



# Determination of solution structures of proteins up to 40 kDa using CS-Rosetta with sparse NMR data from deuterated samples

Oliver F. Lange<sup>a,b,1,2</sup>, Paolo Rossi<sup>c,1</sup>, Nikolaos G. Sgourakis<sup>d</sup>, Yifan Song<sup>d</sup>, Hsiao-Wei Lee<sup>e</sup>, James M. Aramini<sup>c</sup>, Asli Ertekin<sup>c</sup>, Rong Xiao<sup>c</sup>, Thomas B. Acton<sup>c</sup>, Gaetano T. Montelione<sup>c,f</sup>, and David Baker<sup>d,g</sup>

<sup>a</sup>Biomolecular NMR and Munich Center for Integrated Protein Science, Department Chemie, Technische Universität München, 85747 Garching, Germany; <sup>b</sup>Institute of Structural Biology, Helmholtz Zentrum München, 85764 Neuherberg, Germany; <sup>c</sup>Center for Advanced Biotechnology and Medicine, Department of Molecular Biology and Biochemistry, and Northeast Structural Genomics Consortium Rutgers, State University of New Jersey, Piscataway, NJ 08854; <sup>d</sup>Department of Biochemistry, University of Washington, Seattle, WA 98195; <sup>e</sup>Complex Carbohydrate Research Center, University of Georgia, Athens, GA 30602; <sup>f</sup>Department of Biochemistry, Robert Wood Johnson Medical School, University of Medicine and Dentistry of New Jersey, Piscataway, NJ 08854; and <sup>g</sup>Howard Hughes Medical Institute, University of Washington, Seattle, WA 98195

Edited by Adriaan Bax, National Institutes of Health, Bethesda, MD, and approved May 22, 2012 (received for review March 1, 2012)

**We have developed an approach for determining NMR structures of proteins over 20 kDa that utilizes sparse distance restraints obtained using transverse relaxation optimized spectroscopy experiments on perdeuterated samples to guide RASREC Rosetta NMR structure calculations. The method was tested on 11 proteins ranging from 15 to 40 kDa, seven of which were previously unsolved. The RASREC Rosetta models were in good agreement with models obtained using traditional NMR methods with larger restraint sets. In five cases X-ray structures were determined or were available, allowing comparison of the accuracy of the Rosetta models and conventional NMR models. In all five cases, the Rosetta models were more similar to the X-ray structures over both the backbone and side-chain conformations than the “best effort” structures determined by conventional methods. The incorporation of sparse distance restraints into RASREC Rosetta allows routine determination of high-quality solution NMR structures for proteins up to 40 kDa, and should be broadly useful in structural biology.**

nuclear magnetic resonance | sparse data | maltose binding protein | structural genomics | genetic algorithms

Advances in hardware, sample preparation, pulse sequence development, and refinement techniques have expanded the size and complexity of proteins accessible to structure determination by solution-state NMR to include proteins that, until recently, were exclusively the realm of X-ray crystallography (1–3). However, despite a number of landmark studies (4–7), only a small percentage of structures solved by NMR and deposited in the Protein Data Bank exceed 20 kDa in molecular weight. Larger structures need to be assembled by combining structural information from individual domains, and require additional techniques to elucidate the spatial arrangement, such as shape fitting (5) and/or paramagnetic restraints (8).

The 20-kDa general limit coincides with the two fundamental problems in solution-state NMR: resonance overlap and progressive increase in the transverse relaxation rate ( $1/T_2$ ). As the size of a molecule increases, so does the rotational correlation time and, consequently, the efficiency of  $^1\text{H}$ – $^1\text{H}$  relaxation mechanisms. One way to suppress these effects is to incorporate deuterium into the protein sample, diluting the  $^1\text{H}$ – $^1\text{H}$  relaxation networks and increasing  $^{13}\text{C}$  and  $^{15}\text{N}$  relaxation times, resulting in sharper line widths and dramatic improvement of the signal-to-noise ratios (2, 9, 10). Perdeuteration is generally required for studies of larger proteins (11–14), particularly membrane proteins (15, 16).

Unfortunately, deuteration also eliminates the majority of  $^1\text{H}$ – $^1\text{H}$  NOEs, the main source of long-range distance information in solution-state NMR. Several methods have emerged for reintroducing protons at selected sites to function as distance

probes in the structure (11, 17). Methyl groups of isoleucine  $\delta 1$ , leucine, and valine side chains are straightforward to label with  $^{13}\text{C}$  and  $^1\text{H}$  isotopes in an otherwise deuterated protein sample (12, 13). As methyl groups are often found in the core of proteins, “ile-leu-val (ILV) labeling” combined with back-exchange of backbone and side-chain amide protons allows identification of extensive networks of  $\text{CH}_3$ – $\text{CH}_3$  and  $\text{CH}_3$ – $\text{H}^{\text{N}}$ , as well as  $\text{H}^{\text{N}}$ – $\text{H}^{\text{N}}$  restraints. However, while such an ILV-labeling strategy has provided correct fold determination for proteins of up to approximately 80 kDa (4, 7), the overall sparseness of these long-range restraints limits the accuracy of structural details.

Recently, we showed that the iterative RASREC CS-Rosetta methodology (integrating sparse NMR data, a detailed all-atom energy function, and advanced sampling techniques) (18) has considerable promise for the determination of medium- and larger-sized protein structures (19). We were able to determine structures for proteins up to 25 kDa using only backbone amide–amide ( $\text{H}^{\text{N}}$ – $\text{H}^{\text{N}}$ ) NOEs, residual dipolar couplings, and chemical shifts. Nevertheless, in some cases the  $\text{H}^{\text{N}}$ – $\text{H}^{\text{N}}$  backbone-only approach is not sufficiently robust. In particular, the placement of helices is difficult, because backbone  $\text{H}^{\text{N}}$ – $\text{H}^{\text{N}}$  NOEs generally do not yield tertiary structure restraints in helical regions.

Here, we demonstrate that high-quality 3D structures of proteins in the 20–40 kDa range can be routinely determined within the CS-Rosetta framework using a relatively small number of sparse NOE restraints obtained using deuterated ILV–methyl protonated samples. The strategy leverages methyl–methyl ( $\text{CH}_3$ – $\text{CH}_3$ ), methyl–amide ( $\text{CH}_3$ – $\text{H}^{\text{N}}$ ), and amide–amide ( $\text{H}^{\text{N}}$ – $\text{H}^{\text{N}}$ ) NOE contacts in conjunction with backbone chemical shift (CS) and residual dipolar coupling (RDC) data to determine protein structures using the RASREC CS-Rosetta protocol (18). In most cases, high-quality structures were obtained from these datasets using a semiautomated NOE cross-peak assignment procedure requiring minimal manual assignment efforts. In cases where

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

The authors declare no conflict of interest.

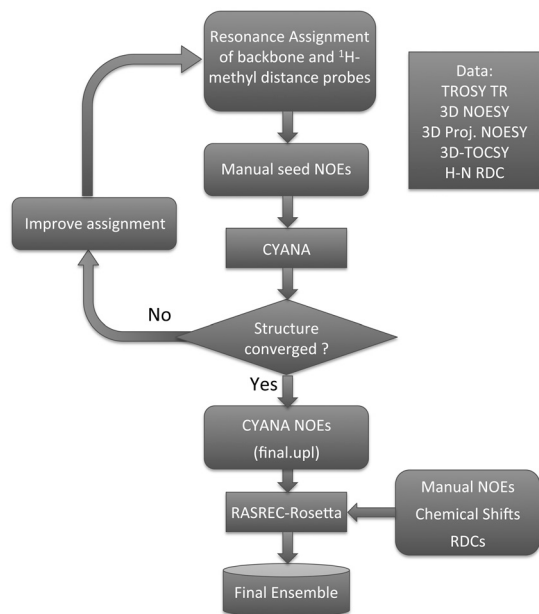
This article is a PNAS Direct Submission.

Data deposition: The atomic coordinates were deposited in the Protein Data Bank, [www.pdb.org](http://www.pdb.org) (PDB ID codes 2kw5, 2loy, 2lmd, 2kzn, 2lnu, and 2lok) and NESG target codes: SgR145, WR73, HR4660B, SR10, HmR11, and HsR50; the NMR chemical shifts were deposited in the BioMagResBank, [www.bmrb.wisc.edu](http://www.bmrb.wisc.edu) (accession nos. 16806, 16833, 18112, 17008, 18180, and 18215). The software is available at [www.csrosetta.org](http://www.csrosetta.org).

<sup>1</sup>O.F.L. and P.R. contributed equally to this work.

<sup>2</sup>To whom correspondence should be addressed. E-mail: [oliver.lange@tum.de](mailto:oliver.lange@tum.de).

This article contains supporting information online at [www.pnas.org/lookup/suppl/doi:10.1073/pnas.1203013109/-DCSupplemental](http://www.pnas.org/lookup/suppl/doi:10.1073/pnas.1203013109/-DCSupplemental).



**Fig. 1.** Flow chart illustrating the new NMR structure determination protocol. Expert NMR data analysis is shown as rounded rectangles, whereas structural modeling based on interpreted data (such as chemical shifts, peak lists, and RDCs) is in boxed shapes.

both X-ray and conventional NMR structures are available, the RASREC CS-Rosetta structures appear to be more accurate than the conventionally determined NMR structures.

## Results

The RASREC CS-Rosetta approach utilizes NMR data on uniformly deuterated ILV-methyl protonated proteins as outlined in Fig. 1. A small number of manually assigned seed NOE contacts (see *SI Appendix, Fig. S1 and Table S1*) are used as input to CYANA (20) or one of the other available automated NOESY assignment programs, such as AutoStructure (21), ARIA (22), or UNIO (23). The resulting, automatically assigned NOE distance restraints, along with backbone chemical shifts, RDC data, and the manually assigned seed restraints, are subsequently used within the RASREC CS-Rosetta protocol (18).

Rather than simply using synthetic/simulated data, we have rigorously developed and tested our approach on real protein NMR data, most of which was recorded specifically for this project, going from sample to structure for nine protein targets ranging in size from 18 to 40 kDa, including seven “blind” targets of previously unknown structure. Details of the experimental NMR data collection and analysis strategy are described in *Methods*.

NMR data were collected specifically for this study on seven protein targets of the Northeast Structural Genomics Consortium (NESG), designated by NESG IDs (24): SR10, SgR145, WR73, HsR50, HmR11, HR4660B, and ER690 [maltose binding protein (MBP) complexed to  $\beta$ -cyclodextrin]. For two targets (the N- and C-terminal domains of BamC), collaborators shared their data (25), and for two additional targets [sensory rhodopsin (26) and MBP (4)], the data were previously published. The latter two restraint sets were filtered to include only those restraints that would be obtained using ILV-methyl protonated samples.

## Comparison with “Best-Effort” Conventional Structure Determination Methods.

Four of the targets (SR10, SgR145, rhodopsin, and MBP) have both a conventionally determined solution NMR structure and an X-ray crystal structure. These structures allowed us to compare the results of the RASREC CS-Rosetta method with (i) conventional automated NMR structure analysis using CYANA (27) and (ii) manual best-effort NMR analysis, which

involves an iterative combination of manual and automated analysis. In this study, we assume that the X-ray structure is an accurate representation of the dominant solution structure; accordingly, the rmsd of atomic coordinates between NMR and X-ray structure provides a measure of the accuracy of the NMR structure. This view is supported by the NMR data (*SI Appendix, Table S2*). Based on these criteria, RASREC CS-Rosetta consistently outperforms conventional automated NMR analysis using the CYANA program and the same NMR data (Table 1). The resulting RASREC CS-Rosetta structures typically have mean backbone  $C_{\alpha}$  rmsd relative to the corresponding X-ray crystal structure of  $<2.0$  Å. Remarkably, although RASREC CS-Rosetta is largely automated, it also generally outperforms the published best-effort manually refined NMR structures (Table 1).

MBP (370 residues,  $\tau_c$  approximately 18 ns at 37°C) is a two-domain protein that dynamically samples open and closed conformations in the absence of ligand (28). Although high-quality NMR data for MBP can only be obtained when complexed to  $\beta$ -cyclodextrin, no ligand protein contacts were employed for structure calculation. The low-energy RASREC CS-Rosetta structures sample the full conformational range and opening angles (Fig. 2*A* and *B*), which explains the relatively high rmsd to the crystal structure observed for the full structure [3.1 Å  $C_{\alpha}$ -rmsd using manually refined restraints (*SI Appendix, Fig. S2*) and 5.0 Å using automatically assigned restraints (Table 1)]. The  $C_{\alpha}$ -rmsd to the crystal structure for individual domains are lower: 3.0 Å and 1.9 Å, respectively, using the published restraints (4), and 2.0 Å and 4.1 Å using automatically assigned restraints obtained from data on MBP collected specifically for this study (Table 1 and Fig. 2*C* and *D*). Remarkably, using the published restraints (4), the RASREC CS-Rosetta structures for each domain are closer to the reference crystal structure (1ez9) (29) (3.0 Å and 1.9 Å for N- and C-terminal domains, respectively) than the best-effort manually refined structure determined using an even larger set of RDC restraints (3.1 Å and 3.0 Å, respectively) including five RDC vectors (instead of just N-H) and CSA restraints, as well as hydrogen-bond and backbone dihedral angle restraints. RASREC CS-Rosetta also converges using just the expert-derived backbone  $H^N$ - $H^N$  NOEs and a single set of  $^1H$ - $^{15}N$  RDCs (*SI Appendix, Table S3*).

Sensory rhodopsin (225 residues) is a membrane protein for which NOESY restraints have been generated using an ILVAMT (ILV-Ala, Met, and Thr)-labeled, deuterated sample (26). The information content of this restraint set is lower than expected: Only 13% of proton-proton contacts ( $<8$  Å) are represented in the assigned cross-peaks after expert manual analysis, compared to an average of approximately 20% for seeded automatic assignments and approximately 50% for expert assignment for the other targets (*SI Appendix, Table S7*). The low information content is probably caused by the slow molecular tumbling of the approximately 70 kDa protein-detergent complex (24 ns at 50°C) (Table 1) and the presence of intense residual detergent signals that made manual analysis of the spectra challenging (26). Using this suboptimal restraint set, conventional methods were not able to obtain well-packed structures, as documented in *SI Appendix, Fig. S8B* of ref. 26. Conversely, RASREC-Rosetta yields well-packed structures demonstrating that Rosetta is more robust than conventional methods when facing problematic data. The ten lowest-energy structures obtained with RASREC CS-Rosetta using an implicit membrane model (*SI Appendix: Methods*) superimpose with 1.7 Å  $C_{\alpha}$ -rmsd to the X-ray reference structure (Fig. 3*A*). RASREC calculations using only the NOEs that would be obtained using an ILV-labeled sample yielded models with 1.8 Å  $C_{\alpha}$ -rmsd to the crystal structure (1h68) (30). Furthermore, the RASREC CS-Rosetta structure that was solved with 215 long-range ( $|i - j| \geq 4$ ) NOESY restraints from ILVAMT is essentially equivalent in accuracy to the final deposited NMR structure that used 1,536 long-range ( $|i - j| \geq 5$ ), 1,131 medium-range

**Table 1. Iterative CS-Rosetta-based approach yields more accurate structures than conventional methods for ILV-labeled protein samples**

Target	No. of residues	MW(kDa)	$\tau_c$ *(ns)	Reference X-ray crystal structure PDB_id	Residue ranges used for rmsd analysis	Low/median/high (Å) <sup>†</sup> rmsd to X-ray structure		
						RASREC CS-Rosetta	Conventional NMR	
							Automated analysis using CYANA	Deposited coordinates
SR10	141	18	9	3e0o	13–25, 36–105, 111–141 <sup>‡</sup>	1.1/1.5/2.0 <sup>§</sup>	2.7/3.1/3.8	2.4/2.9/3.6
SgR145	177	22	12	3mer	21–170, 188–196 <sup>¶</sup>	1.2/1.4/1.6 <sup>§</sup>	3.7/4.7/6.2	2.4/2.6/3.6
Rhodopsin	225	26	24	1h68	4–210	1.4/1.7/2.7 <sup>  </sup>	-	1.5/1.6/1.7 <sup>**</sup>
MBP <sup>††</sup>	FULL	370	41		1–370	4.1/5.0/5.7	7.8/12.3/17.2	-
	NTD	182(370)		1ez9	1–111, 260–327	1.8/2.0/2.5	1.9/2.7/3.2	-
	CTD	178(370)			113–258, 335–370	2.8/4.1/4.7	8.2/9.0/11.8	-
MBP <sup>††</sup>	FULL	370	41		1–370	2.4/3.1/3.2	-	3.4/3.6/3.8 <sup>§§</sup>
	NTD	182(370)		1ez9	1–111, 260–327	2.6/3.0/3.3	-	2.7/3.1/3.5 <sup>§§</sup>
	CTD	178(370)			113–258, 335–370	1.2/1.9/2.1	-	2.8/3.0/3.3 <sup>§§</sup>

\*Rotational correlation times ( $\tau_c$ ) in ns were experimentally determined from <sup>15</sup>N T<sub>1</sub> and T<sub>2</sub> (CPMG) measurements (44) conducted at 800 MHz and 298 K (310 K for MBP) or as given in ref. 26 for sensory rhodopsin.

<sup>†</sup>Backbone C $\alpha$ -rmsd to the X-ray crystal structure were calculated for the 10 lowest-energy models for RASREC CS-Rosetta, or for all structures in the deposited NMR ensembles. Displayed are the lowest, the median, and the highest C $\alpha$ -rmsd.

<sup>‡</sup>Loop residues for which chemical shifts were missing, or where TALOS+ predicts high flexibility, were excluded from analysis.

<sup>§</sup>These results were obtained with the current release version of the protocol, available (as version 1.0) at: <http://www.csrosetta.org>. The main improvement stems from a different scheme to map methyl restraints onto the low-resolution protein model, as described in *SI Appendix: Methods*. The original protocol (used for results without this footnote and results in Table 2) resulted in median C $\alpha$ -rmsd of 2.0 Å and 1.9 Å for targets SR10 and SgR145, respectively.

<sup>¶</sup>Excluded flexible residues; missing electron density in X-ray data at residues 1–20 and 155–164.

<sup>||</sup>ILVAMT-labeled sample.

<sup>\*\*</sup>Deposited NMR structure (Protein Data Bank ID: 2ksy) based on double labeled sample was obtained using considerably more restraints; 1,536 long-range ( $|i-j| \geq 5$ ) for the conventional calculation vs. 185 long-range restraints for the RASREC Rosetta calculation based on the ILVAMT sample.

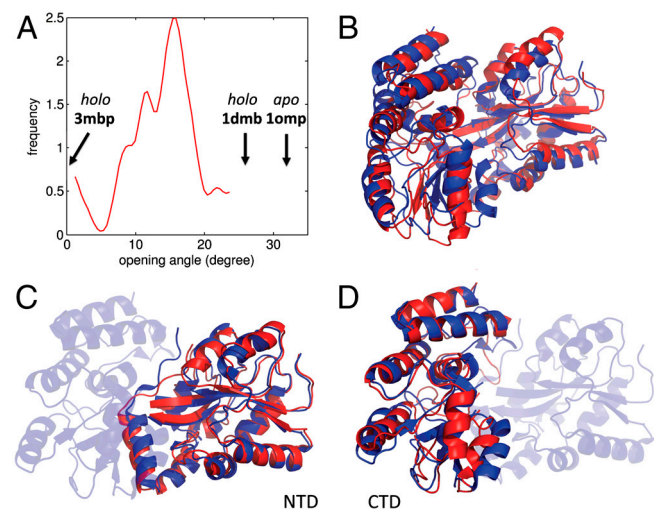
<sup>††</sup>Using experimental data from this work (ER690 ILV sample).

<sup>‡‡</sup>Data from Mueller, et al. (4).

<sup>§§</sup>Results were obtained with significantly more restraints: Five RDC vectors (instead of just N-H), CSA restraints, hydrogen-bond, and backbone dihedral angle restraints. Using only three (vs. five) one-bond RDC vectors (C $\alpha$ -C, N-C, and N-H) and the hydrogen-bond restraints in addition to the NOEs, we obtained median rmsd of 3.0 Å, 2.3 Å, and 2.2 Å for the full-protein, N-terminal domain (NTD), and C-terminal domain (CTD), respectively.

( $|i-j| \leq 4$ ), and 1,336 sequential restraints from additional NMR data and expert analysis (Table 1).

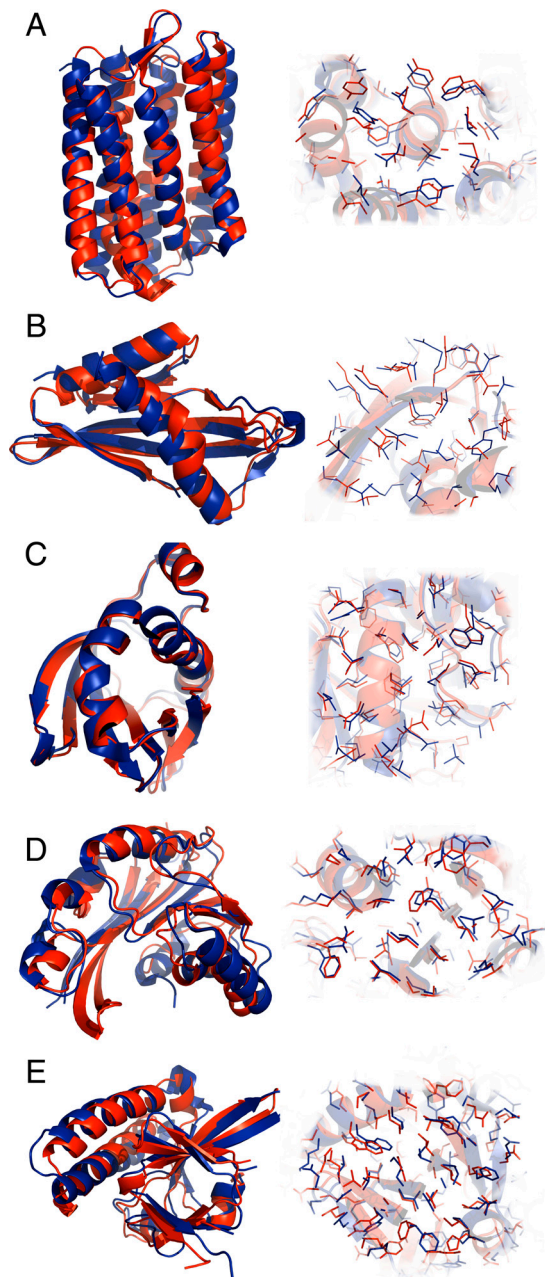
#### Determination of Previously Unsolved Structures by RASREC CS-Rosetta. To test the robustness and reliability of RASREC CS-Rosetta



**Fig. 2.** RASREC Rosetta results for maltose-binding protein. The calculations used experimental data collected in this study (ER690 ILV sample). Shown are structural superpositions of the RASREC CS-Rosetta structure (red) with the X-ray crystal structure of the *holo* protein, 1ez9 (blue). (A) Histogram of the opening angles of the 50 lowest-energy structures. (B) The RASREC CS-Rosetta structure with the smallest opening angle within the 10 lowest by Rosetta energy is superimposed on the most-closed crystal structure (3 mbp). Because the opening angle is heterogeneous (see A), we show in C and D superpositions of the N- and C-terminal domains, respectively, for the structure which best fits the RDC data among the 10 lowest by Rosetta energy.

on larger proteins, we also carried out calculations on six proteins for which the structure was not known prior to our analysis. For four of the proteins (SgR145, WR73, HsR50, and HR4660B) we used data from ILV-methyl protonated, *U*- [<sup>2</sup>H, <sup>13</sup>C, <sup>15</sup>N] labeled protein samples; for the remaining two (HmR11 and BamC) we used data from fully protonated <sup>13</sup>C, <sup>15</sup>N-enriched samples. NOE restraints were obtained using seeded automated NOESY cross-peak assignments as described in *Methods*, and N-H RDCs were also used in the structure calculations. For BamC only the 62 long-range expert seed assignments were used. In parallel, a best-effort experimental structure was determined for each of the blind targets using conventional methodology that included both manual and automated NOE cross-peak assignments and extensive manual refinement of the NOESY peak list. For BamC and SgR145, crystal structures were determined only after the RASREC CS-Rosetta NMR structures were completed.

The agreement between the blind RASREC CS-Rosetta structures and the corresponding reference structure (the crystal structure for BamC and SgR145, and the conventionally determined NMR structure for the remaining proteins) is very good for all but one of the cases (Table 2). The C $\alpha$ -rmsd to the corresponding reference structures range from 1.1–3.9 Å (Table 2 and *SI Appendix, Table S5* for restriction to converged regions). The RASREC CS-Rosetta NMR structures of both domains of BamC were published previously (25). Subsequently, the X-ray structures of both domains became available (31). Superpositions of the RASREC CS-Rosetta NMR and X-ray structures of the individual domains are shown in Fig. 3 B and C; the C $\alpha$ -rmsds are 2.6 Å and 1.1 Å, respectively (Table 2). For SgR145, an X-ray crystal structure was also solved by the NESG consortium using molecular replacement and manual model building only after the RASREC calculations were completed. A superposition of the RASREC NMR and X-ray crystal structures (1.9 Å C $\alpha$ -rmsd) is shown in Fig. 3D.



**Fig. 3.** RASREC CS-Rosetta results for five target proteins: (A) sensory rhodopsin, (B) BamC N-terminal domain, (C) BamC C-terminal domain, (D) SgR145, and (E) WR73. Rosetta structures (red) are superimposed with reference structures (blue). Each sub-figure shows the cartoon representation of the lowest-energy structure (*Left*) and close up of the core region that illustrates accuracy of side chains (*Right*). Protein Data Bank-accession codes of the X-ray reference structures in A–E are 1h68, 2yh6, 2yh5, 3mer, and 2loy, respectively.

The single failure was for HsR50, which is composed of a  $\beta$ -barrel flanked by an approximately 20-residue helix and loops with few isoleucine, valine, or leucine residues outside the barrel region. The ILV dataset yielded 209 restraints for the RASREC CS-Rosetta NMR calculation, which converged (within 3.7 Å) only over the  $\beta$ -barrel. The NMR reference structure could only be achieved with time-consuming analysis of a fully protonated double-labeled dataset. RASREC calculations using the restraint set from the double-labeled sample yielded a structure with 2.2 Å rmsd to the conventional ensemble over the full sequence and 1.6 Å rmsd over the  $\beta$ -barrel portion. The failure with the

HsR50 ILV dataset indicates the sensitivity of the method to uneven distributions of ILV residues throughout the structure; problematic regions could potentially be resolved by adding further methyl probes, such as Ala, Thr, and Met.

The RASREC approach is designed to tolerate a few incorrect seed NOE restraints. For targets SgR145, MBP, and BamC, some of the manual seed assignments turned out to be incorrect (*SI Appendix, Fig. S3*) and were violated in the lowest-energy models. For BamC, 10 out of 64 available long-range restraints were mis-assigned, but remarkably these misleading restraints did not prevent convergence to the correct fold. This demonstrates a considerable robustness of the algorithm against spurious restraints. Although this robustness confirms that the Rosetta force field has a stronger impact on the resulting structures than force fields generally have in conventional structure determination, the RASREC CS-Rosetta structures still fit the ILV–NOESY data well (*SI Appendix, Table S2*).

The RASREC CS-Rosetta NMR structures also have accurate core side-chain conformations, very similar to those in the corresponding X-ray crystal structures (Fig. 3). On average, 96% of the converged core side chains in the RASREC models are in the same  $\chi_1$  rotamer well, and 86% have the same set of rotamer states for all  $\chi$  angles (Table 3).

### Discussion

We have developed a new method for determining the 3D structures of 20–40 kDa proteins that combines sparse  $H^N$ – $H^N$ ,  $H^N$ – $CH_3$ , and  $CH_3$ – $CH_3$  distance restraint information, RDC data, and RASREC CS-Rosetta calculations. The method is particularly useful for determining protein structures with molecular weights >20 kDa, where uniform deuteration with amide and methyl protonation are required but not sufficient to produce a high-quality experimental NMR structure comparable to those of fully protonated and assigned proteins.

We have focused here on datasets with ILV  $^1H$  methyls because these generally provide the most long-range distance information per proton in the protein core. Other labeling schemes allow introduction of  $^1H$  probes on Ala, Met, and Thr (32), and on aromatic protons in Phe, Tyr, and Trp (33). Restraints obtained using these additional probes have proven useful in determining high-quality structures up to 50 kDa (34). However, additional probes also come at the expense of increased spectral crowding, more difficult manual expert analysis, and somewhat higher reagent cost. Our experience suggests that adding Ala, Met, and Thr labeling would provide the best tradeoff between spectral quality and useful long-range contact information; the presence of additional proton probes on Phe, Trp, and Tyr side chains is often detrimental to rapid methyl assignment because of the overlap between  $^1H_N$  and aromatic resonances in the 3D  $^{13}C$  NOESY–HSQC strips, which is particularly severe in large proteins. As a result, multiple samples might be needed to (i) assign methyls and (ii) obtain aromatic-specific contacts. If such restraints are available they can be readily used in RASREC Rosetta and are expected to increase robustness and accuracy of the method. This is especially true in cases where ILV residues are unfavorably distributed, as in HsR50, or in areas of high aromatic density. Otherwise, the improvement caused by the additional data is likely to be relatively small, as seen here for sensory rhodopsin when comparing our results for ILV and ILVAMT labeling.

Comparison to X-ray structures (five datasets, including two blind datasets) demonstrates that the RASREC CS-Rosetta approach generally provides higher accuracy than best-effort conventional analysis methods given the same raw data. The core side chains are also modeled accurately; on average, 96% of the converged side chains adopt the correct  $\chi_1$  rotameric well. Indeed, the RASREC CS-Rosetta NMR structure of target SgR145 (MW 22.4 kDa) was sufficiently accurate to allow phasing of diffraction data by molecular replacement. This finding

**Table 2. Summary of previously unknown protein NMR structures determined with RASREC CS-Rosetta protocol**

	Target	No. of residues	MW(kDa)	$\tau_c^*$ (ns)	Reference structure PDB_id <sup>†</sup>	Residue ranges used for rmsd analysis	C $_{\alpha}$ backbone rmsd to reference structure: low/median/high (Å)
Blind structures‡	BamC-NTD	110(246)	30	7 <sup>**</sup>	2yh6(X-ray)	2–10, 14–102 <sup>§</sup>	2.0/ 2.6/2.8
	BamC-CTD	126(246)	30	8 <sup>**</sup>	2yh5(X-ray)	1–118	0.9/1.1/1.3
	SgR145	197	22.4	12	3mer(X-ray)	21–170, 188–196 <sup>¶</sup>	1.7/1.9/2.9 <sup>  </sup>
	WR73	183	21.6	13	2loy(NMR)	1–37, 66–180 <sup>**</sup>	2.4/2.5/3.2
	HsR50	191	20.5	10	2lok(NMR)	na	unconverged <sup>††</sup>
	HmR11	185	22.1	10	2lnu(NMR)	4–180 <sup>**</sup>	2.9/3.4/4.6
	HR4660B	174	19.5	14	2lmd(NMR)	36–162 <sup>**</sup>	3.4/3.9/4.5

\*Rotational correlation times ( $\tau_c$ ) in ns were experimentally determined from <sup>15</sup>N T<sub>1</sub> and T<sub>2</sub> (CPMG) measurements (44) conducted at 800 MHz and 298 K, or estimated from the molecular assembly weight.

<sup>†</sup>Entries shaded in gray are for targets for which no crystal structure is available for comparison. In these cases, the reference structure is the mean coordinates of the manually refined ensemble of NMR structures determined by conventional methods.

<sup>‡</sup>All targets were solved using RASREC CS-Rosetta before the reference structure became available.

<sup>§</sup>Residues 11–13 are missing in X-ray coordinates. We also performed (blind) calculations of the independent N-terminal domain (NTD) and obtained better convergence with rmsd of 1.7/1.8/2.1 Å for residues 3–8 and 16–109.

<sup>¶</sup>Excluded flexible loop residues 155–164, which have missing electron density in X-ray structure.

<sup>||</sup>This result was obtained using the original protocol as a blind prediction and is thus different from the result reported in Table 1 (<sup>\*\*</sup> footnote).

<sup>\*\*</sup>Excluded residues that fluctuate more than 2 Å in reference NMR ensemble. Note that this does not take fluctuations in the Rosetta ensemble into account. In *SI Appendix, Table S5* shows residue ranges and rmsd to the reference structure when only residues with less than 2Å fluctuation in both ensembles are used.

<sup>††</sup>This structure could not be solved with the ILV approach only. The ILV-RASREC calculation only converged on the central barrel part of the fold where it overlays relatively well (3.7 Å, 58–70, 87–179) with the reference NMR structure.

<sup>†††</sup>Estimated from molecular assembly weight.

might seem surprising given that only a subset of the data used in the conventional analysis was used in the RASREC CS-Rosetta calculations. Relatively accurate models can be obtained with our approach using limited data, likely because (i) the CS-based fragments are reasonably accurate, (ii) the Rosetta all-atom force field is more complete than the CNS/XPLOR force fields typically used for NMR structure refinement (35–38), and (iii) the Rosetta low-resolution force field encapsulates a considerable amount of physical chemistry that leverages the sparse NMR restraints to generate good starting points for Rosetta full-atom refinement. This greatly reduces the chances of obtaining a struc-

ture trapped in nonnative local minima. For the sparse NMR restraint sets resulting from ILV-labeled proteins, the differences between traditional data-driven approaches and our force-field approach appear to be larger than for restraint sets derived for smaller targets from fully protonated samples. The improvement, however, comes at the cost of a much higher computational effort caused by the requirement to sample extensively the rugged energy landscape generated by the Rosetta force fields (*Methods and SI Appendix, Fig. S5*).

In the cases where an X-ray structure was not available to evaluate high-resolution accuracy, the RASREC models had the same overall topology as conventionally determined NMR structures but better side-chain and core packing as judged by traditional knowledge-based validation methods (39, 40). The robust RASREC CS-Rosetta method for determining atomic-resolution NMR solution structures, demonstrated in this study for 20–40 kDa proteins, should have a significant impact in expanding the application of NMR to a broader range of problems in structural biology.

## Methods

**NMR Spectroscopy and Data Analysis.** All NMR data for this study were collected at either 25 °C (SR10, HR4660B, WR73, SgR145, HmR11, and HsR50) or 37 °C (MBP) on Bruker Avance 800-MHz NMR spectrometers equipped with a triple-resonance TXI Cryoprobe. NOESY data were consistently acquired with 300-ms *U*-[<sup>2</sup>H, <sup>13</sup>C, <sup>15</sup>N] samples and 120 ms *U*-[<sup>13</sup>C, <sup>15</sup>N] samples mixing times. Details of data collection and analysis are presented in the *SI Appendix*.

Backbone <sup>13</sup>C, <sup>15</sup>N, and H<sup>N</sup> resonance assignments were determined using standard <sup>2</sup>H-decoupled transverse relaxation optimized spectroscopy-detected triple-resonance methods. Using a set of redundant NOESY spectra, <sup>13</sup>C and <sup>1</sup>H resonances of Ile  $\delta$ 1, Val  $\gamma$ 1, 2 and Leu  $\delta$ 1, 2 methyls were assigned. In each case, a small number of unambiguously assigned NOE interactions were first identified and used to seed the structure-generation process. Well-dispersed/isolated H<sup>N</sup> and the upfield shifted <sup>13</sup>C chemical shifts of Ile  $\delta$ 1 methyl provide ideal starting points for identification of such reliable long-range contacts. Other isolated/shifted Val  $\gamma$ 1, 2 or Leu  $\delta$ 1, 2 methyl resonances in the <sup>13</sup>C-HSQC follow; the analysis of the methyl resonances continues toward the more overlapped regions of the spectrum. Additionally, methyl to side chain tryptophan indole (N<sup>e</sup>1, H<sup>e</sup>1) NOEs were assigned, providing another important source of long-range contacts. Backbone <sup>1</sup>H–<sup>15</sup>N RDCs were measured (or obtained from the literature for MBP) in at least one alignment medium in all test cases except for sensory rhodopsin.

**Table 3. Accuracy of sidechain  $\chi$ 1 rotamers**

	Target	Number of Sidechains		Percentage correct rotamer	
		converged & buried*	correct <sup>†</sup>	$\chi$ 1 only <sup>‡</sup>	all $\chi$ -angles <sup>§</sup>
X-ray reference	SgR145	47	42	89%	82%
	SR10	36	32	86%	79%
	Rhodopsin (ILV)	64	63	98%	93%
	MBP—literature (ILV) <sup>¶</sup>	84	83	99%	96%
	MPB—literature (HN) <sup>¶</sup>	89	88	99%	94%
	MBP—this work (ILV) <sup>  </sup>	85	80	94%	90%
	bamC NTD	18	15	83%	68%
NMR	bamC CTD	33	33	100%	94%
	HR4660B	17	14	82%	80%
	WR73	37	26	70%	64%
	HmR11	29	27	93%	77%

Buried and converged side chains are selected and their adopted rotamer assignment (45) is compared to those in the reference structure (X-ray or structure 1 of NMR ensemble).

\*Side chains that are buried (SASA < 40 Å<sup>2</sup>) and converged ( $\chi$ 1 angle, SD < 10 degrees in 10 low-energy structures).

<sup>†</sup>Subset of rotamers in column 1 (converged and buried) that have a correct  $\chi$ 1 rotamer assignment.

<sup>‡</sup>Ratio of column 2 (correct) and column 1 (converged and buried).

<sup>§</sup>Percentage of side chains that are counted in column 1 (converged and buried) for which all side-chain torsion angles ( $\chi$ 1,..., $\chi$ 4) adopt the same rotamer state as in the reference structure.

<sup>¶</sup>Data taken from the publication by Mueller, et al. (4).

<sup>||</sup>Data collected specifically for this study using Northeast Structural Genomics Consortium sample ER690.

**Automatic NOESY Cross-Peak Assignment.** The manually obtained seed NOE restraints are listed in *SI Appendix, Table S1*. Between 28 and 66 such manual NOE distances per target were included. In addition, RDC data and dihedral angle restraints were provided as input to CYANA structure calculation and NOESY assignment runs (*SI Appendix: Methods*). The resulting upper-distance restraints were used for RASREC CS-Rosetta structure generation as detailed below.

**Structure Generation with RASREC CS-Rosetta.** We have used the RASREC CS-Rosetta method as described previously (18) to determine an ensemble of target structures (*SI Appendix: Methods*).

CYANA upper-distance restraints were separated into the restraints with highest reliability ( $SUP = 1$ ),  $R_{full}$ , and those with lower reliability ( $SUP < 1$ ),  $R_{sup}$  (*SI Appendix, Table S1*). This SUP entry in the CYANA .upl file is equivalent to the quality of a cross-peak assignment given in the .noa output file that is computed as:  $quality = 1.0 - \Pi_i (1.0 - prob(i))$ , where the product runs over all initial assignments of the cross-peak and  $prob(i)$  gives the probability of the individual initial assignment.  $SUP = 1$  is only reached if at least one of the individual assignments is certain i.e.,  $prob(i) = 1$  (41).

The automatic and seed restraints were converted into Rosetta flat-bottom restraints (42) as described in *SI Appendix: Methods*. In order to reduce the possible impact of spurious/incorrect restraints in  $R_{sup}$ , we combine random pairs into ambiguous restraints (43). For random combination, restraints are classified by their sequence separation with  $<5-$ ,  $<20-$ ,  $<50-$ , and  $\geq 50$ -residue separation. Random pairs are formed within each class. For each decoy a new random combination is generated. As expected, including

the automatic CYANA upper-distance restraints into the RASREC CS-Rosetta calculations in addition to the manually assigned seed restraints improves the accuracy for most targets (*SI Appendix, Table S4*).

The method requires substantial computer resources. For instance, target HmR11 requires 7 h on 64 machines of a Linux cluster, which has two quad-core CPUs of 2.93-GHz Intel Xeon 5570 per motherboard (i.e., 512 compute cores). The required time depends on several factors including size, density, and intricateness of the restraints, and fold complexity (see *SI Appendix, Fig. S5*). Although these computer requirements generally exceed the in-lab resources of the average NMR lab, it is not problematic nowadays to allocate such resources e.g., through adjunct computer centers, cloud computing, or a grid project such as the European Grid Infrastructure (<http://www.egi.eu>).

**ACKNOWLEDGMENTS.** We thank Yuanpeng Huang for helpful discussions, Frank DiMaio for help with molecular replacement of SgR145, Daniel Nietlispach for discussions regarding rhodopsin, and Phil Kostenbader for cluster computing support. We thank Department of Energy Innovative and Novel Computational Impact on Theory and Experiment (INCITE) Award for providing access to the Blue Gene/P supercomputer at the Argonne Leadership Computing Facility and to the Juelich Supercomputing Centre for providing access to JUROPA. This work was supported by the Human Frontiers of Science Program (O.F.L.), the DFG grant LA 1817/3-1 (to O.F.L.), National Institutes of Health Grant GM76222 (to D.B.), the Howard Hughes Medical Institute, and the National Institutes of General Medical Science Protein Structure Initiative Program Grant U54 GM-094597 (to G.T.M.).

- Pervushin K, Riek R, Wider G, Wuthrich K (1997) Attenuated T2 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.
- Markley JL, Putterl I, Jardetzky O (1968) High-resolution nuclear magnetic resonance spectra of selectively deuterated staphylococcal nuclease. *Science* 161:1249–1251.
- 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.
- Mueller GA, et al. (2000) Global folds of proteins with low densities of NOEs using residual dipolar couplings: Application to the 370-residue maltodextrin-binding protein. *J Mol Biol* 300:197–212.
- Schwieters CD, et al. (2010) Solution structure of the 128 kDa enzyme I dimer from *Escherichia coli* and its 146 kDa complex with HPr using residual dipolar couplings and small- and wide-angle X-ray scattering. *J Am Chem Soc* 132:13026–13045.
- Frueh DP, et al. (2008) Dynamic thiolation-thioesterase structure of a non-ribosomal peptide synthetase. *Nature* 454:903–906.
- Tugarinov V, Choy WY, Orekhov VY, Kay LE (2005) Solution NMR-derived global fold of a monomeric 82-kDa enzyme. *Proc Natl Acad Sci USA* 102:622–627.
- Gelis I, et al. (2007) Structural basis for signal-sequence recognition by the translocase motor SecA as determined by NMR. *Cell* 131:756–769.
- Grzesiek S, Bax A (1993) Measurement of amide proton exchange rates and NOEs with water in  $^{13}C/^{15}N$ -enriched calcineurin B. *J Biomol NMR* 3:627–638.
- LeMaster DM, Kay LE, Brunger AT, Prestegard JH (1988) Protein dynamics and distance determination by NOE measurements. *FEBS Lett* 236:71–76.
- Rosen MK, et al. (1996) Selective methyl group protonation of perdeuterated proteins. *J Mol Biol* 263:627–636.
- Gardner KH, Rosen MK, Kay LE (1997) Global folds of highly deuterated, methyl-protonated proteins by multidimensional NMR. *Biochemistry* 36:1389–1401.
- Gardner KH, Kay LE (1998) The use of  $^2H$ ,  $^{13}C$ ,  $^{15}N$  multidimensional NMR to study the structure and dynamics of proteins. *Annu Rev Biophys Biomol Struct* 27:357–406.
- Venters RA, et al. (1995) High-level  $^2H/^{13}C/^{15}N$  labeling of proteins for NMR studies. *J Biomol NMR* 5:339–344.
- Fernandez C, Hilty C, Wider G, Guntert P, Wuthrich K (2004) NMR structure of the integral membrane protein OmpX. *J Mol Biol* 336:1211–1221.
- Hiller S, et al. (2008) Solution structure of the integral human membrane protein VDAC-1 in detergent micelles. *Science* 321:1206–1210.
- Kainosho M, et al. (2006) Optimal isotope labelling for NMR protein structure determinations. *Nature* 440:52–57.
- Lange OF, Baker D (2012) Resolution-adapted recombination of structural features significantly improves sampling in restraint-guided structure calculation. *Proteins* 80:884–895.
- Raman S, et al. (2010) NMR structure determination for larger proteins using backbone-only data. *Science* 327:1014–1018.
- Guntert P, Mumenthaler C, Wuthrich K (1997) Torsion angle dynamics for NMR structure calculation with the new program DYANA. *J Mol Biol* 273:283–298.
- 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.
- Rieping W, et al. (2007) ARIA2: Automated NOE assignment and data integration in NMR structure calculation. *Bioinformatics* 23:381–382.
- 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.
- Wunderlich Z, et al. (2004) The protein target list of the Northeast Structural Genomics Consortium. *Proteins* 56:181–187.
- Warner LR, et al. (2011) Structure of the BamC two-domain protein obtained by Rosetta with a limited NMR dataset. *J Mol Biol* 411:83–95.
- Gautier A, Mott HR, Bostock MJ, Kirkpatrick JP, Nietlispach D (2010) Structure determination of the seven-helix transmembrane receptor sensory rhodopsin II by solution NMR spectroscopy. *Nat Struct Mol Biol* 17:768–774.
- Guntert P (2004) Automated NMR structure calculation with CYANA. *Methods Mol Biol* 278:353–378.
- Tang C, Schwieters CD, Clore GM (2007) Open-to-closed transition in apo maltose-binding protein observed by paramagnetic NMR. *Nature* 449:1078–1082.
- Duan X, Quirocho FA (2002) Structural evidence for a dominant role of nonpolar interactions in the binding of a transport/chemosensory receptor to its highly polar ligands. *Biochemistry* 41:706–712.
- Royant A, et al. (2001) X-ray structure of sensory rhodopsin II at 2.1-Å resolution. *Proc Natl Acad Sci USA* 98:10131–10136.
- Albrecht R, Zeth K (2011) Structural basis of outer membrane protein biogenesis in bacteria. *J Biol Chem* 286:27792–27803.
- Isaacson RL, et al. (2007) A new labeling method for methyl transverse relaxation-optimized spectroscopy NMR spectra of alanine residues. *J Am Chem Soc* 129:15428–15429.
- Fesik SW, Gampe RT, Jr, Zuiderweg ER, Kohlbrenner WE, Weigl D (1989) Heteronuclear three-dimensional NMR spectroscopy applied to CMP-KDO synthetase (27.5 kD). *Biochem Biophys Res Commun* 159:842–847.
- Popovych N, Tzeng SR, Tonelli M, Ebricht RH, Kalodimos CG (2009) Structural basis for cAMP-mediated allosteric control of the catabolite activator protein. *Proc Natl Acad Sci USA* 106:6927–6932.
- Schueler-Furman O, Wang C, Bradley P, Misura K, Baker D (2005) Progress in modeling of protein structures and interactions. *Science* 310:638–642.
- Tyka MD, et al. (2011) Alternate states of proteins revealed by detailed energy landscape mapping. *J Mol Biol* 405:607–618.
- Qian B, et al. (2007) High-resolution structure prediction and the crystallographic phase problem. *Nature* 450:259–264.
- Verma A, Wenzel W (2007) Protein structure prediction by all-atom free-energy refinement. *BMC Struct Biol* 7:12–27.
- Ramelot TA, et al. (2009) Improving NMR protein structure quality by Rosetta refinement: A molecular replacement study. *Proteins* 75:147–167.
- Bhattacharya A, Tejero R, Montelione GT (2007) Evaluating protein structures determined by structural genomics consortia. *Proteins* 66:778–795.
- Guntert P (2009) Automated structure determination from NMR spectra. *Eur Biophys J* 38:129–143.
- Leaver-Fay A, et al. (2011) ROSETTA3: An object-oriented software suite for the simulation and design of macromolecules. *Methods Enzymol* 487:545–574.
- 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.
- Farrow NA, et al. (1994) Backbone dynamics of a free and phosphopeptide-complexed Src homology 2 domain studied by  $^{15}N$  NMR relaxation. *Biochemistry* 33:5984–6003.
- Dunbrack RL, Jr (2002) Rotamer libraries in the 21st century. *Curr Opin Struct Biol* 12:431–440.