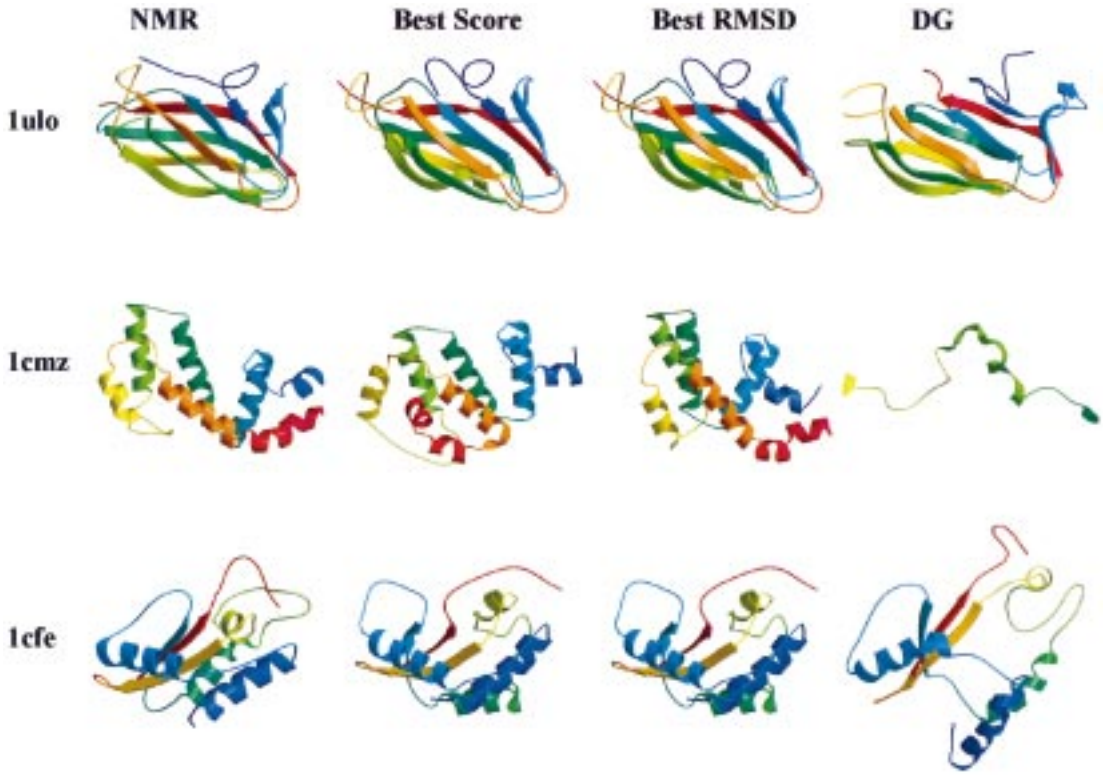

Figure 3. Comparison of Rosetta and DG results for 1ulo, 1cmz, and 1cfe (top to bottom). From left to right: the reference NMR solution structure, the best scoring and best RMSD Rosetta structures, and the best RMSD X-PLOR structure (of 10 structures). Because the 1cmz experimental data set contains no long range constraints, the best DG structure does not form a compact folded structure and does not fit entirely within the figure frame. Despite the absence of long range constraint information, Rosetta generates a best RMSD structure with secondary structure and topology quite similar to the native fold (5.9 Å). Synthetic NOE constraint sets (∼1 constraint per residue) were used to calculate the 1ulo Rosetta and DG structures, whereas experimental (real) NOE constraint sets were used to determine the 1cfe and 1cmz Rosetta and DG structures shown in the figure. Note that the best scoring and best RMSD structures for 1cfe and 1ulo are the same structure.

tic process in the current Rosetta method) and NMR data such as dipolar coupling lends itself to targeted searches of conformational space (Clore et al., 1999; Delaglio et al., 2000). Additional NMR data, such as HN-methyl and methyl-methyl NOEs (Gardner et al., 1997) and residual dipolar couplings, could improve the resolution of the structures considerably. Moreover, including fragments of homologous structures where available would considerably increase the accuracy of the model proteins. Fragments derived from proteins with greater than 25% sequence similarity to the query protein were removed in this de novo study to avoid biasing our results. With the inclusion of such fragments, the method is able to produce and identify models with <1.0 Å accuracy for some of the proteins in our study.

Our methodology combines the strengths of previous approaches (Skolnick et al., 1997; Kolinski and Skolnick, 1998; Debe et al., 1999; Standley et al., 1999; Delaglio et al., 2000). NMR data is used to improve the search of conformational space in two ways: the fragment libraries from which structures are built are made to be consistent with chemical shift and local NOE information, and a sparse NOE constraint score added to the composite Rosetta scoring function enhances identification of structures that closely resemble the native fold (Figure 1, N). The combined Rosetta method could be easily adapted to incorporate constraints derived from other experimental sources such as mass spectrometry cross-linking data (Young et al., 2000), metal binding, disulfide bonds, chemical shift perturbation and line-broadening data from spin-label NMR experiments. Distance constraints derived from sequence homologs of known structure could also be adapted for this approach. Because the method allows structures to be generated for proteins without collecting and assigning large constraint sets, the Rosetta method represents a reasonable av-

enue towards genome-scale structure determination using distance constraints collected using automated techniques (Young et al., 2000). Increased resonance degeneracy, line broadening, and the need to collect large constraint sets severely complicate NMR structure determination of large proteins. By rapidly providing moderate resolution structures using small constraint sets, Rosetta promises to speed up the assignment and high-resolution structure determination of large proteins.

Rosetta

Please contact David Baker at dabaker@u.washington.edu for access to the Rosetta algorithm for use in structure determination.

Acknowledgements

The authors thank Ad Bax, Rachel Klevit, Lewis Kay, and Lawrence McIntosh for generously providing chemical shift assignments, experimentally determined NOE constraints, and NMR solution structures. We also thank Rachel Klevit, Ponni Rajagopal, Carol Rohl and Brian Kuhlman for their helpful comments and discussion, and Brian Kuhlman for the side-chain packing routines.

References

Battiste, J.L. and Wagner, G. (2000) *Biochemistry*, **39**, 5355–5365.

Brunger, A.T. (1992) *X-PLOR version 3.1: A system for X-ray Crystallography and NMR*, Yale University, New Haven, CT.

Clore, G.M., Starich, M.R., Bewley, C.A., Cai, M. and Kuszewski, J. (1999) *J. Am. Chem. Soc.*, **121**, 6513–6514.

Cornilescu, G., Delaglio, F. and Bax, A. (1999) *J. Biomol. NMR*, **13**, 289–302.

Debe, D.A., Carlson, M.J., Sadanobu, J., Chan, S.I. and Goddard III, W.A. (1999) *J. Phys. Chem.*, **B103**, 3001–3008.

Delaglio, F., Kontaxis, G. and Bax, A. (2000) *J. Am. Chem. Soc.*, **122**, 2142–2143.

Doreleijers, J.F., Raves, M.L., Rullmann, T. and Kaptein, R. (1999) *J. Biomol. NMR*, **14**, 123–132.

Gardner, K.H., Rosen, M.K. and Kay, L.E. (1997) *Biochemistry*, **36**, 1389–1401.

Kolinski, A. and Skolnick, J. (1998) *Proteins*, **32**, 475–494.

Kraulis, P.J. (1991) *J. Appl. Crystallogr.*, **24**, 946–950.

Kuhlman, B. and Baker, D. (2000) *Proc. Natl. Acad. Sci. USA*, **97**, 10383–10388.

Merritt, E.A. and Murphy, M.E.P. (1994) *Acta Crystallogr.*, **D50**, 869–873.

Rosen, M.K., Gardner, K.H., Willis, R.C., Paris, W.E., Pawson, T. and Kay, L.E. (1996) *J. Mol. Biol.*, **263**, 627–636.

Simons, K.T., Bonneau, R., Ruczinski, I. and Baker, D. (1999a) *Proteins*, Suppl., 171–176.

Simons, K.T., Kooperberg, C., Huang, E. and Baker, D. (1997) *J. Mol. Biol.*, **268**, 209–225.

Simons, K.T., Ruczinski, I., Kooperberg, C., Fox, B.A., Bystroff, C. and Baker, D. (1999b) *Proteins*, **34**, 82–95.

Skolnick, J., Kolinski, A. and Ortiz, A.R. (1997) *J. Mol. Biol.*, **265**, 217–241.

Standley, D.M., Eyrich, V.A., Felts, A.K., Friesner, R.A. and McDermott, A.E. (1999) *J. Mol. Biol.*, **285**, 1691–1710.

Venters, R.A., Huang, C.C., Farmer II, B.T., Trolard, R., Spicer, L.D. and Fierke, C.A. (1995) *J. Biomol. NMR*, **5**, 339–344.

Young, M.M., Tang, N., Hempel, J.C., Oshiro, C.M., Taylor, E.W., Kuntz, I.D., Gibson, B.W. and Dollinger, G. (2000) *Proc. Natl. Acad Sci. USA*, **97**, 5802–5806.