NMR Structure Determination for Larger Proteins Using Backbone-Only Data

Srivatsan Raman, *et al.*

Science **327**, 1014 (2010);
DOI: 10.1126/science.1183649

This copy is for your personal, non-commercial use only.

If you wish to distribute this article to others, you can order high-quality copies for your colleagues, clients, or customers by clicking here.

Permission to republish or repurpose articles or portions of articles can be obtained by following the guidelines here.

The following resources related to this article are available online at www.sciencemag.org (this information is current as of March 1, 2010):

Updated information and services, including high-resolution figures, can be found in the online version of this article at:
http://www.sciencemag.org/cgi/content/full/327/5968/1014

Supporting Online Material can be found at:
http://www.sciencemag.org/cgi/content/full/science.1183649/DC1

This article **cites 26 articles**, 6 of which can be accessed for free:
http://www.sciencemag.org/cgi/content/full/327/5968/1014#otherarticles

This article appears in the following **subject collections**:
Biochemistry
http://www.sciencemag.org/cgi/collection/biochem

Science (print ISSN 0036-8075; online ISSN 1095-9203) is published weekly, except the last week in December, by the American Association for the Advancement of Science, 1200 New York Avenue NW, Washington, DC 20005. Copyright 2010 by the American Association for the Advancement of Science; all rights reserved. The title *Science* is a registered trademark of AAAS.

NMR Structure Determination for Larger Proteins Using Backbone-Only Data

Srivatsan Raman,[1]*† Oliver F. Lange,[1]* Paolo Rossi,[2] Michael Tyka,[1] Xu Wang,[3] James Aramini,[2] Gaohua Liu,[2] Theresa A. Ramelot,[4] Alexander Eletsky,[5] Thomas Szyperski,[5] Michael A. Kennedy,[4] James Prestegard,[3] Gaetano T. Montelione,[2] David Baker[1,6]‡

Conventional protein structure determination from nuclear magnetic resonance data relies heavily on side-chain proton-to-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 nuclear magnetic resonance (NMR) information on the side chains for proteins up to 25 kilodaltons by incorporating backbone chemical shifts, residual dipolar couplings, and amide proton distances into the Rosetta protein structure modeling methodology. These data, which are too sparse for conventional methods, serve only to guide conformational search toward 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 kilodaltons and should enable routine NMR structure determination for larger proteins.

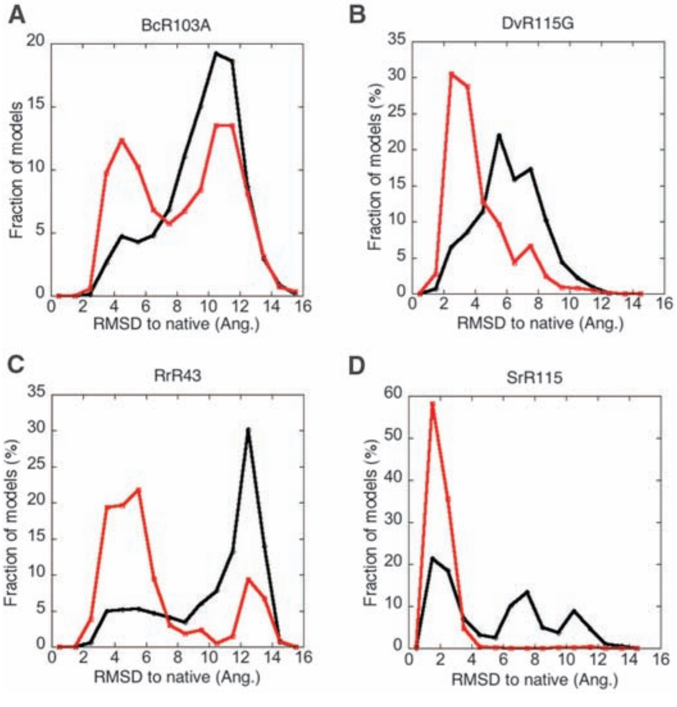
The first step in protein structure determination by nuclear magnetic resonance (NMR) is chemical-shift assignment for the backbone atoms. In contrast to the subsequent assignment of the side chains, this process is now rapid, reliable, and largely automated (*1*–*5*). Global backbone structural information complementing the local structure information provided by backbone chemical-shift assignments (*6*, *7*) can be obtained from H^N-H^N nuclear Overhauser effect spectroscopy (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 side chains, except for selectively protonated side-chain moieties (*13*). The difficulty in determining accurate structures with no or limited side-chain information is a major bottleneck that currently prevents routine application of NMR to larger (>15 kD) systems (*14*).

Here we show that structures of proteins up to 200 residues (23 kD) can be determined with the use of information from backbone (H^N, N, C^α, C^β, C′) NMR data by taking advantage of the conformational sampling and all-atom energy function in the Rosetta structure prediction methodology (*15*), which, for small proteins in favorable cases, can produce atomic accuracy models starting from sequence information alone (*16*, *17*). Structure prediction in Rosetta proceeds in two steps: (i) a low-resolution exploration phase using Monte Carlo fragment assembly and a coarse-grained energy function, and (ii) a computationally expensive refinement phase that cycles between combinatorial side-chain optimization and gradient-based minimization of all torsional degrees of freedom in a physically realistic all-atom forcefield (*16*). 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 they are very seldom sampled in unbiased trajectories. Incorporating NMR chemical-shift information in the selection of the fragments used in the exploration phase [chemical shift (CS)–Rosetta] (*18*, *19*) 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-kD) proteins, the performance of CS-Rosetta is very target-dependent: Structures sufficiently close to the native structure for the energy to drop substantially 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 to 200 residue) proteins. For every attempted Monte Carlo move, the alignment tensor is calculated by singular value decomposition (*20*), and the decision to accept or reject the conformation is biased by the change in the agreement between the back-calculated and experimental couplings (*21*). Incorporation of RDCs dramatically improved convergence on the correct structure in a benchmark of 11 α, β, and α/β proteins ranging in size from 62 to 166 residues (Fig. 1, Table 1, and fig. S1). As indicated in Table 1, CS-RDC-Rosetta consistently generates accurate models for proteins up to 120 residues and, in favorable cases, for larger proteins.

For proteins with more than 120 residues, conformational sampling becomes limiting, even for the CS-RDC-Rosetta protocol, and the low-

Fig. 1. Impact of RDC data on conformational search. Lines depict RMSD histograms of structures selected in the lowest 10th percentile of coarse-grained energy for ensembles generated with the use of CS-Rosetta (black) or CS-RDC-Rosetta (red). (**A**) BcR103A, (**B**) DvR115G, (**C**) RrR43, and (**D**) SrR115C. Ang., angstrom.



---

[1] Department of Biochemistry, University of Washington, Seattle, WA 98195, USA. [2] Department of Molecular Biology and Biochemistry, Center for Advanced Biotechnology and Medicine, and Northeast Structural Genomics Consortium, Rutgers University, Piscataway, NJ 08854, USA. [3] Complex Carbohydrate Research Center, University of Georgia, Athens, GA 30602, USA. [4] Department of Chemistry and Biochemistry and Northeast Structural Genomics Consortium, Miami University, Oxford, OH 45056, USA. [5] Department of Chemistry, State University of New York at Buffalo, Buffalo, NY 14260, USA. [6] Howard Hughes Medical Institute (HHMI), Seattle, WA 98195, USA.

*These authors contributed equally to this work.

†Present address: Department of Genetics, Harvard Medical School, Boston, MA 02115, USA.

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

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 nuclear Overhauser effects (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 (*22*). The new protocol, a genetic algorithm, generates hybrid conformations by recombining first β-sheet pairings and, subsequently, fragments of the low-energy structures (*17*). To further enhance sampling, trajectories are seeded with conformations harvested from previous trajectories that led to low-energy conformations (*23*).

The improvement in the model population with increasing generations in the iterative protocol is illustrated in Fig. 2 for the 200-residue ALG13 protein using experimentally determined chemical shift, RDC, and assigned backbone amide H^N-H^N NOE data (*24*). The C^α root mean square deviation (RMSD) to the native structure and the energy improve from generation to generation, and after several rounds, discrimination toward lower RMSD structures is apparent (Fig. 2A, light blue to yellow). After high-resolution refinement (Fig. 2A, orange to red), the lowest-energy structures are close to the native structure. The final low-energy structural ensemble (Fig. 2B) recapitulates the unusual topology in the previously determined NMR structure (*24*) (Fig. 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 (Fig. 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 (*24*).

The iterative CS-RDC-NOE protocol was tested further on 12 proteins ranging in size from 120 to 266 residues (Table 1 and fig. S3). For all proteins except 1g68, a considerable part of the structure converges (Table 1). Backbone H^N-H^N NOE data were 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 to 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). Side-chain accuracy was generally quite high in the converged regions (fig. S5).

Table 1. Accuracy of models generated with backbone-only NMR data.

Protein name*	Native PDB ID	Topology	Number of residues/ number of residues converged in computed structure	Median RMSD to native over converged region† (Å)	Median GDT-TS among lowest-energy models‡	Depth of converged ensemble energy minimum§	Median energy change resulting from inclusion of experimental data‖
				Noniterative			
GmR137	2k5p	a/b	62/47	2.6	95.4	−32.5¶	−1.8
TR80	2jxt	a/b	78/73	1.5	84.9	−16.7¶	−0.3
DvR115G#	2kct	B	86/66	1.4	80.0	−24.3	−0.7
LkR15	2k3d	a/b	92/74	2.0	85.4	−18.0¶	−1.2
BcR103A	2kd1	B	100/65	3.4	61.3	−22.7	−1.3
SrR115C#	2kcl	A	100/95	1.4	86.1	−25.1¶	0.7§§
MaR214A#	2kbn	B	102/96	2.1	82.1	−43.9	−0.6
RrR43	2k0m	a/b	104/82	2.1	66.8	−12.9	2.9§§
BcR268F#	2k5w	A	118/115	1.4	78.4	−45.7¶	−1.6
ER553	2k1s	a/b	143/115	5.2	46.1§§	−5.1	−2.5
ARF1	2k5u	a/b	166/141	2.6	73.3	−21.6	−9.5
				Iterative			
AtT7#	2ki8	a/b	122/98	3.0	70.0	−37.5	−12.5
ER541**	2jyx	a/b	124/115	2.5	76.7	−31.6	−9.9
X-ray**	1f21	a/b	142/122	9.4	76.6	−28.5	3.5§§
ER553	2k1s	a/b	143/136	1.9	85.2	−38.5	−15.4
BtR324B**	2kd7	B	150/148	2.4	79.3	−51.6	−29.1
X-ray**	1i1b	B	151/111‡‡	2.5	71.1	−53.3	−25.5
X-ray**	1i1b_2††	B	151/133‡‡	1.7	84.8	−100.9	−30.5
X-ray**	2rn2	a/b	155/76	3.1	72.9	−67.5	−24.0
X-ray**	5pnt	a/b	157/134	3.0	71.6	−34.4	−3.0
X-ray**	1s0p	A	160/116	4.3	70.8	−19.9	−10.4
ARF1	2k5u	a/b	166/122	2.5	77.2	−28.6	−8.7
X-ray**	2z2i	a/b	179/143	1.8	77.7	−46.5	−21.6
ALG13	2jzc	a/b	201/155‡‡	3.4	63.7	−77.8	−12.8
X-ray**	1sua	a/b	263/173	6.2	57.0	−43.5	−26.5
X-ray**	1g68	a/b	266/119	3.2	41.4§§	−36.3	−25.1

*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 11 rows were generated with the CS-RDC-Rosetta protocol and the remaining with the iterative CS-RDC-Rosetta protocol. †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 noniterative protocol, the residues were selected with the FindCore algorithm (*26*, *27*) based on the conventional NMR ensemble. ‡The global distance test–total score (GDT-TS) is the average number of Cα superimposable within 1, 2, 3, 4, and 7 Å, respectively (*28*). Shown is the median of GDT-TS scores computed for each pair of structures out of the 10 lowest-energy models. §Energy difference between the median energy of the 10 lowest-energy models and the 10 lowest-energy models that differ by at least 7 Å RMSD from the lowest-energy model (Rosetta energy units). ‖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 (Rosetta energy units). ¶Energy gap computed with 4 Å cutoff radius (instead of 7 Å). #Blind test case. **Partially or fully synthetic data were used (see table S2). ††All pairs of H^N protons within 5 Å generated an H^N-H^N NOE distance constraint of 6 Å (*17*). ‡‡Results are shown with a reduced convergence cutoff of 3 Å (with cutoff 4 Å, 151, 151, and 176 residues converge and yield a median RMSD of 3.5, 2.3, and 4.9 Å for 1i1b, 1i1b_2, and 2jzc, respectively). §§Violation of validation criterion.

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 (Fig. 3, A to D), whereas for a fifth, convergence was not observed, and blind structure determination was instead carried out with the iterative CS-RDC-NOE protocol (Fig. 3E). In

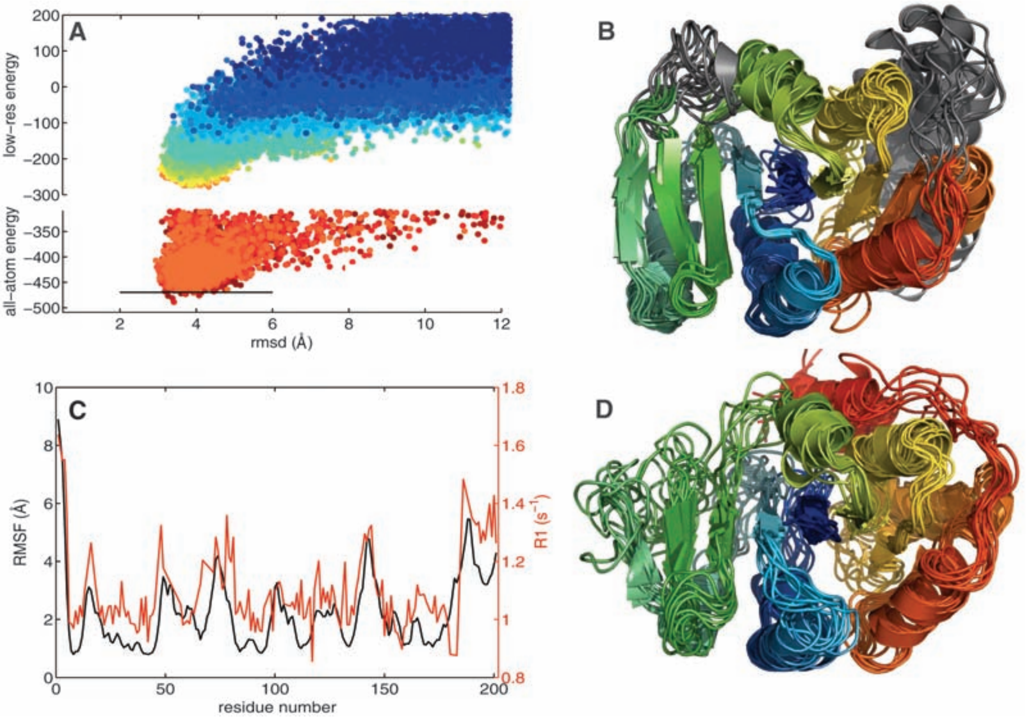


Fig. 2. Determination of ALG13 structure from backbone NMR data with Rosetta. (**A**) RMSDs and energies of structures generated in batches of 2000 during the iterative protocol. Each generation of structures (color code: blue to red, corresponds to number of generation) is based on information from previous runs (*17*). Strong convergence is reached already in the computational 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 Å root mean square fluctuations (RMSF) was reached in the 50 lowest-energy Rosetta models (residues 5 to 70, 81 to 139, 151 to 180). (**B**) Ensemble of 10 lowest-energy Rosetta structures [below line in (A)]. Regions with more than 3 Å RMSF are depicted in gray. (**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). The relaxation data were not used in the structure calculation. 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 (PDB ID: 2jzc).

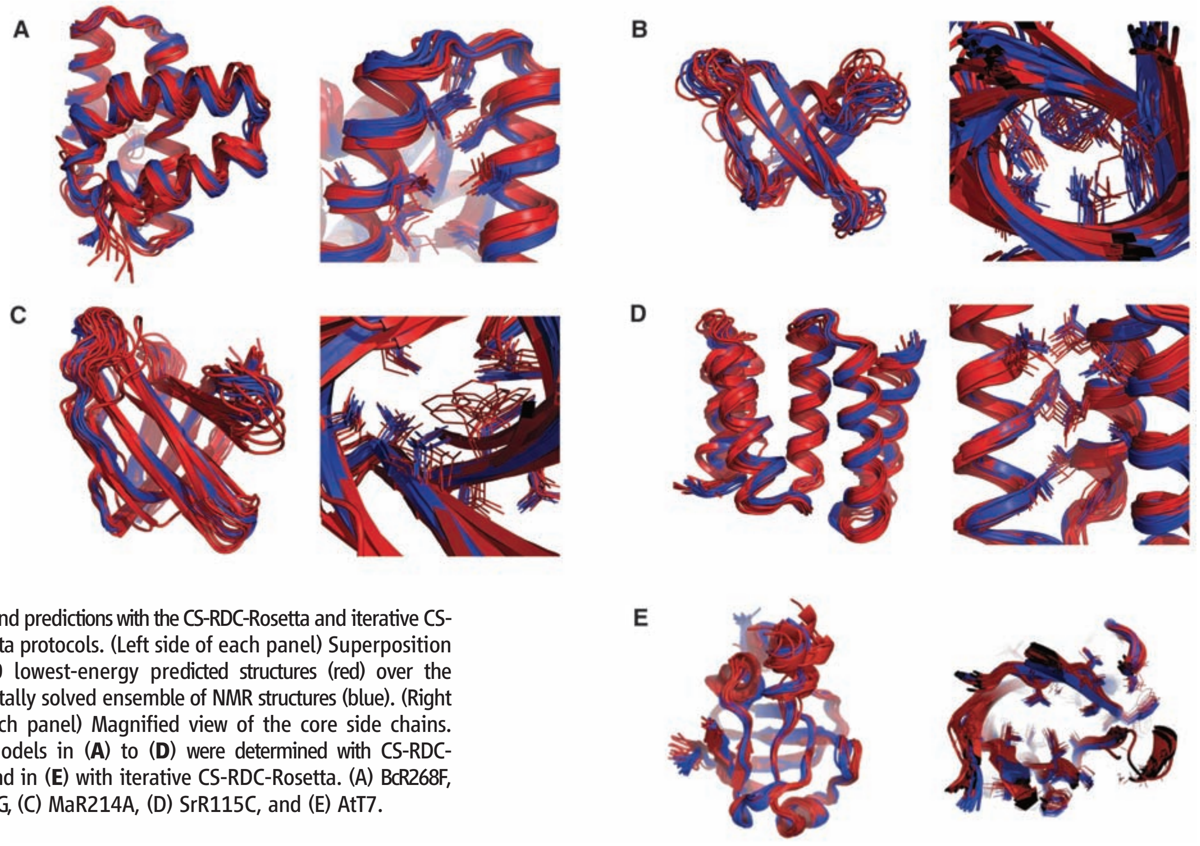


Fig. 3. Blind predictions with the CS-RDC-Rosetta and iterative CS-RDC-Rosetta protocols. (Left side of each panel) Superposition of the 10 lowest-energy predicted structures (red) over the experimentally solved ensemble of NMR structures (blue). (Right side of each panel) Magnified view of the core side chains. Rosetta models in (**A**) to (**D**) were determined with CS-RDC-Rosetta, and in (**E**) with iterative CS-RDC-Rosetta. (A) BcR268F, (B) DvR115G, (C) MaR214A, (D) SrR115C, and (E) AtT7.

all five cases (Table 1), the resulting Rosetta-determined structure is very similar to the conventionally determined NMR solution structure over both the backbone (Fig. 3, left side of each panel) and the core side chains (Fig. 3, right side of each panel), which is notable because no experimental side-chain information is used in the Rosetta protocol; the details of core packing are determined by the Rosetta all-atom energy function (magnified views of core side chains are shown for all of the remaining targets in fig. S6).

Thus, our methodology is able to generate accurate structures of proteins up to ~25 kD 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 are 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: (i) the detailed physical chemistry implicit in the Rosetta all-atom energy function and (ii) the experimental NMR data. As illustrated in Fig. 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), though smoother, is degenerate and lacks resolution. Because 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 Fig. 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 (Fig. 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 toward the global minimum (Fig. 4, A and B, arrow 2).

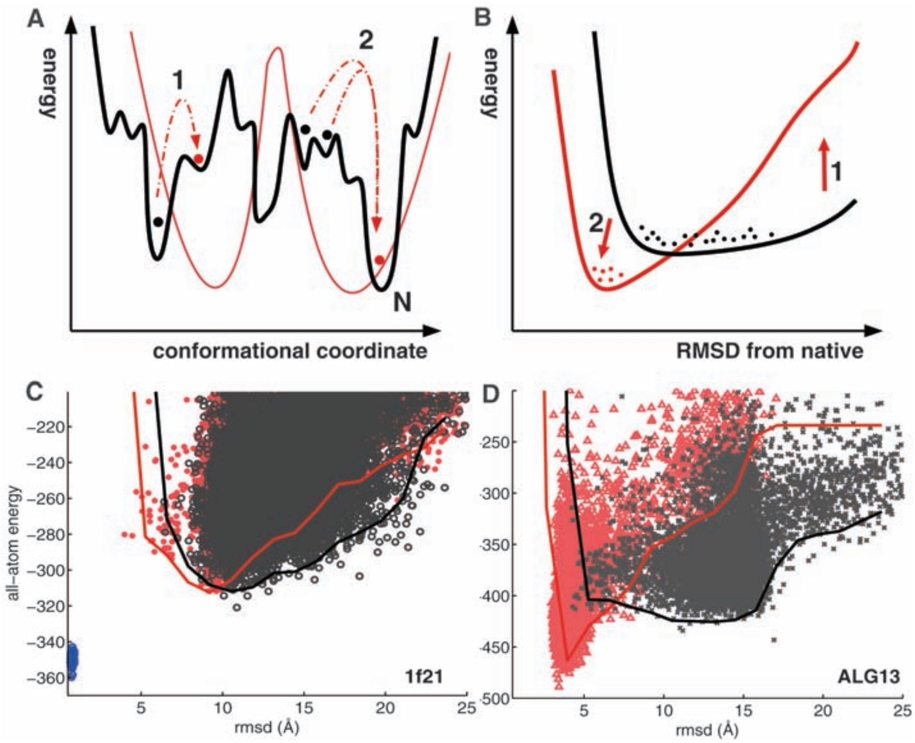


Fig. 4. Effect of incorporation of experimental data on energy minimization. (**A**) The Rosetta all-atom energy (black line) has many local minima, making minimization difficult, but the global minimum is generally close to the native structure (N). The experimental bias (red line), though smoother, 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), whereas 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, the experimental data can guide sampling toward the global minimum close to the native structure (arrow 2), and thus, constrained optimization can result in lower-energy conformations than unconstrained optimization, whereas biased optimization is less effective than unconstrained optimization distant from the native structure leading to higher energies (arrow 1). (**C** and **D**) In contrast, all-atom energy and RMSD of final Rosetta ensemble from iterative refinement, with and without experimental data, are shown. 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).

Better optimization in the presence of experimental data (Fig. 4B) is unlikely to occur if there is no sampling near the correct structure, as 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, columns 6 to 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 clearly be lower in energy than all significantly different (RMSD > 7 Å) structures; this was true 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 residues), the data in fact guide the trajectories to lower-energy structures than those obtained by unconstrained optimization (Fig. 4D and Table 1, column 8). As argued above, this is a strong indicator that the correct structure has been found.

When all three criteria were satisfied for the 20 proteins in our test set, 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 Å to the native, fails the third criterion: The energy is higher rather than lower when the experimental data are included in the optimization (Fig. 4C and Table 1, column 8). Because 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) (*17*). For almost all of 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-kD molecular-weight range, the computed structures are not completely converged and have large disordered regions. This is clearly a sampling problem because the native structure has lower energy (Fig. 4C and fig. 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 using additional experimental data (1ilb_2 in fig. S3) or by improved sampling. Though at present the former is the more reliable solution, the latter will probably 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 (fig. S4). In the new approach, the bottlenecks of side-chain 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 from three- or four-dimensional experiments, which restrict possible β-strand registers. Advantages of the approach are that ^{1}H,^{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 kD 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 (*25*) for which the available data are generally quite sparse.

References and Notes

1. D. E. Zimmerman *et al.*, *J. Mol. Biol.* **269**, 592 (1997).
2. C. Bartels, P. Güntert, M. Billeter, K. Wüthrich, *J. Comput. Chem.* **18**, 139 (1998).
3. M. C. Baran, Y. J. Huang, H. N. Moseley, G. T. Montelione, *Chem. Rev.* **104**, 3541 (2004).
4. Y. S. Jung, M. Zweckstetter, *J. Biomol. NMR* **30**, 11 (2004).
5. W. Lee, W. M. Westler, A. Bahrami, H. R. Eghbalnia, J. L. Markley, *Bioinformatics* **25**, 2085 (2009).
6. M. Berjanskii *et al.*, *Nucleic Acids Res.* **37** (Web Server issue), W670 (2009).
7. Y. Shen, F. Delaglio, G. Cornilescu, A. Bax, *J. Biomol. NMR* **44**, 213 (2009).
8. G. Kontaxis, F. Delaglio, A. Bax, *Methods Enzymol.* **394**, 42 (2005).
9. I. Bertini, C. Luchinat, G. Parigi, R. Pierattelli, *Dalton Trans.* **29**, 3782 (2008).
10. J. H. Prestegard, C. M. Bougault, A. I. Kishore, *Chem. Rev.* **104**, 3519 (2004).
11. S. Grzesiek, A. Bax, *J. Biomol. NMR* **3**, 627 (1993).
12. D. M. LeMaster, F. M. Richards, *Biochemistry* **27**, 142 (1988).
13. K. H. Gardner, M. K. Rosen, L. E. Kay, *Biochemistry* **36**, 1389 (1997).
14. G. Wagner, *J. Biomol. NMR* **3**, 375 (1993).
15. R. Das, D. Baker, *Ann. Rev. Biochem.* **77**, 363 (2008).
16. P. Bradley, K. M. S. Misura, D. Baker, *Science* **309**, 1868 (2005).
17. Materials and methods are available as supporting material on *Science* Online.
18. A. Cavalli, X. Salvatella, C. M. Dobson, M. Vendruscolo, *Proc. Natl. Acad. Sci. U.S.A.* **104**, 9615 (2007).
19. Y. Shen *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **105**, 4685 (2008).
20. J. A. Losonczi, M. Andrec, M. W. Fischer, J. H. Prestegard, *J. Magn. Reson.* **138**, 334 (1999).
21. C. A. Rohl, D. Baker, *J. Am. Chem. Soc.* **124**, 2723 (2002).
22. B. Qian *et al.*, *Nature* **450**, 259 (2007).
23. T. J. Brunette, O. Brock, *Proteins* **73**, 958 (2008).
24. X. Wang, T. Weldeghiorghis, G. Zhang, B. Imperiali, J. H. Prestegard, *Structure* **16**, 965 (2008).
25. H. van Ingen, D. M. Korzhnev, L. E. Kay, *J. Phys. Chem. B* **113**, 9968 (2009).
26. D. A. Snyder, G. T. Montelione, *Proteins* **59**, 673 (2005).
27. The FindCore algorithm is available at http://fps.nesg.org.
28. A. Zemla, *Nucleic Acids Res.* **31**, 3370 (2003).
29. We are thankful to the U.S. Department of Energy Innovative and Novel Computational Impact on Theory and Experiment 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 Y. Shen and A. Bax for fruitful discussions; Y. J. Huang and Y. Tang for their contribution during preliminary studies using sparse NOE constraints with CS-Rosetta; S. Bansal, H.-w. Lee, and Y. Liu for collection of RDC data, A. Lemak for providing the Crystallography and NMR System 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; M.T. developed the all-atom refinement protocol; S.R., O.F.L. and D.B. designed and performed research for energy based structure validation; X.W and J.P analyzed the ALG13 ensemble; J.A, G.L, T.R, A.E, M.K, and T.S provided blind NMR data sets; and 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 (O.F.L.), NIH grant GM76222 (D.B.), the HHMI, the National Institutes of General Medical Science Protein Structure Initiative program grant U54 GM074958 (G.T.M.), and the Research Resource grant RR005351 (J.P.). M.T. holds a Sir Henry Wellcome Postdoctoral Fellowship. RDC and Paramagnetic Relaxation Enhancement data as deposited in the Protein Data Bank (PDB) with accession code 2jzc.

Supporting Online Material

www.sciencemag.org/cgi/content/full/science.1183649/DC1
Materials and Methods
SOM Text
Figs. S1 to S5
Tables S1 to S3
References

21 October 2009; accepted 14 January 2010
Published online 4 February 2010;
10.1126/science.1183649
Include this information when citing this paper.

Limits of Predictability in Human Mobility

Chaoming Song,[1,2] Zehui Qu,[1,2,3] Nicholas Blumm,[1,2] Albert-László Barabási[1,2]*

A range of applications, from predicting the spread of human and electronic viruses to city planning and resource management in mobile communications, depend on our ability to foresee the whereabouts and mobility of individuals, raising a fundamental question: To what degree is human behavior predictable? Here we explore the limits of predictability in human dynamics by studying the mobility patterns of anonymized mobile phone users. By measuring the entropy of each individual's trajectory, we find a 93% potential predictability in user mobility across the whole user base. Despite the significant differences in the travel patterns, we find a remarkable lack of variability in predictability, which is largely independent of the distance users cover on a regular basis.

When it comes to the emerging field of human dynamics, there is a fundamental gap between our intuition and the current modeling paradigms. Indeed, although we rarely perceive any of our actions to be random, from the perspective of an outside observer who is unaware of our motivations and schedule, our activity pattern can easily appear random and unpredictable. Therefore, current models of human activity are fundamentally stochastic (*1*) from Erlang's formula (*2*) used in telephony to Lévy-walk models describing human mobility (*3*–*7*) and their applications in viral dynamics (*8*–*10*), queuing models capturing human communication patterns (*11*–*13*), and models capturing body balancing (*14*) or panic (*15*). Yet the probabilistic nature of the existing modeling framework raises fundamental questions: What is the role of randomness in human behavior and to what degree are individual human actions predictable? Our goal here is to quantify

---

[1] Center for Complex Network Research, Departments of Physics, Biology, and Computer Science, Northeastern University, Boston, MA 02115, USA. [2] Department of Medicine, Harvard Medical School, and Center for Cancer Systems Biology, Dana-Farber Cancer Institute, Boston, MA 02115, USA. [3] School of Computer Science and Engineering, University of Electric Science and Technology of China, Chengdu 610054, China.

*To whom correspondence should be addressed. E-mail: alb@neu.edu