Consistent blind protein structure generation from NMR chemical shift data

Yang Shen*, Oliver Lange†, Frank Delaglio*, Paolo Rossi‡, James M. Aramini‡, Gaohua Liu‡, Alexander Eletsky§, Yibing Wu§, Kiran K. Singarapu§, Alexander Lemak¶, Alexandr Ignatchenko¶, Cheryl H. Arrowsmith¶, Thomas Szyperski§, Gaetano T. Montelione‡, David Baker†‖, and Ad Bax*‖

*Laboratory of Chemical Physics, National Institute of Diabetes and Digestive and Kidney Diseases, National Institutes of Health, Bethesda, MD 20892; †Department of Biochemistry and Howard Hughes Medical Institute, University of Washington, Seattle, WA 98195; ‡Center for Advanced Biotechnology and Medicine, Department of Molecular Biology and Biochemistry, and Northeast Structural Genomics Consortium, Rutgers, The State University of New Jersey, and Robert Wood Johnson Medical School, Piscataway, NJ 08854; §Departments of Chemistry and Structural Biology and Northeast Structural Genomics Consortium, University at Buffalo, State University of New York, Buffalo, NY 14260; and ¶Ontario Cancer Institute, Department of Medical Biophysics, and Northeast Structural Genomics Consortium, University of Toronto, Toronto, ON, Canada M5G IL5

Contributed by Ad Bax, January 10, 2008 (sent for review December 14, 2007)

Protein NMR chemical shifts are highly sensitive to local structure. A robust protocol is described that exploits this relation for *de novo* protein structure generation, using as input experimental parameters the $^{13}C^{\alpha}$, $^{13}C^{\beta}$, $^{13}C'$, ^{15}N, $^{1}H^{\alpha}$ and $^{1}H^{N}$ NMR chemical shifts. These shifts are generally available at the early stage of the traditional NMR structure determination process, before the collection and analysis of structural restraints. The chemical shift based structure determination protocol uses an empirically optimized procedure to select protein fragments from the Protein Data Bank, in conjunction with the standard ROSETTA Monte Carlo assembly and relaxation methods. Evaluation of 16 proteins, varying in size from 56 to 129 residues, yielded full-atom models that have 0.7–1.8 Å root mean square deviations for the backbone atoms relative to the experimentally determined x-ray or NMR structures. The strategy also has been successfully applied in a blind manner to nine protein targets with molecular masses up to 15.4 kDa, whose conventional NMR structure determination was conducted in parallel by the Northeast Structural Genomics Consortium. This protocol potentially provides a new direction for high-throughput NMR structure determination.

molecular fragment replacement | protein structure prediction | ROSETTA | structural genomics

Over the past two decades, NMR spectroscopy has become an established complement to x-ray crystallography for determination of the three-dimensional (3D) structures of proteins at atomic resolution. The vast majority of current protein NMR structure studies today rely on the expression of recombinant protein with uniform enrichment of ^{13}C and ^{15}N stable isotopes. The first stage of the structure determination process then involves assignment of the ^{1}H, ^{15}N, and ^{13}C NMR resonances of the polypeptide backbone atoms and often can be carried out quite rapidly, using small amounts of protein. A number of procedures have been introduced in recent years that can greatly expedite this resonance assignment process (1, 2). Chemical shift values, obtained from this assignment process, reflect a wide array of structural factors including backbone and side-chain conformations, secondary structure, hydrogen bond strength, and the position of aromatic rings (3–8).

The second stage of the NMR structure determination process involves assignment of the side chain resonances and collection of structural data, including interproton distance restraints from multidimensional nuclear Overhauser enhancement (NOE) spectra. Sensitivity of such experiments tends to be lower, therefore requiring stable, relatively concentrated samples or lengthier data acquisitions, and spectra exhibit more resonance overlap, complicating analysis. Although procedures have been introduced for automated interpretation of the thousands of cross peaks in such NOE spectra, their success hinges upon the quality of the spectral data. As a result, side chain assignment together with collection and analysis of the NOE data commonly remains the limiting and most time-consuming step in the NMR structure determination process.

Multiple NMR approaches have been proposed in recent years that are all aimed at circumventing the need for collection and iterative analysis of NOE data. In one such method, orientations of bond vectors are determined from measurement of residual dipolar couplings (RDCs) (9, 10). Searching of the protein structure database (PDB) for fragments approximately compatible with these couplings and the experimental chemical shifts then yields substructures that subsequently can be assembled using a molecular fragment replacement (MFR) method into a model for the target protein (11). Completeness of the RDC data is a prerequisite when building a protein with this approach. The feasibility of alternate RDC-based procedures has also been demonstrated, but to date none of these has advanced to a robust approach for routine structure determination.

Chemical shift data can also be used to guide selection of fragments (6), which can be used in conjunction with protein sequence information and a reasonable force field to build up 3D structure models (12). Two recent studies have extended this approach considerably (13, 14). In particular, the CHESHIRE method, introduced by Vendruscolo and coworkers (13), generates all-atom structures and "refines" the model generated from selected fragments using a force field similar to the standard ones used in classical molecular dynamics simulations, while simultaneously optimizing agreement with the experimental chemical shifts. This CHESHIRE approach, demonstrated for 11 proteins in the size range of 46–123 residues, yielded results remarkably close (1.3–1.8 Å backbone atom rmsd; 2.1–2.6 Å rmsd for all atoms) to structures previously determined using conventional x-ray crystallography or NMR methods.

It has long been recognized that the 3D structure of a protein is directly related to its amino acid sequence (15). *De novo* structure predictions from solely the sequence thus provide another pathway to generate protein structural models. Among those, ROSETTA is one of the most successful programs for obtaining atomic level 3D structures of small proteins (16). For each small segment of the query protein, ROSETTA selects two hundred fragments from the crystallographic structural database that are similar in amino acid

Author contributions: Y.S., D.B., and A.B. designed research; Y.S. and O.L. performed research; F.D. contributed new reagents/analytic tools; Y.S., P.R., J.M.A., G.L., A.E., Y.W., K.K.S., A.L., and A.I. analyzed data; and Y.S., C.H.A., T.S., G.T.M., D.B., and A.B. wrote the paper.

The authors declare no conflict of interest.

See Commentary on page 4533.

‖To whom correspondence may be addressed. Email: dabaker@u.washington.edu or bax@nih.gov.

This article contains supporting information online at www.pnas.org/cgi/content/full/0800256105/DC1.

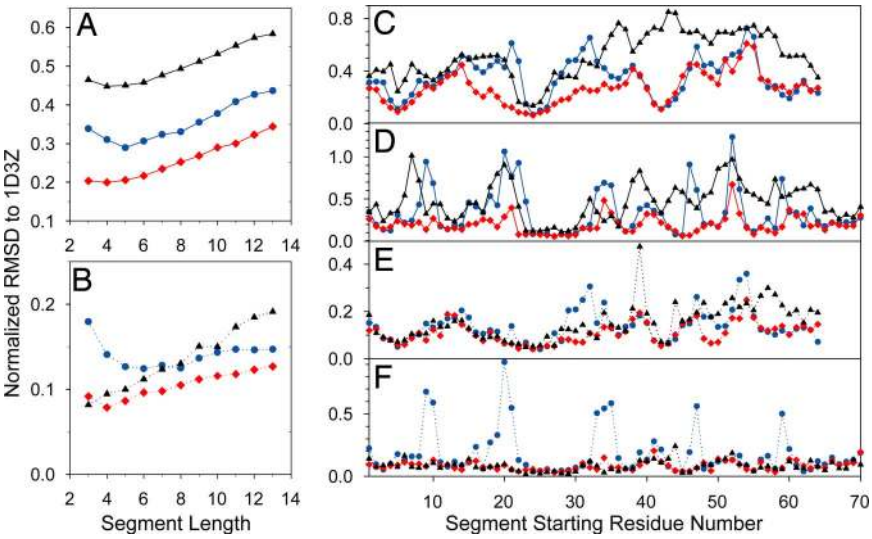

Fig. 1. Plots of normalized accuracy of database fragments selected for ubiquitin. For each ubiquitin segment, 200 fragment candidates of the same length were selected using either the standard ROSETTA procedure (filled triangles), or an MFR search of the 5665-protein structural database, assigned by the programs DC (filled circles) or SPARTA (filled diamonds). For all panels, coordinate rmsds (N, C^{α}, and C′) between query segment and selected fragments are normalized with respect to randomly selected fragments. (*A* and *B*) Average (*A*) and lowest (*B*) normalized rmsd of 200 selected fragments, as a function of fragment size, relative to the x-ray coordinates of the corresponding ubiquitin segment, averaged over all (overlapped) consecutive segments. (*C* and *D*) Average normalized rmsd of 200 nine-residue (*C*) and three-residue (*D*) fragments relative to the x-ray coordinates, as a function of position in the ubiquitin sequence. (*E* and *F*) Lowest normalized rmsd of any of these selected nine-residue (*E*) or three-residue (*F*) fragments.

sequence and hence representative of the conformations the peptide segment is likely to sample during folding. A Monte Carlo based assembly process then uses these fragments to search for compact, low energy folds. The ROSETTA full atom refinement protocol, which employs Monte Carlo minimization coupled with a detailed all-atom force field, is then used to search for low energy structures with close complementary side chain packing in the vicinity of the starting model (17). Adding the structural information contained in experimentally determined NMR chemical shifts holds promise to greatly improve the structural accuracy of selected fragments, and thereby to improve ROSETTA performance without any significant change in the basic structure or functioning of this well established program.

Here, we demonstrate the robustness of this chemical-shift-ROSETTA (CS-ROSETTA) approach for 16 proteins, whose experimental structures had previously been determined by x-ray or NMR methods. Although the structural coordinates of this test set were not used during the *de novo* structure generation process, it cannot be excluded *a priori* that optimization of the procedure itself could have developed a bias favoring this set of 16 proteins. Therefore, we also present the application of CS-ROSETTA to a set of nine proteins, under study in the Northeast Structural Genomics (NESG) consortium (www.nesg.org), for which only chemical shift assignments but no structural coordinates were available. Subsequent comparison with experimental NMR-derived coordinates confirms the close similarity between the two sets of structures, thereby independently validating the CS-ROSETTA approach.

Results

Generating new protein structures by CS-ROSETTA involves two separate stages. First, polypeptide fragments are selected from a protein structural database, based on the combined use of $^{13}C^{\alpha}$, $^{13}C^{\beta}$, $^{13}C'$, ^{15}N, $^{1}H^{\alpha}$, and $^{1}H^{N}$ chemical shifts and the amino acid sequence pattern. In the second stage, these fragments are used for *de novo* structure generation, using the standard ROSETTA protocol. Below, we first evaluate the improvement in structural accuracy of the selected fragments which results from use of a recently improved correlation between protein chemical shift and local structure, and then discuss the application of CS-ROSETTA to structure generation.

Effects of Improved Chemical Shift Prediction on Fragment Accuracy.

The recently developed SPARTA program (8) predicts chemical shifts of $^{13}C^{\alpha}$, $^{13}C^{\beta}$, $^{13}C'$, $^{1}H^{\alpha}$ and backbone ^{15}N and $^{1}H^{N}$ atoms for proteins of known structure. This program, initially trained on a set of 200 proteins for which high-resolution x-ray structures as well as complete chemical shift assignments were available, was subsequently used for "assigning" hypothetical chemical shifts to a set of 5,665 proteins for which high-resolution x-ray structures were available in the PDB. Below, we refer to this set simply as "the structural database." Importantly, SPARTA not only offers a small, ≈10% (8), improvement in accuracy for the predicted chemical shifts compared with the best alternate program, SHIFTX (18), it also derives an individual uncertainty for each chemical shift assigned to the structural database. When searching this structural database for fragments with chemical shifts similar to those of a segment in the query protein, this uncertainty is taken into account. It is also worth noting that this search aims for similarity in the so-called secondary chemical shifts, which represent the deviations of the chemical shifts from their random coil values. The correlation between local structure and secondary shift to a good approximation is independent of residue type (6), and the use of secondary shift thus removes the otherwise dominant impact of residue type when searching the structural database. A separate, weaker factor is used to favor selection of database fragments that are similar in amino acid sequence to that of the query segment.

Despite the small magnitude of the improvement in accuracy of SPARTA chemical shift prediction over alternate programs, this modest advance strongly narrows down the selection of fragments compatible with the chemical shifts of the query segment. The impact of narrowing the search on the structural accuracy of the selected fragments is evaluated for the small proteins ubiquitin and GB3 [Fig. 1 and supporting information (SI) Fig. 5], using a structural database assigned with either SPARTA or the program DC (19). As can be seen in Fig. 1, the backbone coordinate error obtained for fragments selected from the DC-assigned database is 25–65% higher than for the SPARTA-assigned database. Success or failure of the ROSETTA assembly method depends not only on the average quality of the selected fragments, but is limited also by the accuracy of the best fragments, each of which is likely to be "tried" many times during the Monte Carlo procedure. Therefore, we also compare the backbone coordinate rmsd for the best fragments selected from the database with either type of chemical shift assignment and, for reference, sequence similarity only (Fig. 1 and SI Fig. 5). Sequence similarity information alone provides less structural restraint than sequence plus chemical shift, and therefore results in a wider distribution of selected peptide conformations. This wider sampling results in a significant loss in the average quality of selected fragments (Fig. 1 *A*, *C*, and *D*), but is robust in that the best fragment of the selected ensemble is never far from the

Table 1. Accuracy of CS-ROSETTA structures for 16 proteins used during optimization

Protein name	PDB ID	N_α/N_β*	N_{res}†	N_{cs}‡	$rmsd_{bb}$§, Å	$rmsd_{all}$¶, Å
GB3	2OED	14/26	56	332	0.69	1.40
CspA	1MJC	0/33	70	405	1.57	2.19
Calbindin	4ICB	47/0	75	435	1.20	2.01
Ubiquitin	1D3Z	18/25	76	426	0.75	1.35
XcR50	1TTZ	28/16	76	352	1.53	2.30
DinI	1GHH	36/21	81	463	1.76	2.29
HPr	1POH	29/23	85	419	1.01	1.79
MrR16	1YWX	23/35	88	514	1.52	2.28
TM1112	1O5U	10/52	89	524	1.51	2.22
PHS018	2GLW	20/41	92	531	1.28	2.08
HR2106‖	2HZ5	37/25	96	470	1.65	2.42
TM1442	1SBO	41/23	110	647	1.09	1.88
Vc0424	1NXI	55/25	114	679	1.72	2.51
Spo0F	1SRR	55/25	121	590	1.24	2.02
Profilin	1PRQ	41/41	125	595	1.71	2.34
Apo_Ifabp	1LFO	15/70	129	688	1.64	2.18

Additional information in SI Table 3.

*Number of residues in α-helix and β-strand.

†Number of total residues. N- and C-terminal flexible tails are excluded from RMSD calculation (see SI Table 3).

‡Total number of CS-ROSETTA input chemical shifts.

§rmsd (C^α, C′, and N) of the lowest-energy model to the experimental structure.

¶rmsd (all non-H atoms) of the lowest-energy model to the experimental structure.

‖Protein HR2106 is a homodimer, only the monomer conformation is calculated and analyzed in this work.

reference structure, in particular for short fragments (Fig. 1 *B*, *E*, and *F*).

Evaluating CS-ROSETTA for Proteins of Known Structure. A set of 16 small proteins, for which published chemical shifts and coordinates derived from x-ray diffraction or NMR data were available, was used to optimize the CS-ROSETTA protocol and evaluate its applicability. These proteins range in size from 56 to 129 residues and include different fold types (α-, β-, and α/β-folds) and topological complexities (Table 1 and SI Table 3). For each protein, fragments were selected from our structural database using the MFR module of the NMRPipe software package, which compares the experimental chemical shifts with those of the $\approx 10^6$ fragments contained in the SPARTA-assigned database. The MFR-selected fragments were used as input for ROSETTA Monte Carlo fragment assembly and subsequent full atom refinement (see *Methods*).

For each test protein, ≈10,000–20,000 trial structures were generated by CS-ROSETTA. Using MFR-selected fragments, lower energy structures are obtained than when using standard ROSETTA fragments (SI Fig. 6). When plotting the all-atom energy, as evaluated by ROSETTA versus the difference in C^α coordinates relative to the x-ray/NMR reference structures (Fig. 2 and SI Fig. 7), for most proteins the backbone coordinates of the all-atom models with the lowest total energy deviate very little, ≈1–2 Å, from the reference structures. However, for a few proteins (e.g., histidine phosphocarrier protein, or HPr, Fig. 2*C*), models deviating *ca* 3 Å from the correct structure score more favorably than models that fall closer, ≈1 Å, to the reference structure, both in terms of total energy and in terms of the number of structures in the cluster. However, the standard ROSETTA energy scoring does not yet take into account the agreement between the structural models and the experimental chemical shifts. Each model is assembled starting from fragments that, on average, agree reasonably well with experimental chemical shifts, but because chemical shift is not used as a restraint during the ROSETTA Monte Carlo assembly process and its subsequent refinement, the chemical shifts predicted by SPARTA for these models can deviate substantially from the experimental input values. Interestingly, a significant correlation is found between the goodness of the model, and how well it agrees with the experimental chemical shifts (SI Fig. 8). As shown in Fig. 2 *A′–D′* and SI Fig. 7, after inclusion of a chemical shift term in the empirical all-atom energy function (Eq. **1** in *Methods*), the lowest energy models consistently fall close to the reference structure. Note that the use of chemical shifts to rescore the energy function only becomes beneficial when ROSETTA structures are close to the experimental structure. For models that deviate by >5 Å from the true structure, agreement with chemical shifts no longer is a useful discriminator (SI Fig. 8).

Results for the full set of 16 test proteins (Table 1 and SI Table 3) show that, in all cases, the backbone atomic coordinates of the lowest-energy predicted models are within 0.7–1.8 Å from their experimental x-ray or NMR reference structures (Fig. 3 and SI Fig. 9). Even when considering all nonhydrogen atoms of these structures, remarkably close agreement (1.4–2.5 Å) is found. The latter is likely due to the reasonably accurate description of hydrogen bonding, side chain packing, polar solvation, and backbone and side chain torsional energy by the ROSETTA all atom force field (23, 24, 27).

For proteins with incomplete backbone chemical shift assignments, such as XcR50, HR2106, Spo0F and profilin, CS-ROSETTA still yields high quality models (SI Table 3). If a given type of nucleus, for example, $^{13}C'$, is absent from the assignment table, this simply results in a slight decrease in accuracy of the fragments selected from the database, but which nevertheless

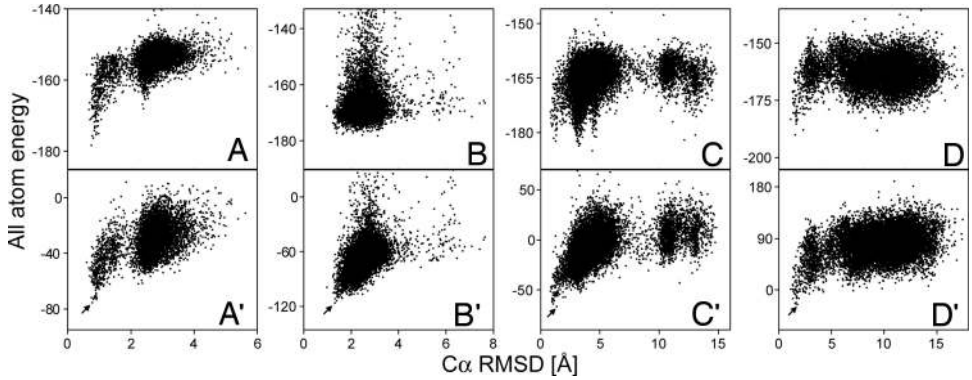

Fig. 2. Plots of ROSETTA all atom energy versus C^α rmsd relative to the experimental structures for four representative test proteins. (*A–D*) Standard ROSETTA all atom energy. (*A′–D′*) ROSETTA energy, rescored by using the experimental chemical shifts (Eq. **1**). (*A*) Ubiquitin. (*B*) Calbindin. (*C*) HPr. (*D*) TM1112. For *A′–D′*, the model with the lowest energy, marked by an arrow, is shown in Fig. 3 or SI Fig. 9.

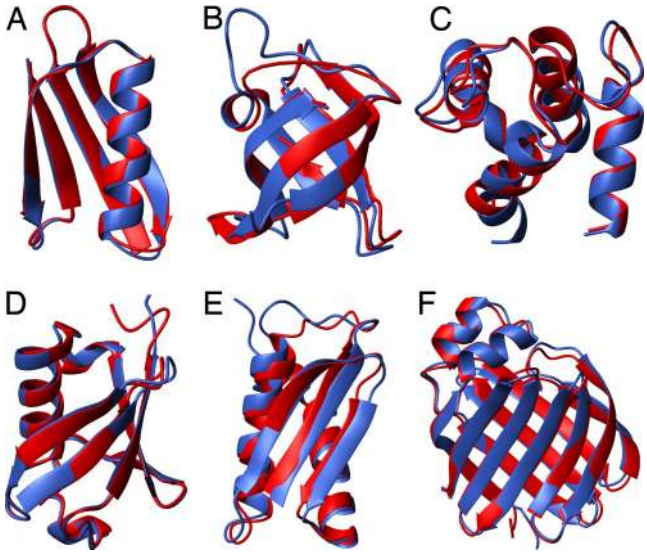

Fig. 3. Backbone ribbon representations (32) of the lowest-energy CS-ROSETTA structure (red) superimposed on the experimental x-ray/NMR structures (blue), with superposition optimized for ordered residues, as defined in the footnote to SI Table 3. (*A*) GB3. (*B*) CspA. (*C*) Calbindin. (*D*) Ubiquitin. (*E*) DinI. (*F*) Apo_lafbp. Overlays of the 10 remaining structures are shown in SI Fig. 9.

remain far superior than what could be obtained on the basis of sequence information alone. On the other hand, if for a contiguous region along the protein backbone all chemical shifts are missing, for example as a result of conformational exchange on a micro- to millisecond time scale, CS-ROSETTA effectively reverts to standard ROSETTA structure prediction for these residues. For the remainder of the protein, it takes advantage of the experimental chemical shifts.

Convergence and Acceptance of Predicted Models. The CS-ROSETTA results obtained for the 16 proteins suggest that the ROSETTA all-atom energy is funneled with respect to the C^{α} rmsd from the experimental structures, resulting in convergence to a unique structural model (Fig. 2 and SI Fig. 7). In the absence of a reference structure, a decision on whether the CS-ROSETTA structure generation process has converged instead is based on how well the coordinates of the lowest energy structures agree with one another. For this purpose, the ROSETTA all-atom energy is plotted as a function of the C^{α} rmsd relative to the model with the lowest energy. Indeed, such diagrams show the presence of at least 10 low-energy structures for each of the 16 test proteins that have a low C^{α} rmsd relative to the lowest-energy structure (SI Fig. 10).

In contrast, when inspecting in the same manner the energies of the models predicted for several larger proteins (SI Table 4), no clustering around the lowest energy structure is observed and the backbone coordinates of other low energy structures differ by more than 4–5 Å from those of the lowest-energy model (SI Fig. 11). Such large divergence in structure for the lowest energy models indicates that the structure determination process has not converged, and in these cases a reliable structure model cannot be obtained from the calculations. In practice, we find that convergence rapidly decreases with increasing protein size, and that for proteins larger than ≈130 residues the CS-ROSETTA approach starts to fail. Convergence is also adversely affected by the presence of long, disordered loops in the protein.

Blind Structure Generation for In-Progress Structural Genomics Proteins. To further evaluate the applicability of CS-ROSETTA to structural genomics and high throughput NMR, structures were generated for nine proteins, under study by investigators of the NESG consortium, for which backbone chemical shifts but no coordinates were available. Vice versa, with one exception (see footnote to SI Table 5), none of the structural information derived from CS-ROSETTA was sent to NESG until after the structures of these targets had been determined using their standard, NOE-based procedures (20, 21). The nine proteins range in molecular mass from 7.8 to 15.4 kDa and span a range of topological complexities (Table 2 and SI Table 5). For each of these proteins, the 10 structures with the lowest ROSETTA total energy, after rescoring to ensure agreement with the experimental chemical shifts, are selected for further evaluation. Comparison of these models to the experimental NMR structures indicates the results to be strikingly similar (Fig. 4, Table 2, SI Fig. 12, and SI Table 5). The two different methods identify identical folds and very similar secondary structure, with one caveat: The CS-ROSETTA structures had a tendency to slightly lengthen the elements of secondary structure and include residues that were clearly disordered as judged by the NMR data. For example, CS-ROSETTA extended the N-terminal helix of StR82 by two residues, and for 8 of the 10 lowest energy structures, it suggested the presence of a short helical segment within a disordered extended loop region. The ROSETTA all-atom energy function favors the formation of intramolecular hydrogen bonds, and a tendency to generate secondary structure for disordered regions is therefore not surprising. This caveat was subsequently addressed by using the recently introduced ′random coil index′, or RCI (22), which positively identifies regions of disordered structure on the basis of near-random-coil chemical shifts. These residues are now flagged in CS-ROSETTA, such that their hydrogen bonding no longer contributes to the ROSETTA energy term.

When comparing the two sets of structures, and focusing on the ordered regions, remarkably good agreement is seen with backbone coordinate rmsd values <1 Å for six of the nine proteins, the other three differing by less than ≈2 Å. Standard structure validation parameters, such as the G-factor obtained with the program PROCHECK (23), or the MolProbity (24) clash and Ramachandran plot scores generally rank the CS-ROSETTA structures comparable or higher in quality than the experimental NMR structures (SI Table 5). DP scores, which compare short distances in the structure with the experimental NOESY peak list (25), are also reasonably good for most of the CS-ROSETTA structures (SI Table 5). However, these DP scores are 10–40% better for the

Table 2. Statistics on CS-ROSETTA structures of nine structural genomics proteins

Protein name	N_{res}*	PDB ID	$rmsd_{bb}$†, Å	$rmsd_{all}$‡, Å	DP§, %
RpT7	65	2JTV	0.64	1.29	69
StR82	69	2JT1	0.57	1.14	65
RhR95	72	2JVM	0.66	1.18	55
NeT4	73	2JV8	0.70	1.42	57
TR80	78	2JXT	0.69	1.27	67
VfR117	80	2JVW	0.60	1.40	37
PsR211	100	2JVA	2.07	2.34	57
AtR23	101	2JYA	1.10	1.81	60
NeR45A	147	2JXN	2.03	2.85	53

Additional information in SI Table 5.

*Number of total residues.

†rmsd (C^{α}, C′, and N) of the mean coordinates of 10 lowest-energy models to the mean coordinates of the experimental NMR structure. Residues in disordered regions (see SI Table 5) are excluded from rmsd calculation.

‡rmsd (all non-H atoms) of the mean coordinates of 10 lowest-energy models to the mean coordinates of the experimental structure.

§DP scores measure the agreement between the structure and the NOESY peak list, as defined in ref. 25.

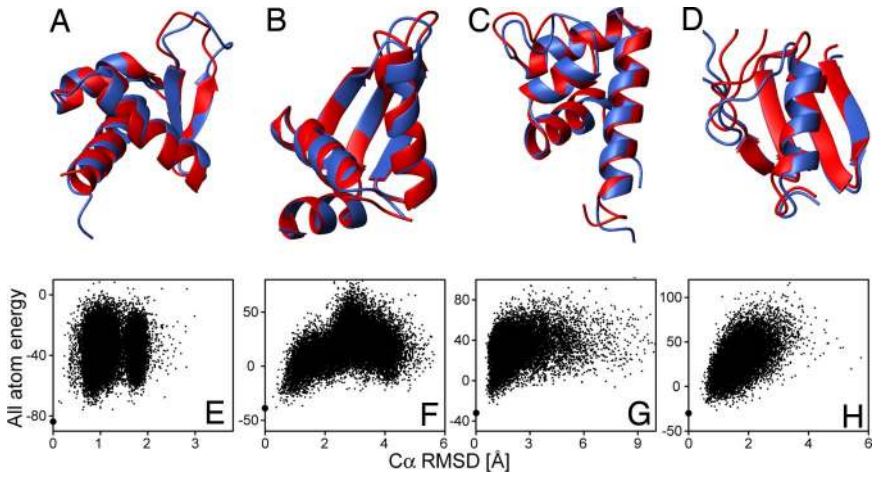

Fig. 4. Results from blind CS-ROSETTA structure generation for four structural genomics targets (Table 2). The remaining five are in SI Fig. 12. (*A–D*) Superposition of lowest-energy CS-ROSETTA models (red) with experimental NMR structures (blue), with superposition optimized for ordered residues, as defined in the footnote to SI Table 5. (*E–H*) Plots of rescored (Eq. **1**) ROSETTA all-atom energy versus C^{α} rmsd relative to the lowest-energy model (bold dot on vertical axis). (*A* and *E*) StR82. (*B* and *F*) RpT7. (*C* and *G*) VfR117. (*D* and *H*) NeT4.

experimentally determined NMR structures, refined against these same experimental NOESY data, than for the predicted structures (SI Table 5). Differences in DP scores are found to result mostly from small differences in core side chain packing.

Discussion

It has long been recognized that chemical shifts are strongly influenced by local conformation. The CHESHIRE method, recently introduced by Vendruscolo and coworkers (13), exploits this relation and became the first program to generate near-atomic resolution structures from chemical shifts. The CS-ROSETTA method described in the present study is based on the same concept, but combines the well established ROSETTA structure prediction program (16, 17) with a recently enhanced empirical relation between structure and chemical shift (8), which allows selection of database fragments that better match the structure of the unknown protein. The method was calibrated using 16 proteins of known structure, and then successfully tested for nine proteins under study in the NESG structural genomics program. The improvement in fragment selection and the incorporation of high resolution refinement are responsible for the significant increase in model quality compared with an earlier integration of chemical shift data with ROSETTA (12).

All 16 proteins used to initially develop the CS-ROSETTA program are of relatively simple topology. Upon completion of this work, we became aware of an NMR structure recently solved for corona virus protein *nsp1* (26) (116 residues excluding disordered tails) which exhibits a novel, highly unusual fold that requires a complex "folding from the center" pathway. The standard ROSETTA structure assembly protocol has difficulty generating such nearly knot-forming proteins, and it is therefore not surprising that CS-ROSETTA was unable to obtain a converged low energy fold. Failure to reach the convergence threshold (see *Methods*) for this protein provided a clear indication that no structural conclusions could be drawn from the CS-ROSETTA results.

CS-ROSETTA was successfully applied, in a blind manner, to determine the structures of nine in progress structural genomics targets with sizes in the 65–129 residue range, yielding structural models that are highly consistent with their independently solved experimental NMR structures. These structural genomics target proteins tend to include larger unstructured loop regions than the globular proteins for which CS-ROSETTA originally was optimized.

Successful application of CS-ROSETTA so far remains limited to relatively small proteins, not larger than ≈15 kDa. Although this size threshold is substantially higher than for conventional ROSETTA structure prediction, it remains well below the 25–30 kDa size limit of protein structures that can be studied in a relatively standard manner by triple resonance NMR spectroscopy. A number of variations and extensions of CS-ROSETTA are currently being explored to extend its limit to this larger size range, where conventional NMR structure determination can become very time-intensive. Preliminary results indicate that knowledge regarding a very small set of long range H^N–H^N or H^N–CH_3 NOE interactions, which typically can easily be extracted from a 3D NOE data set on a perdeuterated protein sample with protonated amide and/or methyl protons, provides a considerable boost in this direction, and such experimental information can readily be added to the standard ROSETTA structure generation algorithm (12, 17).

ROSETTA structure determination requires large amounts of computer time for generating a sufficient number (10,000–20,000) of all-atom models, needed to ensure "convergence" to the lowest energy model. Our study used the Berkeley Open Infrastructure for Network Computing (BOINC) for this computationally demanding work, taking advantage of idle time on thousands of personal computers world-wide. However, computations for a single protein can also be carried out locally, requiring ≈1 day on a cluster of 100–200 CPUs, and in favorable cases a smaller number of models suffices to reach convergence (SI Fig. 13).

Compared with conventional NMR protein structure determination, CS-ROSETTA offers considerable time savings, both in terms of measurement time and in terms of spectral analysis. CS-ROSETTA uses only backbone and $^{13}C^{\beta}$ chemical shifts, thereby obviating the need for side chain assignments as well as the collection and interpretation of NOE data. Although the time required for these last two steps varies greatly depending on the protein, the quality of the data, and the expertise of the experimentalist, they always take considerably longer than the time needed for backbone assignments. Thus, we estimate that CS-ROSETTA yields a time savings of at least 50%.

Perhaps even more important than the potential acceleration of the NMR structure determination process is CS-ROSETTA's applicability to the study of systems not amenable by conventional NMR (13). These include structures of short-lived, unstable proteins, or systems where a minor state is in dynamic equilibrium with a more populated state, in which case collection of structural restraints for the minor component can be prohibitively difficult.

Methods

Structural Database and Fragment Searching. Details regarding the construction of the structural database, the selection of fragments, and identification of flexible regions are provided as SI.

Generating Protein Structures from Fragments using ROSETTA. The regular ROSETTA Monte Carlo fragment assembly method (16, 27, 28) is used in this work to generate full atom models, including the steps used for generation of a low-resolution backbone conformation, as well as the final refinement stage where a full-atom model is generated. During generation of the backbone fold, ROSETTA adopts a simple representation of the protein chain in which only the backbone heavy atoms and the ''centroid'' of side chains are explicitly considered. Starting from a fully extended chain, these backbone folds are generated by 360,000 steps of Monte Carlo fragment replacement while minimizing an empirical energy term that primarily includes van der Waals packing, hydrogen bonding and desolvation terms, and where the ϕ, ψ, and ω backbone torsion angles of a randomly selected three- or nine-residue fragment of the protein chain are replaced with the torsion angles from one of the 200 corresponding, randomly selected MFR candidates. The move is accepted according to the score calculated for the new conformation using the Metropolis criterion (29). In the subsequent high-resolution all-atom model generation, side chain conformations are first added to the low-resolution backbone model using a Monte Carlo simulated annealing search (30) through a backbone-dependent rotamer library (31). This all atom model is then refined using a Monte Carlo minimization protocol in which each attempted move consists of (*i*) a random perturbation of one or more backbone torsion angle, (*ii*) reoptimization of the side chain rotamer conformations and (*iii*) gradient-based optimization of the backbone and side chain torsion angles. These compound moves are accepted or rejected based on the Metropolis criterion; the number of attempted moves is equal to twice the number of residues in the protein. The all-atom energy function includes a Lennard-Jones potential, an orientation dependent hydrogen bonding potential, an implicit solvation model, and a knowledge-based side chain and backbone torsional potential. Details on the energy function and methods are described in (16, 17). For each protein, 10,000–20,000 all atom models are generated, and for the 16 test proteins, Fig. 2 and SI Fig. 7 show the energies plotted against the C^{α} rmsd relative to the known test protein structure.

All CS-ROSETTA protein models were generated using ROSETTA@home (http://boinc.bakerlab.org/rosetta/) supported by the BOINC server.

Selection of All-Atom Models Using Energies and Chemical Shifts. The ROSETTA all-atom models resulting from the above procedure were evaluated further in terms of the fitness with respect to the experimental chemical shifts. For each all-atom model, the backbone $^{13}C^{\alpha}$, $^{13}C^{\beta}$, $^{13}C'$, ^{15}N, $^{1}H^{\alpha}$, and $^{1}H^{N}$ chemical shifts were predicted using the SPARTA program. The outcome of this was used to adjust the ROSETTA full atom energy according to:

$$E' = E + c \times \chi^2_{cs}, \qquad [1a]$$

where

$$\chi^2_{cs} = \sum_i \sum_j (\delta^{exp}_{i,j} - \delta^{pred}_{i,j})^2 / \sigma^2_{i,j}, \qquad [1b]$$

where $\delta^{pred}_{i,j}$ refers to the SPARTA-predicted backbone chemical shift ($i = {}^{13}C^{\alpha}$, $^{13}C^{\beta}$, $^{13}C'$, ^{15}N, $^{1}H^{\alpha}$, and $^{1}H^{N}$) from the all-atom model for a given residue j, $\delta^{exp}_{i,j}$ is the experimental chemical shift, $\sigma_{i,j}$ is the uncertainty of $\delta^{pred}_{i,j}$, and c is a weighting factor set to 0.25.

Extraction of the Best Models. For each protein, a total of 10,000–20,000 all-atom models were generated from the MFR-selected fragments by the ROSETTA assembly and relaxation protocol. The 5,000 lowest-energy models were taken and their all-atom energies were adjusted according to Eq 1. A plot of the ROSETTA all-atom energy against the C^{α} rmsd relative to the lowest-energy model (Fig. 4 and SI Fig. 12) was then used to evaluate convergence, and to select the 10 lowest-energy models.

Criteria for Convergence and Accepting Models. For all predicted models of each protein the ROSETTA all-atom energy, rescored by Eq. **1**, is plotted against its C^{α} rmsd from the lowest-energy model. If the low energy models cluster within less than ≈2 Å from the model with the lowest energy, the structure prediction is deemed successful and the 10 lowest energy models are accepted.

Software. CS-ROSETTA, which includes SPARTA, MFR scripts, a complete example for GB3, and the structural database used in this work, can be freely downloaded from http://spin.niddk.nih.gov/bax/software/CSROSETTA/index.html. ROSETTA can be downloaded from www.rosettacommons.org/software.

ACKNOWLEDGMENTS. This work was funded by the Intramural Research Program of the National Institute of Diabetes and Digestive and Kidney Diseases, National Institutes of Health (NIH); Intramural AIDS-Targeted Antiviral Program of the Office of the Director, NIH; the National Institute of General Medical Sciences (NIGMS), NIH, The Human Frontier Science Program (to G.L.), and the Howard Hughes Medical Institutes (to D.B.); the Ontario Research and Development Challenge Fund, and the Genome Canada and Canada Research Chairs Programs. Northeast Structural Genomics Consortium is supported by the NIGMS Protein Structure Initiative, grant U54-GM074958. We also thank ROSETTA@home participants and the BOINC project for contributing computing power.

1. Atreya HS, Szyperski T (2005) Rapid NMR data collection. *Methods Enzymol* 394:78–108.
2. Freeman R, Kupce E (2003) New methods for fast multidimensional NMR. *J Biomol NMR* 27:101–113.
3. Wagner G, Pardi A, Wüthrich K (1983) Hydrogen-bond length and H-1-NMR chemical-shifts in proteins. *J Am Chem Soc* 105:5948–5949.
4. Williamson MP, Asakura T (1993) Empirical comparisons of models for chemical-shift calculation in proteins. *J Magn Reson B* 101:63–71.
5. Case DA (1995) Calibration of ring-current effects in proteins and nucleic acids. *J Biomol NMR* 6:341–346.
6. Cornilescu G, Delaglio F, Bax A (1999) Protein backbone angle restraints from searching a database for chemical shift and sequence homology. *J Biomol NMR* 13:289–302.
7. Xu XP, Case DA (2001) Automated prediction of N-15, C-13(alpha), C-13(beta) and C-13′ chemical shifts in proteins using a density functional database. *J Biomol NMR* 21:321–333.
8. Shen Y, Bax A (2007) Protein backbone chemical shifts predicted from searching a database for torsion angle and sequence homology. *J Biomol NMR* 38:289–302.
9. Tolman JR, Flanagan JM, Kennedy MA, Prestegard JH (1995) Nuclear magnetic dipole interactions in field-oriented proteins: Information for structure determination in solution. *Proc Natl Acad Sci USA* 92:9279–9283.
10. Tjandra N, Bax A (1997) Direct measurement of distances and angles in biomolecules by NMR in a dilute liquid crystalline medium. *Science* 278:1111–1114.
11. Delaglio F, Kontaxis G, Bax A (2000) Protein structure determination using molecular fragment replacement and NMR dipolar couplings. *J Am Chem Soc* 122:2142–2143.
12. Bowers PM, Strauss CEM, Baker D (2000) De novo protein structure determination using sparse NMR data. *J Biomol NMR* 18:311–318.
13. Cavalli A, Salvatella X, Dobson CM, Vendruscolo M (2007) Protein structure determination from NMR chemical shifts. *Proc Natl Acad Sci USA* 104:9615–9620.
14. Gong HP, Shen Y, Rose GD (2007) Building native protein conformation from NMR backbone chemical shifts using Monte Carlo fragment assembly. *Protein Sci* 16:1515–1521.
15. Anfinsen CB, Haber E, Sela M, White FH (1961) Kinetics of formation of native ribonuclease during oxidation of reduced polypeptide chain. *Proc Natl Acad Sci USA* 47:1309-1314.
16. Bradley P, Misura KMS, Baker D (2005) Toward high-resolution de novo structure prediction for small proteins. *Science* 309:1868–1871.
17. Rohl CA, Strauss CEM, Misura KMS, Baker D (2004) Protein structure prediction using rosetta. *Methods Enzymol* 383:66–93.
18. Neal S, Nip AM, Zhang HY, Wishart DS (2003) Rapid and accurate calculation of protein H-1, C-13 and N-15 chemical shifts. *J Biomol NMR* 26:215–240.
19. Kontaxis G, Delaglio F, Bax A (2005) Molecular fragment replacement approach to protein structure determination by chemical shift and dipolar homology database mining. *Methods Enzymol* 394:42–78.
20. Liu GH, *et al.* (2005) NMR data collection and analysis protocol for high-throughput protein structure determination. *Proc Natl Acad Sci USA* 102:10487–10492.
21. Grishaev A, *et al.* (2005) ABACUS, a direct method for protein NMR structure computation via assembly of fragments. *Proteins* 61:36–43.
22. Berjanskii M, Wishart DS (2006) NMR: Prediction of protein flexibility. *Nature Protocols* 1:683–688.
23. Laskowski RA, Rullmann JAC, MacArthur MW, Kaptein R, Thornton JM (1996) AQUA and Procheck NMR: Programs for checking the quality of protein structures solved by NMR. *J Biomol NMR* 8:477–486.
24. Word JM, Bateman RC, Presley BK, Lovell SC, Richardson DC (2000) Exploring steric constraints on protein mutations using MAGE/PROBE. *Protein Sci* 9:2251–2259.
25. 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.
26. Almeida MS, Johnson MA, Herrmann T, Geralt M, Wüthrich K (2007) Novel beta-barrel fold in the nuclear magnetic resonance structure of the replicase nonstructural protein 1 from the severe acute respiratory syndrome coronavirus. *J Virol* 81:3151–3161.
27. Tsai J, *et al.* (2003) An improved protein decoy set for testing energy functions for protein structure prediction. *Proteins* 53:76–87.
28. Misura KMS, Baker D (2005) Progress and challenges in high-resolution refinement of protein structure models. *Proteins* 59:15–29.
29. Metropolis N, Rosenbluth AW, Rosenbluth MN, Teller AH, Teller E (1953) Equation of state calculations by fast computing machines. *J Chem Phys* 21:1087–1092.
30. Kuhlman B, Baker D (2000) Native protein sequences are close to optimal for their structures. *Proc Natl Acad Sci USA* 97:10383–10388.
31. Dunbrack RL, Cohen FE (1997) Bayesian statistical analysis of protein side-chain rotamer preferences. *Protein Sci* 6:1661–1681.
32. Koradi R, Billeter M, Wüthrich K (1996) MOLMOL: A program for display and analysis of macromolecular structures. *J Mol Graphics* 14:51–55.