



Published in final edited form as:

*Science*. 2010 February 19; 327(5968): 1014–1018. doi:10.1126/science.1183649.

## NMR Structure Determination for Larger Proteins Using Backbone-Only Data

Srivatsan Raman<sup>1,7,\*</sup>, Oliver F. Lange<sup>1,\*</sup>, Paolo Rossi<sup>2</sup>, Michael Tyka<sup>1</sup>, Xu Wang<sup>3</sup>, James Aramini<sup>2</sup>, Gaohua Liu<sup>2</sup>, Theresa Ramelot<sup>5</sup>, Alexander Eletsy<sup>6</sup>, Thomas Szyperski<sup>6</sup>, Michael Kennedy<sup>5</sup>, James Prestegard<sup>3</sup>, Gaetano T. Montelione<sup>2</sup>, and David Baker<sup>1,4,#</sup>

<sup>1</sup>Department of Biochemistry, University of Washington, Seattle, WA 98195, USA

<sup>2</sup>Center for Advanced Biotechnology and Medicine, Department of Molecular Biology and Biochemistry, and Northeast Structural Genomics Consortium, Rutgers University, Piscataway, NJ 08854

<sup>3</sup>Complex Carbohydrate Research Center, University of Georgia, Athens, GA 30602

<sup>4</sup>Howard Hughes Medical Institute, Seattle, WA 98195

<sup>5</sup>Department of Chemistry and Biochemistry and Northeast Structural Genomics Consortium, Miami University, Oxford, OH

<sup>6</sup>Department of Chemistry, State University of New York at Buffalo, Buffalo, New York 14260

### Abstract

Conventional protein structure determination from nuclear magnetic resonance data relies heavily on side-chain proton-proton distances. The necessary side-chain resonance assignment, however, is labor intensive and prone to error. Here we show that structures can be accurately determined without NMR information on the sidechains for proteins up to 25 kDa by incorporating backbone chemical shifts, residual dipolar couplings, and amide proton distances into the Rosetta protein structure modelling methodology. These data, which are too sparse for conventional methods, serve only to guide conformational search towards the lowest energy conformations in the folding landscape; the details of the computed models are determined by the physical chemistry implicit in the Rosetta all atom energy function. The new method is not hindered by the deuteration required to suppress nuclear relaxation processes for proteins greater than 15 kDa, and should enable routine NMR structure determination for larger proteins.

The first step in protein structure determination by NMR is chemical shift assignment for the backbone atoms. In contrast to the subsequent assignment of the sidechains, this is now rapid, reliable, and largely automated (1–5). Global backbone structural information complementing the local structure information provided by backbone chemical shift assignments (6,7), can be obtained from H<sup>N</sup>-H<sup>N</sup> NOESY, residual dipolar coupling (RDC) (8), and other (9,10) experiments. For larger proteins, deuteration becomes necessary to circumvent the efficient spin relaxation properties resulting from their higher rotational correlation times(11,12), but removing protons also eliminates long range NOESY information from sidechains except for selectively protonated sidechain moieties (13). The difficulty in determining accurate structures with no or limited sidechain information is a

<sup>#</sup>To whom correspondence should be addressed. Phone: (206) 543-1295, Fax: (206) 685-1792, dabaker@u.washington.edu.

<sup>7</sup>Current address: Department of Genetics, Harvard Medical School, Boston, MA 02115

\*S.R. and O.F.L. contributed equally to this work

major bottleneck that currently prevents routine application of NMR to larger (> 15 kDa) systems(14).

Here we show that structures of proteins up to 200 residues (23 kDa) can be determined using information from backbone ( $H^N$ , N,  $C^\alpha$ ,  $C^\beta$ , C') NMR data by taking advantage of the conformational sampling and all atom energy function in the Rosetta structure prediction methodology, which for small proteins in favorable cases can produce atomic accuracy models starting from sequence information alone(15). Structure prediction in Rosetta proceeds in two steps; first a low resolution exploration phase using Monte-Carlo fragment assembly and a coarse-grained energy function, and second a computationally expensive refinement phase which cycles between combinatorial sidechain optimization and gradient-based minimization of all torsional degrees of freedom in a physically-realistic all-atom forcefield(15). The primary obstacle to Rosetta structure prediction from amino acid sequence information alone is conformational sampling; native structures almost always have lower energies than non-native conformations, but are very seldom sampled in unbiased trajectories. Incorporating NMR chemical shift information in the selection of the fragments used in the exploration phase (CS-Rosetta) (16,17) provides a robust approach to determining accurate structures of small (< 100 residue) proteins using only backbone and  $^{13}C^\beta$  chemical shift data. For larger (> 12 kDa) proteins, the performance of CS-Rosetta is very target dependent: structures sufficiently close to the native structure for the energy to drop significantly may be generated rarely or not at all.

We investigated whether RDC data, which provide long range information on the orientations between bond vectors, can guide the low resolution search closer to the native structure and overcome the sampling problem for larger (100 – 200 residue) proteins. For every attempted Monte Carlo move, the alignment tensor is calculated by singular value decomposition(18), and the decision to accept or reject the conformation is biased by the change in the agreement between the back-calculated and experimental couplings(19). Incorporation of RDCs dramatically improved convergence on the correct structure in a benchmark of 11 alpha, beta and alpha/beta proteins ranging in size from 62 to 166 residues (Figs 1, S1 and Table 1). As indicated in Table 1, CSRDC-Rosetta consistently generates accurate models for proteins up to 120 residues, and in favorable cases for larger proteins.

For proteins with over 120 residues, conformational sampling becomes limiting even for the CS-RDC-Rosetta protocol and the low energy ensemble is not always close to the native structure. To further focus sampling, we developed an iterative refinement protocol that incorporates assigned backbone  $H^N$ - $H^N$  NOEs in addition to backbone RDCs. As in the previously described “rebuild and refine” protocol, a pool of diverse low energy conformations is maintained and the highest energy structures in the pool are periodically replaced with offspring(20). The new protocol, a genetic algorithm, generates hybrid conformations by recombining first beta sheet pairings and subsequently fragments of the low energy structures (see methods). To further enhance sampling, trajectories are seeded with conformations harvested from previous trajectories that led to low energy conformations(21).

The improvement in the model population with increasing generations in the iterative protocol is illustrated in Figure 2 for the 200-residue ALG13 protein using experimentally determined chemical shift, RDC, and assigned backbone amide  $H^N$ - $H^N$  NOE data(22). The  $C^\alpha$ -RMSD to the native structure and the energy improve from generation to generation, and after several rounds, discrimination towards lower RMSD structures is apparent (Figure 2a, cyan to yellow). After high resolution refinement (Figure 2a, orange to red), the lowest energy structures are close to the native structure. The final low energy structural ensemble (Figure 2b) recapitulates the unusual topology in the previously determined NMR

structure(22) (Figure 2d) to within 3.4 Å RMSD (Table 1). The Rosetta ensemble fits independent RDC data as well as the NMR structure, and the backbone variation in the ensemble is correlated with backbone dynamics as probed by the R1 relaxation rate (Figure 2c). The iterative-CS-RDC-NOE-Rosetta models of ALG13 thus appear to be comparable in quality to the previously published structure that required substantial effort, including preparation of selectively methyl and aromatic-protonated samples(22).

The iterative-CS-RDC-NOE protocol was tested further on 12 proteins with sizes ranging from 120 to 266 residues (Table 1 and Fig S3). For all proteins but 1g68 a considerable part of the structure converges (Table 1). Backbone  $H^N$ - $H^N$  NOE data was required for convergence of 2z2i, 1i1b, arf1, 2rn2 and 1sua but not for 5pnt, 1s0p, 1f21 and er553. The RMSDs to the native structures over the converged regions range from 1.7–4.3 Å with the exceptions of 1sua and 1f21. For 1f21 high accuracy (1.6 Å) was reached for a 92 residue subset (Fig S3). Sidechain accuracy was generally quite high in the converged regions (Fig S5).

We carried out a blind test of the new methods on five data sets generated in the Northeast Structural Genomics (NESG) Center before conventional NMR structures were determined. For four of the proteins, the CS-RDC protocol converged (Figure 3a–d), while for a fifth, convergence was not observed and blind structure determination was instead carried out using the iterative CS-RDC-NOE protocol (Figure 3e). In all five cases (Table I) the resulting Rosetta determined structure is very similar to the conventionally determined NMR solution structure over both the backbone (Figure 3, left panels) and the core sidechains (Figure 3, right panels), which is notable because no experimental sidechain information is used in the Rosetta protocol; the details of core packing are determined by the Rosetta all atom energy function.

Thus, our methodology is able to generate accurate structures of proteins up to ~25 kDa from sparse NMR data without side-chain assignments. To be useful in practice, it is important that there be a means of assessing the reliability of the computed models. Cross-validation with independently collected data is an excellent way to do this, but truly independent data may not always be available, and if the available data are already sparse, it may not be possible to remove a subset for independent validation.

Our approach to structure validation is based on the interplay between the two contributing sources of structural information—the detailed physical chemistry implicit in the Rosetta all atom energy function, and the experimental NMR data. As illustrated in Figure 4a, the all atom energy landscape (black) is rugged with many local minima, making optimization difficult. The experimental bias based on backbone NMR data (red), although smoother, is degenerate and lacks resolution. Since the constrained minimization of a function will almost always result in higher function values than unconstrained minimization, NMR data constrained optimization in general should result in higher energy structures than bias-free optimization (arrow 1 in Figure 4a). This scenario may hold for traditional structure determination in which the search is almost completely driven by the experimental data. However, if the two sources of information are in concordance, the bias from the experimental data can have two favorable effects (Figure 4b)—first, optimization far from the native minimum is impeded, resulting in an upward shift of the energy of non-native structures (arrow 1), and second, optimization near the native minimum is improved as the data guide the search towards the global minimum (Figure 4a–b, arrow 2).

The scenario illustrated in Figure 4b is unlikely to occur if there is no sampling near the correct structure: the experimental data and the energy function will almost never independently favor the same incorrect structure. Hence, we propose the following three

criteria for evaluating the reliability of a calculated structure (Table 1, column 6–8). First, the calculation should converge — the lowest energy conformations should be very similar to each other over a large fraction of the structure. For both the CS-RDC-Rosetta and the iterative protocol, whenever the calculation converged for more than 60% of the structure, the RMSD to native over this region was less than 4 Å (Table 1, column 6). Second, the converged structures should be clearly lower in energy than all significantly different (RMSD greater than 7 Å) structures; this was the case for nearly all of our test cases (Table 1, column 7). Third, the structures generated with experimental data should be at least as low in energy as those generated without experimental data; for none of the successful calculations does the energy increase significantly when the experimental data are included in the optimization (Table 1, column 8). For larger proteins (>120) the data in fact guide the trajectories to lower energy structures than obtained by unconstrained optimization (Figure 4d and Table 1 column 8)— as argued above, this is a strong indicator that the correct structure has been found.

For the twenty proteins in our test set, when all three criteria were satisfied the low energy ensemble resembles the independently determined structures. Importantly, the clear structure calculation failure — 1f21 — which converged to a wrong conformation with an RMSD of 9.4 to the native, fails the third criterion: the energy is higher rather than lower when the experimental data are included in the optimization (Figure 4c and Table 1, column 8). Since we had only one such failure, we simulated additional failures by deleting all near native structures from the model populations and computed the three metrics described above for these 'fake' minima, (Table S1, cf. Methods). For almost all the proteins, these constructed pathological cases again fail the third criterion: they have higher energies in the experimentally biased optimization.

For the proteins in our set in the ~30 kDa molecular weight range, the computed structures are not completely converged and have large disordered regions. This is clearly a sampling problem since the native structure has lower energy (Figure 4d, S3); even with the NMR data as a guide, Rosetta trajectories fail to sample very close to the native state. Increased convergence on the low energy native state can be achieved either by collecting and utilizing additional experimental data (1ilb\_2 in Fig S3;) or by improved sampling. While at present the former is the more reliable solution, the latter will likely become increasingly competitive as the cost of computing decreases and conformational search algorithms improve.

We have shown that accurate structures can be computed for a wide range of proteins using backbone only NMR data. These results suggest a change in the traditional NOE-constraint-based approach to NMR structure determination (Suppl. Fig. S4). In the new approach, the bottlenecks of sidechain chemical shift assignment and NOESY assignment are eliminated, and instead, more backbone information is collected: RDCs in one or more media, and a small number of unambiguous  $H^N-H^N$  constraints are collected using 3D or 4D experiments, which restrict the number of  $\beta$ -strand registers. Advantages of the approach are that  $^1H, ^{15}N$ -based NOE and RDC data quality is relatively unaffected in slower tumbling larger proteins and that the analysis of resonance and NOESY peak assignments can be done in a largely automated fashion with fewer opportunities for error. The approach is compatible with deuteration necessary for proteins greater than 15 kDa, and for larger proteins can be extended to include methyl- NOEs on selectively protonated samples. The method should also enable a more complete structural characterization of transiently populated states(24) for which the available data are generally quite sparse.

## Summary

Protein structures can be determined using the limited NMR information obtainable for larger proteins

## Acknowledgments

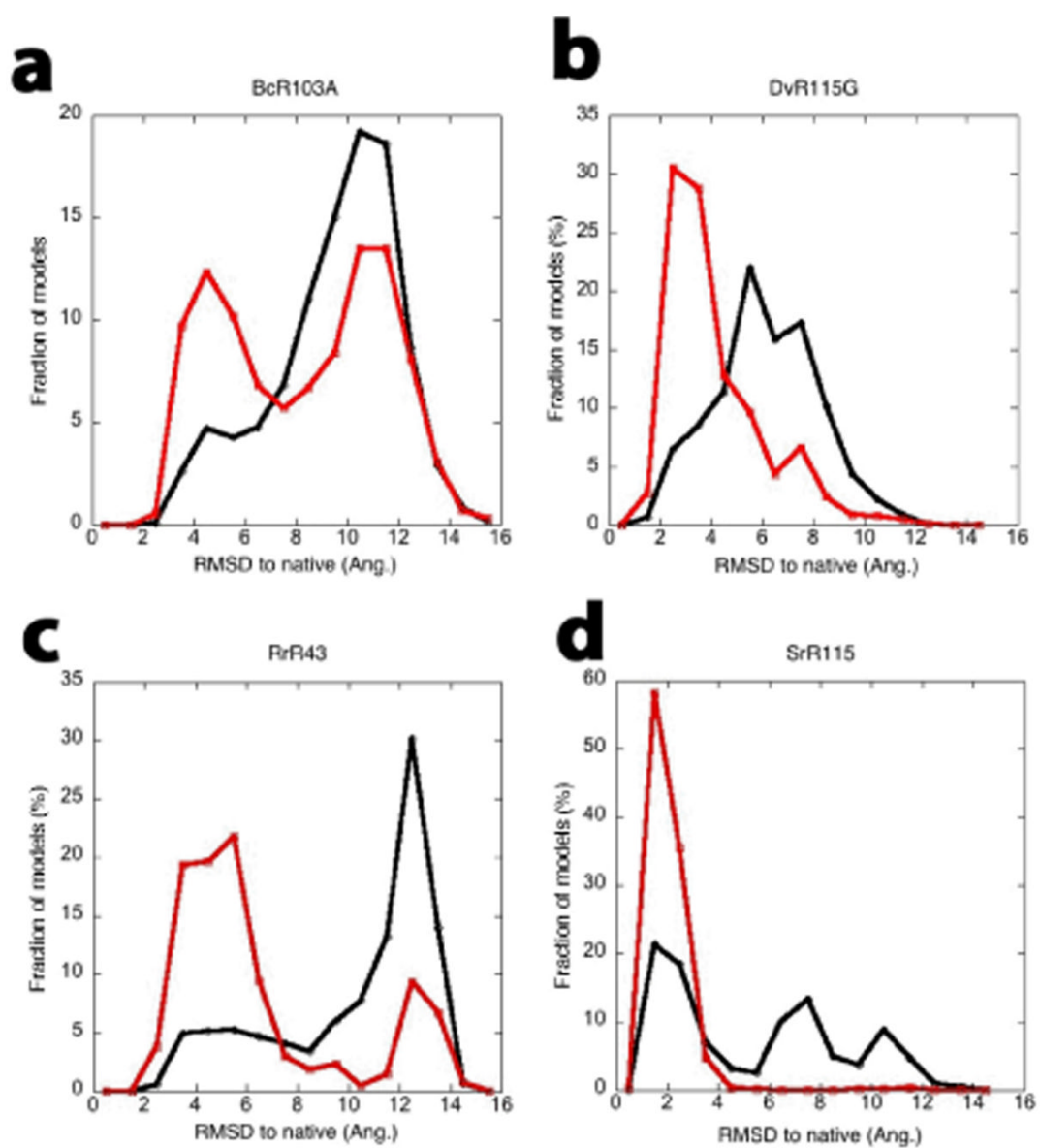
We are thankful to the DoE INCITE award for providing access to the Blue Gene/P supercomputer at the Argonne Leadership Computing Facility and to Rosetta@home participants for their generous contributions of computing power. We thank Yang Shen and Ad Bax for fruitful discussions, Y. Janet Huang and Yeufeng Tang for their contribution during preliminary studies using sparse NOE constraints with CS-Rosetta, Sonal Bansal, Hsiau-wei Lee, and Yizhou Liu for collection of RDC data, Alexander Lemak for providing the CNS RDC refinement protocol, and the NESG consortium for access to other unpublished NMR data that has facilitated methods development. S.R., O.F.L., P.R., G.T.M. and D.B. designed research, S.R. designed and tested the CS-RDC-Rosetta protocol, O.F.L. designed and tested the iterative CS-RDC-NOE-Rosetta protocol, S.R., O.F.L. and D.B. designed and performed research for energy based structure validation, X.W and J.P analysed the ALG13 ensemble, J.A, G.L, T.R, A.E, M.K, T.S provided blind NMR datasets, S.R., O.F.L., P.R., G.T.M., and D.B. wrote the manuscript. This work was supported by the Human Frontiers of Science Program (to O.F.L.), by the National Institutes of Health grant GM76222 (to D.B), the HHMI, the National Institutes of General Medical Science Protein Structure Initiative program, grants U54 GM074958 (to G.T.M) and the Research Resource grant RR005351 (to J.P).

## References

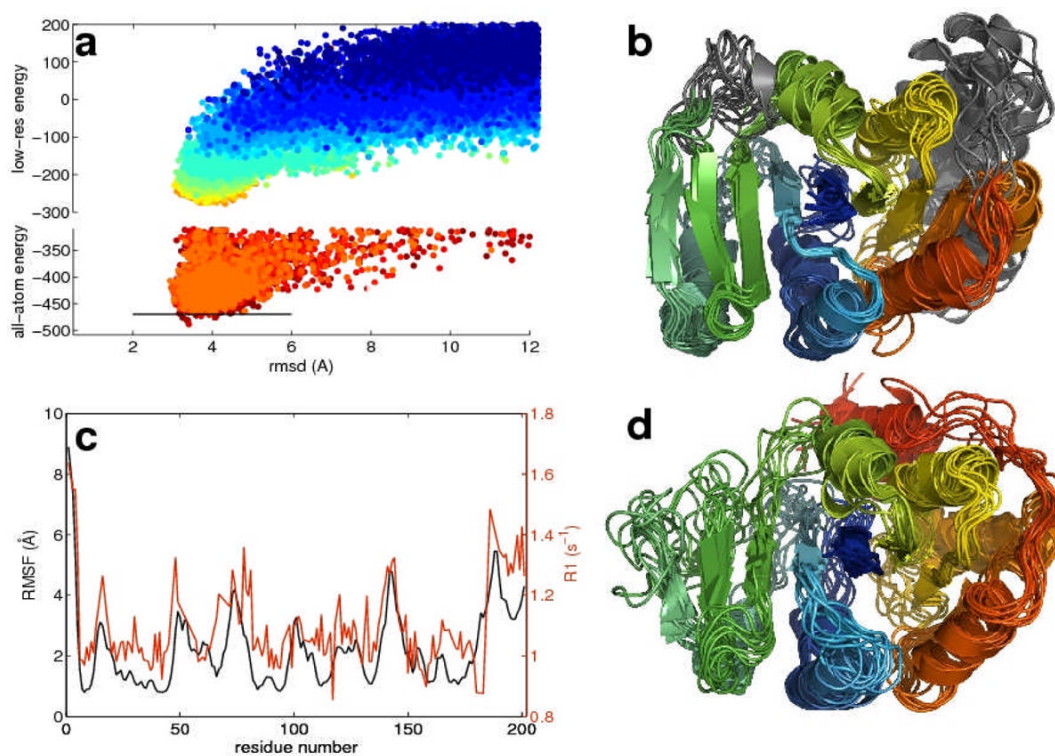
1. Zimmerman DE, et al. *J Mol Biol* Jun 20;1997 269:592. [PubMed: 9217263]
2. Bartels C, Guntert P, Billeter M, Wuthrich K. *J Comp Chem* 1998;18:139.
3. Baran MC, Huang YJ, Moseley HN, Montelione GT. *Chem Rev* Aug;2004 104:3541. [PubMed: 15303826]
4. Jung YS, Zweckstetter M. *J Biomol NMR* Sep;2004 30:11. [PubMed: 15452431]
5. Lee W, Westler WM, Bahrami A, Eghbalnia HR, Markley JL. *Bioinformatics* Aug 15;2009 25:2085. [PubMed: 19497931]
6. Berjanskii M, et al. *Nucleic Acids Res* Jul 1;2009 37:W670. [PubMed: 19406927]
7. Shen Y, Delaglio F, Cornilescu G, Bax A. *J Biomol NMR* Aug;2009 44:213. [PubMed: 19548092]
8. Kontaxis G, Delaglio F, Bax A. *Methods Enzymol* 2005;394:42. [PubMed: 15808217]
9. Bertini I, Luchinat C, Parigi G, Pierattelli R. *Dalton Trans* Aug 7;2008 :3782. [PubMed: 18629397]
10. Prestegard JH, Bougault CM, Kishore AI. *Chem Rev* Aug;2004 104:3519. [PubMed: 15303825]
11. Grzesiek S, Bax A. *J Biomol NMR* Nov;1993 3:627. [PubMed: 8111229]
12. LeMaster DM, Richards FM. *Biochemistry* Jan 12;1988 27:142. [PubMed: 3280013]
13. Gardner KH, Rosen MK, Kay LE. *Biochemistry* Feb 11;1997 36:1389. [PubMed: 9063887]
14. Wagner G. *J Biomol NMR* Jul;1993 3:375. [PubMed: 8400829]
15. Bradley P, Misura KM, Baker D. *Science* Sep 16;2005 309:1868. [PubMed: 16166519]
16. Cavalli A, Salvatella X, Dobson CM, Vendruscolo M. *Proc Natl Acad Sci U S A* Jun 5;2007 104:9615. [PubMed: 17535901]
17. Shen Y, et al. *Proc Natl Acad Sci U S A* Mar 25;2008 105:4685. [PubMed: 18326625]
18. Losonczi JA, Andrec M, Fischer MW, Prestegard JH. *J Magn Reson* Jun;1999 138:334. [PubMed: 10341140]
19. Rohl CA, Baker D. *J Am Chem Soc* Mar 20;2002 124:2723. [PubMed: 11890823]
20. Qian B, et al. *Nature* Nov 8;2007 450:259. [PubMed: 17934447]
21. Brunette TJ, Brock O. *Proteins* Dec;2008 73:958. [PubMed: 18536015]
22. Wang X, Weldeghiorghis T, Zhang G, Imperiali B, Prestegard JH. *Structure* Jun;2008 16:965. [PubMed: 18547528]
23. Zheng D, et al. *Protein Sci* Jun;2003 12:1232. [PubMed: 12761394]
24. van Ingen H, Korzhnev DM, Kay LE. *J Phys Chem B* Jul 23;2009 113:9968. [PubMed: 19569643]
25. Snyder DA, Montelione GT. *Proteins* Jun 1;2005 59:673. [PubMed: 15822099]

26. Zemla A. Nucleic Acids Res Jul 1;2003 31:3370. [PubMed: 12824330]





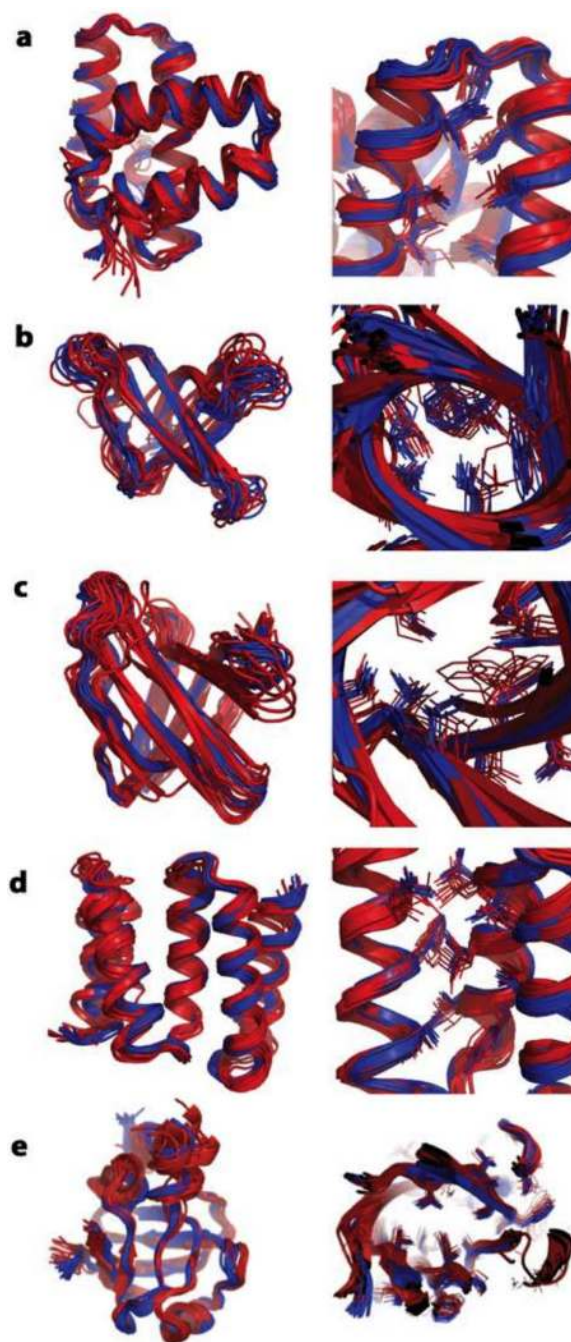
**Figure1.** Impact of RDC data on conformational search. Lines depict RMSD histograms for the lowest low-resolution energy 10% of structures generated using CS-Rosetta(black) or CS-RDC-Rosetta(red). **(a)** BcR103A **(b)** DvR115G **(b)** RrR43 **(d)** SrR115C



**Figure 2.**

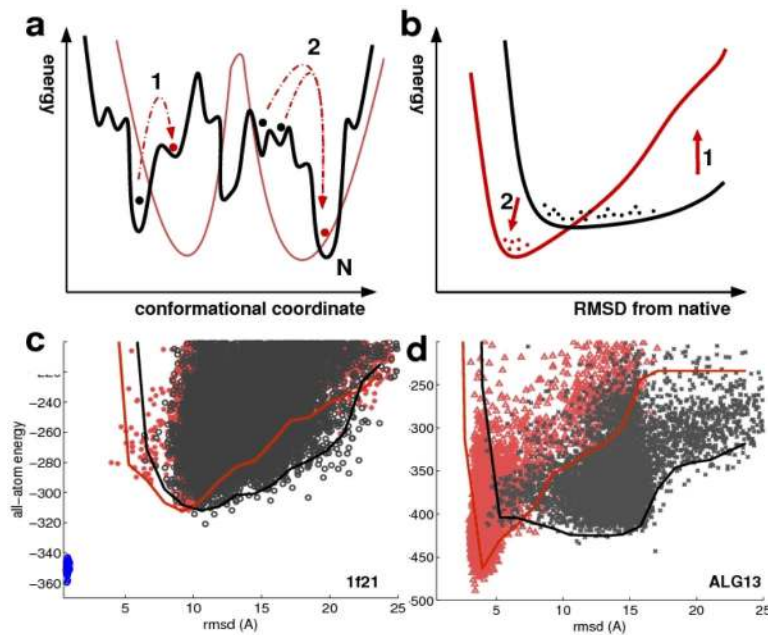
Determination of ALG13 structure from backbone NMR data with Rosetta. **(a)** RMSDs and energies of structures generated in batches of 2000 during the iterative protocol. Each generation of structures (color code: blue to red, corresponds to number of generation) is based on information from previous runs (cf. Methods). Strong convergence is reached already in the computationally less expensive low-resolution mode. The last generations (orange to red) increase both the precision and accuracy of the ensemble, by refining the structures within the Rosetta all-atom energy. The RMSD is computed over the residues for which convergence within  $3\text{\AA}$  root mean square fluctuations (RMSF) was reached in the 50 lowest energy Rosetta models (5–70, 81–139, 151–180). **(b)** Ensemble of 10 lowest energy Rosetta structures (below line in panel a). Regions with more than  $3\text{\AA}$  RMSF are colored in grey. **(c)** Comparison of the RMSF at each residue in the low energy Rosetta ensemble to NMR R1 relaxation rate (Red, relaxation rates; black, RMSF in Rosetta ensemble). Regions variable in the low energy structures exhibit increased dynamics in solution; these data were not used in the structure calculation. **(d)** NMR solution ensemble based on side-chain NOEs, RDC and PRE data as deposited in the Protein Data Bank (PDB code: 2jzc).





**Figure 3.**

Blind predictions with the CS-RDC-Rosetta and iterative CS-RDC-Rosetta protocols. Left panels: superposition of the lowest energy 10 predicted structures (red) over the experimentally solved ensemble of NMR structures (blue); right panels: magnified view of the core side-chains. Rosetta models in panels (a–d) were determined with CS-RDC-Rosetta and in (e) with iterative CS-RDC-Rosetta. (a) BcR268F (b) DvR115G (c) MaR214A (d) SrR115C (e) AtT7



**Figure 4.**

Effect of incorporation of experimental data on energy minimization. **(a)** The Rosetta all atom energy (black line) has many local minima making minimization difficult, but the global minimum is generally close to the native structure (N). The experimental bias (red line), while smoother, has degeneracies and lacks resolution because the data are sparse. Local minima of the all-atom energy and the experimental bias are uncorrelated far away from the native structure but coincide close to the native structure. Accordingly, far from the global minimum, including the experimental data during optimization usually results in higher energies (arrow 1), while close to the native structure (N), including the data results in lower energies (arrow 2). **(b)** Lines represent the lowest energies sampled by structures at various RMSDs after optimization in the absence (black line) or presence (red line) of experimental data. Generally, the all-atom energy and experimental data are in concordance for conformations close to the native protein structure but not for conformations far from the native structure. If this concordance condition is met, close to the native structure the experimental data can guide sampling towards the global minimum (arrow 2) and thus constrained optimization can result in lower energy conformations than unconstrained optimization distant from the native structure leading to higher energies (arrow 1). In contrast, **(c–d)** All-atom energy and RMSD of final Rosetta ensemble from iterative refinement with and without experimental data. Lines represent the median of the 10 lowest energy models per RMSD-bin. **(c)** 1f21 – an unsuccessful calculation; biased optimization with RDC data (red) yields similar energies as unbiased optimization (black); there is a large remaining energy gap to the native structure (blue dots). **(d)** ALG13 – a successful calculation; biased optimization with the experimental data (red) results in lower energies than unbiased optimization (black).

Table 1

Accuracy of models generated with backbone-only NMR data and metrics for validation. Figures marked in bold indicate violation of validation criterion.

Protein Name <sup>1</sup>	Native PDB ID	Topology	Number of residues/ Number of residues converged in computed structure	Median RMSD to native over converged region <sup>2</sup> (Å)	Median GDT-TS among lowest energy models <sup>3</sup>	Depth of converged ensemble energy minimum <sup>7</sup>	Median energy change resulting from inclusion of experimental data <sup>8</sup>
GmR137r,*	2k5p	a/b	62/47	2.6	95.4	-32.5	-1.8
TR80r,*	2jxt	a/b	78/73	1.5	84.9	-16.7	-0.3
DvR115G <sup>r,b</sup>	2ket	B	86/66	1.4	80.0	-24.3	-0.7
LkR15r,*	2k3d	a/b	92/74	2.0	85.4	-18.0	-1.2
BcR103A <sup>r</sup>	2kd1	B	100/65	3.4	61.3	-22.7	-1.3
SrR115C <sup>r,b,*</sup>	2kcl	A	100/95	1.4	86.1	-25.1	<b>0.7</b>
MaR214A <sup>r,b</sup>	2kbn	B	102/96	2.1	82.1	-43.9	-0.6
RrR43 <sup>r</sup>	2k0m	a/b	104/82	2.1	66.8	-12.9	<b>2.9</b>
BcR268F <sup>r,b,*</sup>	2k5w	A	118/115	1.4	78.4	-45.7	-1.6
ERS53 <sup>r</sup>	2k1s	a/b	143/115	5.2	<b>46.1</b>	-5.1	-2.5
ARF1 <sup>r</sup>	2k5u	a/b	166/141	2.6	73.3	-21.6	-9.5
AlT7 <sup>r,b</sup>	2ki8	a/b	122/98	3.0	70.0	-37.5	-12.5
ERS41 <sup>s</sup>	2jyx	a/b	124/115	2.5	76.7	-31.6	-9.9
X-ray <sup>s</sup>	1f21	a/b	142/122	9.4	76.6	-28.5	<b>3.5</b>
ERS53 <sup>r</sup>	2k1s	a/b	143/136	1.9	85.2	-38.5	-15.4
BtR324B <sup>s</sup>	2kd7	B	150/148	2.4	79.3	-51.6	-29.1
X-ray <sup>s</sup>	1i1b	B	151/115	2.5	71.1	-53.3	-25.5
X-ray <sup>s</sup>	1i1b_2 <sup>d</sup>	B	151/133 <sup>5</sup>	1.7	84.8	-100.9	-30.5
X-ray <sup>s</sup>	2rn2	a/b	155/76	3.1	72.9	-67.5	-24.0
X-ray <sup>s</sup>	5pnt	a/b	157/134	3.0	71.6	-34.4	-3.0
X-ray <sup>s</sup>	1s0p	A	160/116	4.3	70.8	-19.9	-10.4

Protein Name <sup>1</sup>	Native PDB ID	Topology	Numbr of residues/ Number of residues converged in computed structure	Median RMSD to native over converged region <sup>2</sup> (Å)	Median GDT-TS among lowest energy models <sup>3</sup>	Depth of converged ensemble energy minimum <sup>7</sup>	Median energy change resulting from inclusion of experimental data <sup>8</sup>
ARF1 <sup>r</sup>	2k5u	a/b	166/122	2.5	77.2	-28.6	-8.7
X-ray <sup>s</sup>	2z2i	a/b	179/143	1.8	77.7	-46.5	-21.6
ALG13 <sup>r</sup>	2jzc	a/b	201/156 <sup>6</sup>	3.4	63.7	-77.8	-12.8
X-ray <sup>s</sup>	1sua	a/b	263/173	6.2	57.0	-43.5	-26.5
X-ray <sup>s</sup>	1g68	a/b	266/119	3.2	<b>41.4</b>	-36.3	-25.1

<sup>r</sup> r – real experimental data was used

<sup>s</sup> s – partially or fully synthetic data was used. See Supplementary Table S2 for more details.

<sup>b</sup> Blind test case

<sup>1</sup> NESG codes are used for protein structures obtained with conventional NMR methods in the NESG, and PDB codes for the remaining proteins. The results shown in the top eleven rows were generated with the CS-RDC-Rosetta protocol and the remaining with the iterative CS-RDC-Rosetta protocol.

<sup>2</sup> For the iterative protocol, residues were considered converged if they are members of the largest set of residues that is superimposable within 4 Å. For the non-iterative protocol the residues were selected with the FindCore(25) algorithm <http://fps.nesg.org/> based on the conventional NMR ensemble

<sup>3</sup> The GDT-TS (Global Distance Test – Total Score) is the average number of C-alphas superimposable within 1,2,3,4 and 7Å, respectively(26) Shown is the median of GDTTS scores computed for each pair of structures in the lowest energy 10.

<sup>4</sup> For this structure calculation all pairs of H<sup>N</sup> protons within 5 Å generated an H<sup>N</sup>-H<sup>N</sup> NOE distance constraint of 6 Å (cf. Methods)

<sup>5</sup> For this ensemble all residues converged within 4 Å with a median RMSD to the native of 2.3 Å and 3.5 Å for 1i1b\_2 and 1i1b, respectively.

<sup>6</sup> Converged residues within 3 Å due to the high flexibility in the reference NMR structure. ( with cutoff of 4 Å 176 residues converge and yield a median RMSD of 4.9 )

<sup>7</sup> Energy difference between the median energy of the 10 lowest energy models and 10 lowest energy models which differ by at least 7Å RMSD

\* (\* 4Å RMSD) from lowest energy model. Values are in Rosetta energy units ( ~0.5kcal/mol ).

<sup>8</sup> Energy difference between the median energy of the 10 lowest energy models obtained with RDC and/or NOE data and the median energy of the 10 lowest energy CS-Rosetta models. Values are in Rosetta energy units (~0.5 kcal/mol).