Accurate protein structure modeling using sparse NMR data and homologous structure information

James M. Thompson[a,1], Nikolaos G. Sgourakis[a,1], Gaohua Liu[b], Paolo Rossi[b], Yuefeng Tang[b], Jeffrey L. Mills[c], Thomas Szyperski[c], Gaetano T. Montelione[b], and David Baker[a,d,2]

[a]Department of Biochemistry, University of Washington, Seattle, WA 98195; [b]Department of Molecular Biology and Biochemistry, Center for Advanced Biotechnology and Medicine, Northeast Structural Genomics Consortium, Rutgers, The State University of New Jersey, and Robert Wood Johnson Medical School, University of Medicine and Dentistry of New Jersey, Piscataway, NJ 08854; [c]Department of Chemistry, Northeast Structural Genomics Consortium, The State University of New York at Buffalo, Buffalo, NY 14260; and [d]Howard Hughes Medical Institute, University of Washington, Seattle, WA 98195

Edited by Adriaan Bax, National Institutes of Health, Bethesda, MD, and approved April 20, 2012 (received for review February 13, 2012)

While information from homologous structures plays a central role in X-ray structure determination by molecular replacement, such information is rarely used in NMR structure determination because it can be incorrect, both locally and globally, when evolutionary relationships are inferred incorrectly or there has been considerable evolutionary structural divergence. Here we describe a method that allows robust modeling of protein structures of up to 225 residues by combining $^{1}H^{N}$, ^{13}C, and ^{15}N backbone and $^{13}C\beta$ chemical shift data, distance restraints derived from homologous structures, and a physically realistic all-atom energy function. Accurate models are distinguished from inaccurate models generated using incorrect sequence alignments by requiring that (*i*) the all-atom energies of models generated using the restraints are lower than models generated in unrestrained calculations and (*ii*) the low-energy structures converge to within 2.0 Å backbone rmsd over 75% of the protein. Benchmark calculations on known structures and blind targets show that the method can accurately model protein structures, even with very remote homology information, to a backbone rmsd of 1.2–1.9 Å relative to the conventional determined NMR ensembles and of 0.9–1.6 Å relative to X-ray structures for well-defined regions of the protein structures. This approach facilitates the accurate modeling of protein structures using backbone chemical shift data without need for side-chain resonance assignments and extensive analysis of NOESY cross-peak assignments.

biochemistry | biophysics | computational biology | nuclear magnetic resonance | structural genomics

In recent years, the application of multidimensional data collection techniques in isotopically enriched proteins (1) as well as development of selective labeling schemes in perdeuterated samples (2–5) and other methodological improvements (6, 7) have allowed the application of NMR methods to larger proteins (8–10). Conventional NMR structure determination relies primarily on the availability of distance restraints from NOESY experiments, which requires time-consuming experiments, including extensive analysis of side-chain resonance assignments and laborious assignments of most of the observed NOESY cross-peak resonances. While automated assignment methods (11–13) have greatly stream-lined the process (14), the assignment of NOE cross-peaks in spectra of larger proteins presents a significant challenge due to increased spectral overlap, line broadening, and low signal-to-noise ratios, rendering existing automated assignment methods ineffective in the absence of a preliminary structural model. Accurate structures can be generated for small proteins (up to 100–120 residues) using chemical shift information to guide structure prediction calculations (15, 16), but additional NMR data, including backbone NOEs and residual dipolar coupling (RDC) data, are required to obtained converged structures for larger proteins (17) and protein oligomers (18). Such data are often hard to collect and analyze, which hinders the automation of these methods. The development of assignment-independent methods is a powerful alternative to conventional NMR structure determination (19, 20), but such methods have only been applied for small, globular proteins. Finally, even with the assignment of NOE data available, the use of NOE restraints for structural analysis suffers from problems such as conformational pinning (21) and spin diffusion (22), while ensemble dynamics cannot be adequately represented using conventional methods due to the optimization of a single target function to the experimental data (21). The solution of these drawbacks is nontrivial and can be ameliorated using inferential structure determination techniques (23) and ensemble averaging methods (24–26), but these calculations are inherently limited to smaller-size systems due to the large number of free parameters.

Considerable information is often available in the structures of evolutionarily related proteins. This information has long been used to generate molecular replacement models to phase X-ray diffraction datasets, and with recent advances this can be accomplished even when the evolutionary relationships are quite remote (~15%–25% sequence identity) (27, 28). Although the use of homologous X-ray structures to improve and refine existing NMR structures has been previously described (29), this information has not been used for *de novo* structure modeling. We reasoned that this information should also be useful for guiding CS-Rosetta structure predictions, and investigated the possibility of accurately modeling protein structures using only the chemical shifts of backbone atoms (H^{N}, N, Cα, Cβ, C'), distance restraints derived from homologous proteins of known structure, and the Rosetta sampling methodology and all atom energy function (30) (see *Methods*).

An obvious concern in using evolutionary information in structure modeling is the potential for error due to sequence alignment inaccuracies, both locally (local alignment mismatches) and globally (incorrect fold in the identified template structure). A useful method must be both robust to input alignment errors and have reliable metrics for assessing the accuracy of the resulting models, in particular indicating when a structure calculation is likely to have significant errors. Energy functions have been used to a limited degree for distinguishing correct from incorrect sequence alignments in homology modeling calculations (31), while in molecular replacement methods for X-ray crystallography, the fit to the diffraction data distinguishes correct from incorrect homologous structure information (32). Here we show that the combination of rapidly obtained backbone chemical shift

Author contributions: J.M.T., N.G.S., and D.B. designed research; J.M.T. and N.G.S. performed research; J.M.T., N.G.S., G.L., P.R., Y.T., J.L.M., T.S., and G.T.M. contributed new reagents/analytic tools; J.M.T., N.G.S., T.S., G.T.M., and D.B. analyzed data; and J.M.T., N.G.S., G.T.M., and D.B. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

[1]J.M.T. and N.G.S. contributed equally to this work.

[2]To whom correspondence should be addressed. E-mail: dabaker@u.washington.edu.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1202485109/-/DCSupplemental.

data together with the Rosetta energy functions provides a robust and reliable method for distinguishing correct from incorrect homology information and for generating accurate homology-based models of protein structures.

Results

To investigate the utility of restraints derived from homologous structures in guiding CS-Rosetta calculations, which we refer to here as the CS-HM-Rosetta method, we carried out structure calculations for four proteins of known structure using experimentally measured chemical shifts and between 8 and 12 independent sets of distance restraints based on individual alignments ranging from very remote (10% sequence identity) to significant (30% sequence identity) homology (Fig. 1 and Fig. S1). Several of the input alignments contained significant errors that would, in the absence of experimental information, result in incorrect homology-based models. CS-Rosetta with chemical shift information alone (15) (no homology information) fails to converge to a well-defined structural ensemble in all four cases. As indicated in Fig. 1*A*, structures generated using correct alignments (green) have lower Rosetta full-atom energies than structures generated using either no evolutionary information (blue) or information from incorrect alignments (red). Structures generated using restraints from correct alignments are closer to the native structure (Fig. 1*B*), and the lowest-energy models closely matched the conventionally determined NMR structure (Table 1). The observation of lower energies in the restrained calculations is nontrivial, as restrained optimization generally results in poorer energy minimization than unrestrained optimization. Improved optimization with restraints suggests that the reduced conformational freedom, due to the use of restraints, is more than compensated by improved sampling nearer the low-energy native state (17). Thus, the decrease in energy compared to unrestrained calculations provides a first metric for evaluating the structures generated by the CS-HM-Rosetta method.

We reasoned that calculations based on accurate alignments of the correctly identified fold should converge more strongly than calculations based on incorrect alignments, as restraints derived from incorrect alignments are likely to be in conflict with the experimental chemical shift data and the full-atom energy function. Such conflict should result in a rugged landscape for optimization, as biasing the calculations toward the incorrect fold by the restraints likely results in structures with poorly optimized backbone, side-chain, and solvation energy terms. Indeed, for each case illustrated in Fig. 1 and Fig. S1, the convergence of the calculations increased with increasing alignment quality [Fig. 1*C*; compare accuracy (*x* axis) of blue (poorly converged) and yellow (highly converged) calculations]. Taken together, the decrease in

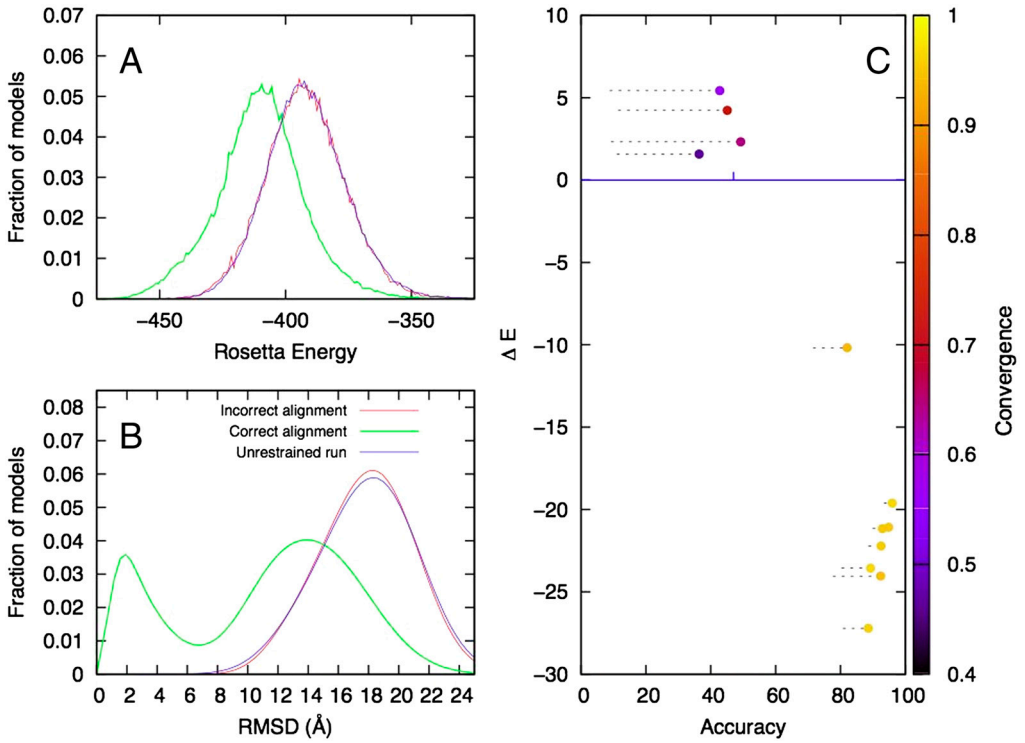

Fig. 1. Influence of alignment quality on CS-HM-Rosetta calculations. (*A*, *B*) Starting from the sequence of the proteasome protomer, remote homology searches were carried out and alignments to two distant homologues (less than 20% sequence identity) were selected which covered at least 70% of the sequence; one of these was basically correct, and the other incorrect (different fold than the native structure). CS-HM-Rosetta calculations were carried out using no homology restraints (blue lines), the correct remote homology information (green lines), or the incorrect remote homology information (red lines). (*A*) Restraints from remote homologues with the correct fold allow Rosetta to more effectively minimize the full atom energy. The energy distribution is shifted to lower energies in the correctly restrained calculations (green) than in the incorrectly restrained (red) or unrestrained (blue) calculations. (*B*) Lower backbone rmsd (relative to the X-ray structure) models are produced using accurate restraints than with no restraints or inaccurate restraints. (*C*) Relationship between calculation accuracy and the extent of convergence and energy drop relative to unconstrained calculations. Each point represents an independent calculation using chemical shift data and restraints from single alignments to different templates with various levels of accuracy. The *x* axis: fraction of residues that superimpose within 2.0 Å backbone rmsd to the native structure; *y* axis: average Rosetta full atom energy (*y* axis) of the lowest energy ten models from each calculation. The color scale represents the extent of convergence of the calculation (the percent of residues that superimpose to an average structure with less than 2.0 Å rmsd). The blue line represents the average Rosetta energy of models built without homology restraints, and the dashed lines show the improvement in accuracy over models built using only homology information (35). Using chemical shift data and accurate homology information gives the most accurate models, and successful simulations consistently yield both lower Rosetta energies and a highly converged structural ensemble. The rmsds on the *x* axis are to the X-ray structures for the monomeric α subunit extracted from the proteasome complex structure (36) (PDB ID 1ya7). Chemical shift assignments were obtained from previous work (9).

Table 1. Convergence and accuracy of models generated with chemical shift data and homology information

Target	PDB code (NMR/X-ray)	Convergence*	Accuracy (NMR/X-ray)	Fold type	Backbone rmsd†
CsR251	**N/A**	**104/126**	**2.26**	**α/β**	**2.0**
HR4403E	2lni	108/122	1.95	α/β	0.4
LpR145J	**2lfc**	**107/138**	**1.95**	**α/β**	**2.0**
HR5460A	2lah	114/146	1.29	α/β	1.0
Tpx	2jsz	122/167	1.69	α/β	1.0
ER553	2k1s	111/143	1.62	α/β	0.9
WR73‡	**2kwb**	**109/139**	**2.52**	**α/β**	**1.8**
CgR26a	**2kpt**	**114/115**	**1.46**	**α/β**	**2.0**
T0475	2k54	91/120	1.29	α/β	1.2
SeR147	2l9p	116/151	1.56	α/β	0.8
RhoA	**NA/1a2b**	**143/181**	**1.60**	**α/β**	**2.0**
Rhodopsin	2ksy/1f88	159/222	1.89/1.48 (1.76)	α	1.5
SgR145‡	2kw5/3mer	134/152	2.37/1.16 (2.38)	α/β	1.3
fgf2	**1bla/1bas**	**98/125**	**1.45/1.41 (0.75)**	β	**2.0**
Antigen-1	1dgq/1zon	147/189	1.23/1.03 (1.04)	α/β	1.0
Ribonuclease	2aas/1kf5	93/124	1.25/0.90 (0.98)	α/β	1.5
Hemoglobin	1vre/1jf4	106/147	1.84/1.64 (1.43)	α	1.3

Structures determined using chemical shifts and restraints from homologs <20% sequence identity (CS-HM-Rosetta, first two columns) are compared to conventionally determined NMR or X-ray structures. All targets reported pass the convergence and energetic validation metrics (see article text). Accuracy is the median backbone rmsd to the NMR/X-ray structure over the converged part of the structure (numbers in parentheses are backbone rmsds between X-ray and NMR structures). Names of blind cases are italicized (first four rows). For all benchmark cases, the standard CS-Rosetta method fails to converge over more than 50% of the target structure. While high-resolution structural ensembles are obtained using remote homology information for 11/17 cases (category 1 in main text), for the remaining 6/17 cases (highlighted in bold) the resulting ensembles are not as well-converged and are perhaps best used as preliminary models toward refinement using additional experimental data (category 2 in the main text). N/A, not applicable.

*Convergence here is the fraction of total residues that superimpose within the rmsd threshold of column 6 in the low-energy CS-HM-Rosetta ensemble.

†Rmsd computed for the backbone of ordered portions of the CS-HM-Rosetta ensemble.

‡Regions disordered in the NMR ensemble were excluded from this analysis; convergence values are reported relative to the adjusted sequence length.

energy relative to unconstrained calculations (*y* axis in Fig. 1*C*; the energy of the unconstrained calculation is indicated by the horizontal blue line) and the convergence clearly discriminate accurate structures (yellow points at bottom right) from inaccurate structures (purple points at top left).

We defined two criteria for assessing structure calculations using homology information: first, the lowest energy 10 models should converge to within 2.0 Å over at least 75% of the protein, and second, the energy should be lower than in unrestrained calculations. Over the 48 simulations performed, structure calculations which satisfied these two criteria (Fig. 2*A*, solid line) were almost always quite close to the independently determined conventional NMR or X-ray crystal structure, while structure calculations which did not satisfy one or both of the criteria were in some cases quite inaccurate (Fig. 2*A*, dotted line). The two validation metrics successfully diagnose even a pathological case where the evolutionary information, despite being based on very high sequence identity, is catastrophically wrong (Fig. S2). A failure to satisfy both criteria using information from homologous structures is a clear indication that additional experimental data are required to solve the structure of a protein. When complementary data such as RDCs are available, they can provide an additional validation metric (Fig. S3).

The joint optimization of the largely orthogonal evolutionary restraints, backbone chemical shift data, and the physically realistic energy function surmounts many inaccuracies (incorrectly placed insertions and deletions, register shifts, etc.) in the input sequence alignments. Models generated with homology information and chemical shift data are almost always better than the models generated using homology information alone (Fig. 1 *C* and Fig. S3, dashed lines connect models generated without and with chemical shift information; and Fig. 2*B*). The improvement in model quality due to the chemical shift information holds over the full range of sequence identity (Fig. 2 *B* and *C*). There are large improvements for very low-identity sequence alignments (yellow and red points, Fig. S4*B*, left-hand section of Fig. 2*C*), but these calculations often fail the two metrics and hence the resulting models are likely to be useful only in the context of additional experimental data (e.g., with additional selected labeling NOESY-derived restraints; manuscript in preparation). For sequence alignments of 20% sequence identity or above, there are also clear improvements over models based on homology information alone (Fig. S5, green and blue points above diagonal in Fig. S4), and calculations that pass the two metrics consistently are reasonably accurate (Fig. 2*B*, *y* axis). Employing the validation criteria rescues (Fig. 2*B*, dashed line, and *C*) the steep fall-off in model accuracy below 20% sequence identity (Fig. 2 *B*, dotted line, and *C*) that has plagued comparative modeling since its inception. Similarly, the calculations performed using chemical shift data alone do not converge to near-native conformations due to the magnitude of the conformational sampling problem and generally result in a low-precision ensemble with higher full-atom energies (blue lines in Fig. 1 and Fig. S1). Overall, the CS-HM-Rosetta method is a clear improvement over previous methods using either chemical shift data alone (CS-Rosetta) or conventional homology modeling.

After establishing the two validation measures, we carried out CS-HM-Rosetta structure calculations on a benchmark of 13 proteins, with experimentally determined structures using backbone chemical shifts and distance restraints from template structures with less than 20% sequence identity to the query (Table 1 and Fig. S4). We focused on alignments with less than 20% sequence identity, as for closer homologues, good models can be generated using homology information alone (Fig. 2 *A* and *B* and Fig. S4). We selected larger proteins where the standard CS-Rosetta method fails to converge due to sampling limitations, even with the use of RDCs. For all of the targets meeting both validation criteria, we obtain models that are consistently very similar (1.25–2.5 Å backbone rmsd) to structures determined previously using conventional NMR protocols. In the five cases where high-resolution X-ray structures are available (bottom of Table 1), the accuracy of the CS-HM-Rosetta ensembles falls within 0.9–1.6 Å backbone rmsd for well-defined regions of the structure. This compares well with backbone rmsds of 0.5–1.5 Å observed in a case study of some 230 pairs of NMR/X-ray crystal structures of identical proteins (33), particularly considering that

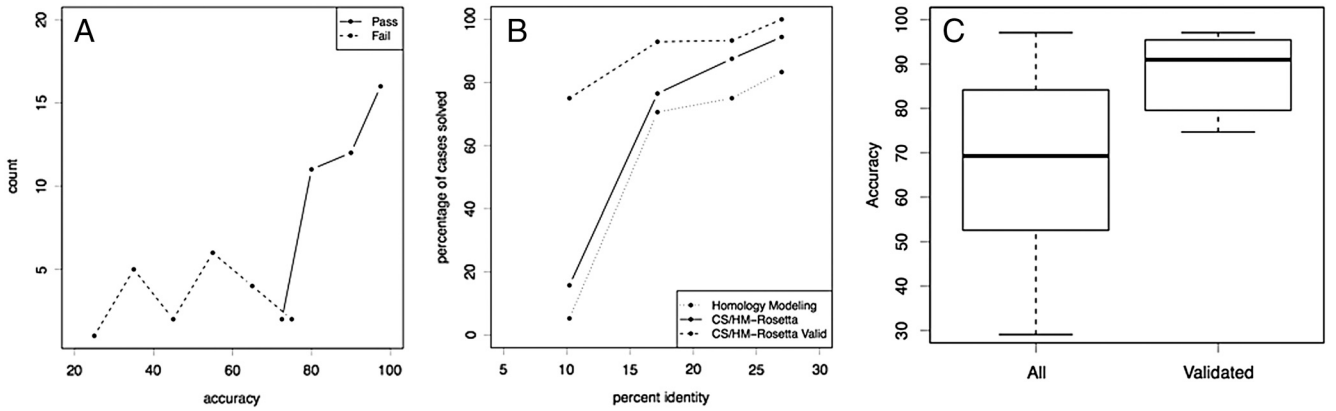

Fig. 2. Validation metrics allow discrimination of accurate from inaccurate calculated structures. Independent calculations were performed using chemical shift data and homology modeling restraints derived from templates of varying evolutionary distance (37) (10–30% sequence identity) using the benchmark targets described in Table 1. From each simulation, the lowest 10 models by Rosetta full atom energy were used as a structural ensemble. (*A*) Calculated structures satisfying both validation criteria are close to the native structure. Ensembles from CS-HM-Rosetta were validated using two metrics: the lowest 10 models must superimpose within 2.5 Å over at least 75% of the structure, and the same models must be lower in Rosetta energy than any models from an unrestrained simulation that fails to converge. Accuracy was measured as the percentage of residues that superimpose to the ordered NMR ensemble. Ensembles that are valid by both criteria (solid line) are significantly more accurate than ensembles that fail at least one of the criteria (dashed line). (Kolmogorov-Smirnov one-sided *P*-value = 1.477*e*-10). (*B*) Calculated structures satisfying both validation criteria avoid the sharp drop in model accuracy at low sequence identity. Cases were deemed successful if they superimposed within 2.5 Å over 75% of the native structure. Each low-energy ensemble was placed in a bin defined by the sequence identity of the homologous structure used in the simulation. At 20% sequence identity, there is a sharp drop-off in accuracy for both homology modeling (dotted line) and CS-HM-Rosetta (solid line), which is rescued by application of the validation criteria (*B*, dashed line). (*C*) Over the full set of alignments from 10–30% sequence identity, the variance in model accuracy is quite high (left bar), but with the application of the validation criteria (right bar), the variance decreases considerably and the average accuracy increases dramatically

most of the proteins assessed in this previous case study were smaller than the proteins used in this CS-HM-Rosetta study. In cases where X-ray structures of the same targets are also available (Table 1), the structural ensemble modeled using the CS-HM-Rosetta method is often closer (in terms of backbone rmsd) to the X-ray structure than the conventionally determined NMR ensemble (Table 1 and Fig. 3*D*). Moreover, the CS-HM-Rosetta structures show a high degree of side-chain convergence to the rotamers observed in the X-ray or NMR structures (Fig. 3 *C* and *D*); indeed, the CS-HM-Rosetta structures have more side-chains in the same rotamer conformation as in the X-ray structures than the conventional NMR structure (Table S1). The observation that highly converged CS-HM-Rosetta ensembles have accurate side-chain conformations suggests the classification of CS-HM-Rosetta ensembles into three categories (Table S2): (1) lower energy after adding restraints and convergence to <1.5 Å over at least 75% of the structure: as well determined as conventional NMR structures; (2) lower energy after adding restraints and convergence between 1.5 and 2.0 Å over 75% of the structure: correct overall structure, but less accurate than conventional NMR; and (3) energy not lower with restraints and convergence worse than 2.0 Å over 75% of the structure: more experimental data required. Class (2) ensembles can be very useful intermediate steps in structure determination using additional experimental data, and could provide valuable starting points for automated NOE-assignment methods.

In order to assess the practical performance of CS-HM-Rosetta, we calculated structures for four mid-size targets whose structures had not yet been published (or in some cases even solved) using backbone chemical shift data (but no other experimental data) from expert laboratories. These particular cases were selected for this study because they do not converge using the standard CS-Rosetta protocol. The low energy converged structures were subsequently sent to the laboratories who had independently solved the same structures by conventional NMR methods, including extensive analysis of side-chain assignments and NOESY data, for evaluation. For all four targets (Table 1 and Fig. 3), the CS-HM-Rosetta calculations converged on an ensemble in good agreement with the structures determined independently in the expert laboratories. According to the validation criteria, two of the ensembles are category 1 (high accuracy) and two are category 2. The CS-HM-Rosetta structures agree as well with unassigned NOESY data as conventional NMR models determined using assigned NOEs (Table S3). When unassigned NOESY data are available, metrics such as the DP score which assess the fit between structure models and unassigned NOESY peak lists can provide a third criterion for assessing

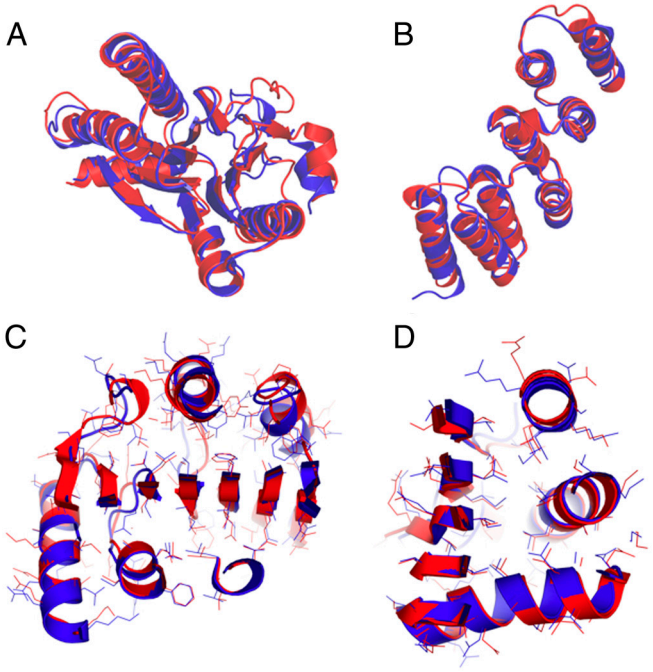

Fig. 3. Structural comparison of high-resolution CS-HM-Rosetta structures with conventionally determined NMR (*A*, *B*, and *D*) and X-ray (*C*) structures. CS-HM-Rosetta structures are shown in red, and conventionally determined structure is shown in blue. (*A*) Proteasome monomer. (*B*) HR5460a. (*C*) SgR145. (*D*) ER553.

model quality (34). The CS-HM-Rosetta models also showed good agreement with the RDC data (Fig. S3).

Discussion

We have shown that reasonably accurate models of proteins up to 25 kDa can be obtained by using remote homology information to guide CS-Rosetta structure calculations (30). For proteins larger than 100 residues, the CS-HM-Rosetta method is a significant improvement over CS-Rosetta. The final structures agree well with structures determined using state-of-the-art conventional NMR methods and X-ray crystallography. The validation metrics (ensemble convergence and improved optimization of the full-atom energy) distinguish between cases where the method produces an accurate structure and cases where the protocol fails due to inadequate sampling or incorrect homology information (arising from inaccuracies in the input sequence alignments), allowing confident utilization of the wealth of structural data currently available for structure determination. Overall, this method is a powerful approach for protein NMR data analysis that does not require determination of side-chain resonance assignments. The CS-HM-Rosetta method provides high-quality models of protein structures, with accurate core side-chain structures, approaching the quality that is obtained with full side-chain assignments and much more extensive NOESY analysis.

Methods

CS-HM-Rosetta Input Data. Alignments to templates were generated using standard methods. Interatomic distance restraints were derived from templates using a previously described method (27) that models restraints as a mixture of Gaussian probability densities, where the weight and width of each restraint is based on the estimated probability that the restraint is correct. In this method, restraints operating on the same pair of atoms are combined from multiple templates by awarding higher weight to restraints that are more likely to be correct. Fragments were derived using the standard CS-Rosetta method (15), which scores candidate fragments by measuring their agreement with the observed sequence profile, predicted secondary structure, and chemical shift data. The use of RDC data from an arbitrary number of alignment media is also supported in a simple input format.

CS-HM-Rosetta Simulations. Full-atom models were constructed using the previously described CS-Rosetta procedure (15, 17), with one modification to include the restraints from template structures as an extra scoring term (27). The protocol consists of a low-resolution stage in which side-chains are represented by a centroid atom with radius depending on the amino acid residue identity. Structures are assembled starting from an extended chain by fragment insertion under a low-resolution force field that favors structural compactness and formation of secondary structure elements. In order to avoid frustration of the optimization process, restraints are gradually incorporated into the protocol by activating restraints between adjacent residues early in the simulation and progressively including restraints between residues more distant in sequence. Low-resolution modeling is followed by a high-resolution refinement stage during which all heavy and hydrogen atoms in the protein are explicitly represented, and the backbone and side-chain torsion angles are refined in the presence of the Rosetta full-atom energy function. The use of RDCs during the low-resolution and high-resolution refinement stages is optionally performed through the addition of a pseudoenergy term that measures the back-calculated RDCs from the current structure with the experimental RDCs. Models are selected according to the Rosetta full-atom energy function (30), and convergence is measured as the proportion of residues superimposable within 2.0 Å among the lowest 10 models by Rosetta energy. CS-Rosetta models were generated using the same procedure without homology restraints, and homology models were generated using Modeller (35). Software and data are available upon request from the authors, and a detailed description of the method is available in the *SI Text* and Table S4.

ACKNOWLEDGMENTS. We thank Microsoft Windows Azure and the users of Rosetta@Home for their generous donation of computing resources. We thank Dr. Yuanpeng Huang, Binchen Mao, and Robert Vernon for helpful discussions and comments on the manuscript. This work was supported by the Howard Hughes Medical Institute and National Institutes of Health (NIH) Grant 1R01-GM092802-01 (D.B.), and the NIH PSI-Biology Grant U54-GM094597 (G.T.M. and T.S.).

1. Kay LE, Clore GM, Bax A, Gronenborn AM (1990) 4-Dimensional heteronuclear triple-resonance NMR-spectroscopy of interleukin-1-beta in solution. *Science* 249:411–414.
2. Xu R, Ayers B, Cowburn D, Muir TW (1999) Chemical ligation of folded recombinant proteins: segmental isotopic labeling of domains for NMR studies. *Proc Natl Acad Sci USA* 96:388–393.
3. Kay LE, Gardner KH (1997) Solution NMR spectroscopy beyond 25 kDa. *Curr Opin Struct Biol* 7:722–731.
4. Goto NK, Gardner KH, Mueller GA, Willis RC, Kay LE (1999) A robust and cost-effective method for the production of Val, Leu, Ile (delta 1) methyl-protonated N-15-, C-13-, H-2-labeled proteins. *J Biomol NMR* 13:369–374.
5. Kainosho M, et al. (2006) Optimal isotope labelling for NMR protein structure determinations. *Nature* 440:52–57.
6. Bertini I, et al. (2011) Solid-state NMR of proteins sedimented by ultracentrifugation. *Proc Natl Acad Sci USA* 108:10396–10399.
7. Pervushin K, Riek R, Wider G, Wuthrich K (1997) Attenuated T-2 relaxation by mutual cancellation of dipole-dipole coupling and chemical shift anisotropy indicates an avenue to NMR structures of very large biological macromolecules in solution. *Proc Natl Acad Sci USA* 94:12366–12371.
8. Fiaux J, Bertelsen EB, Horwich AL, Wuthrich K (2002) NMR analysis of a 900 K GroEL GroES complex. *Nature* 418:207–211.
9. Sprangers R, Kay LE (2007) Quantitative dynamics and binding studies of the 20 S proteasome by NMR. *Nature* 445:618–622.
10. Keramisanou D, et al. (2006) Disorder-order folding transitions underlie catalysis in the helicase motor of SecA. *Nat Struct Mol Biol* 13:594–602.
11. Herrmann T, Guntert P, Wuthrich K (2002) Protein NMR structure determination with automated NOE assignment using the new software CANDID and the torsion angle dynamics algorithm DYANA. *J Mol Biol* 319:209–227.
12. Ikeya T, et al. (2011) Exclusively NOESY-based automated NMR assignment and structure determination of proteins. *J Biomol NMR* 50:137–146.
13. Huang YJ, Tejero R, Powers R, Montelione GT (2006) A topology-constrained distance network algorithm for protein structure determination from NOESY data. *Proteins* 62:587–603.
14. Rosato A, et al. (2012) Blind testing of routine, fully automated determination of protein structures from NMR data. *Structure* 20:227–236.
15. Shen Y, et al. (2008) Consistent blind protein structure generation from NMR chemical shift data. *Proc Natl Acad Sci USA* 105:4685–4690.
16. Cavalli A, Salvatella X, Dobson CM, Vendruscolo M (2007) Protein structure determination from NMR chemical shifts. *Proc Natl Acad Sci USA* 104:9615–9620.
17. Raman S, et al. (2010) NMR structure determination for larger proteins using backbone-only data. *Science* 327:1014–1018.
18. Sgourakis NG, et al. (2011) Determination of the structures of symmetric protein oligomers from NMR chemical shifts and residual dipolar couplings. *J Am Chem Soc* 133:6288–6298.
19. Grishaev A, Llinas M (2002) CLOUDS, a protocol for deriving a molecular proton density via NMR. *Proc Natl Acad Sci USA* 99:6707–6712.
20. Meiler J, Baker D (2003) Rapid protein fold determination using unassigned NMR data. *Proc Natl Acad Sci USA* 100:15404–15409.
21. Snyder DA, Bhattacharya A, Huang YJ, Montelione GT (2005) Assessing precision and accuracy of protein structures derived from NMR data. *Proteins* 59:655–661.
22. Cavanagh J, Fairbrother WJ, Palmer AG, Rance M, Skelton NJ (2007) *Protein NMR Spectroscopy* (Elsevier Academic Press, London), pp 553–678.
23. Rieping W, Habeck M, Nilges M (2005) Inferential structure determination. *Science* 309:303–306.
24. Torda AE, Scheek RM, van Gunsteren WF (1990) Time-averaged nuclear Overhauser effect distance restraints applied to tendamistat. *J Mol Biol* 214:223–235.
25. Richter B, Gsponer J, Varnai P, Salvatella X, Vendruscolo M (2007) The MUMO (minimal under-restraining minimal over-restraining) method for the determination of native state ensembles of proteins. *J Biomol NMR* 37:117–135.
26. Bonvin AM, Brunger AT (1995) Conformational variability of solution nuclear magnetic resonance structures. *J Mol Biol* 250:80–93.
27. Thompson J, Baker D (2011) Incorporation of evolutionary information into Rosetta comparative modeling. *Proteins* 79:2380–2388.
28. Sali A, Blundell TL (1993) Comparative protein modelling by satisfaction of spatial restraints. *J Mol Biol* 234:779–815.
29. Brunner K, Gronwald W, Trenner JM, Neidig KP, Kalbitzer HR (2006) A general method for the unbiased improvement of solution NMR structures by the use of related X-ray data, the AUREMOL-ISIC algorithm. *BMC Struct Biol* 6:14.
30. Bradley P, Misura KM, Baker D (2005) Toward high-resolution de novo structure prediction for small proteins. *Science* 309:1868–1871.
31. Sahasrabudhe PV, Tejero R, Kitao S, Furuichi Y, Montelione GT (1998) Homology modeling of an RNP domain from a human RNA-binding protein: Homology-constrained energy optimization provides a criterion for distinguishing potential sequence alignments. *Proteins* 33:558–566.
32. McCoy AJ, et al. (2007) Phaser crystallographic software. *J Appl Crystallogr* 40:658–674.
33. Andrec M, et al. (2007) A large data set comparison of protein structures determined by crystallography and NMR: statistical test for structural differences and the effect of crystal packing. *Proteins* 69:449–465.

34. Huang YJ, Powers R, Montelione GT (2005) Protein NMR recall, precision, and F-measure scores (RPF scores): structure quality assessment measures based on information retrieval statistics. *J Am Chem Soc* 127:1665–1674.
35. Eswar N, et al. (2007) Comparative protein structure modeling using MODELLER. *Curr Protoc Protein Sci*, Chap 2:Unit 2 9; 10.1002/0471140864.ps0209s50.
36. Forster A, Masters EI, Whitby FG, Robinson H, Hill CP (2005) The 1.9 A structure of a proteasome-11S activator complex and implications for proteasome-PAN/PA700 interactions. *Mol Cell* 18:589–599.
37. Zhang Y, Skolnick J (2005) TM-align: a protein structure alignment algorithm based on the TM-score. *Nucleic Acids Res* 33:2302–2309.