

center of its circular swimming path. These particles move randomly with an apparent diffusion coefficient of  $D = 9.0 \pm 2.0 \mu\text{m}^2/\text{s}$ , measured for isolated spermatozoa. A short-range pairwise attraction, arising from the hydrodynamic forces leading to the observed synchronization (20), and a longer range repulsion, which could be of steric or hydrodynamic origin (21), are assumed (Fig. 4D). Although one cannot describe circular flow by a potential (22), the important features of the observed pattern are captured by our model.

Stochastic simulations of this model (SOM text) also revealed two regimes: a random distribution of particles at low densities with a transition toward a hexagonal array of clusters at a critical particle density (Fig. 4E). Assigning to each particle a spermatozoon circling around that position, we generated simulated movies (3) mimicking the experimental observation (Fig. 4F versus Fig. 1B). Moreover, the order parameter  $\chi$  computed for different simulated sperm densities agreed with the experimentally observed dependency (Fig. 4C). Our numerical results were further supported by a 1D mean-field analysis (SOM text), which indicated the existence of a supercritical pitchfork bifurcation at a critical sperm density (23). This critical density was proportional to the interaction strength and inversely proportional to the diffusion coefficient, the latter being associated with the noise in the system. This analysis demonstrates how the activity of biological processes can be regulated by critical points or bifurcations. For example, ciliary

metachronal waves (16, 24) might be switched on and off by small physiologically controlled changes of the activity of the individual cilia, thereby tuning the critical density for the onset of the metachronal wave.

The only free parameter in our model was the ratio of the maximum interaction potential to the drag coefficient,  $V_0/\gamma = 5 \mu\text{m}^2/\text{s}$ , which was chosen to match the critical density (Fig. 4C). This allowed us to estimate the interaction force between two spermatozoa  $F_{\text{int}} = |\text{grad}(V)| = (V_0/\gamma) \times \gamma/R \sim 0.03 \text{ pN}$  (using  $R = 13 \mu\text{m}$  and  $\gamma = 0.07 \mu\text{N}\cdot\text{s}/\text{m}$  from above). This force is about 1% of the forward propulsion force of spermatozoa  $F_{\text{for}} \sim 5 \text{ pN}$  (25). Although this hydrodynamic interaction force is smaller than typical adhesion forces involved in sperm cooperation (26), it is evidently large enough to coordinate the cells and to regulate large-scale pattern formation in the absence of chemical signals (27).

#### References and Notes

1. I. R. Gibbons, *J. Cell Biol.* **91**, 107s (1981).
2. D. M. Woolley, *Reproduction* **126**, 259 (2003).
3. Materials and methods are available as supporting material on Science Online.
4. H. C. Berg, *Random Walks in Biology* (Princeton Univ. Press, Princeton, NJ, 1993).
5. X. L. Wu, A. Libchaber, *Phys. Rev. Lett.* **84**, 3017 (2000).
6. N. Darnton, L. Turner, K. Breuer, H. C. Berg, *Biophys. J.* **86**, 1863 (2004).
7. A. M. Turing, *Philos. Trans. R. Soc. London Ser. B* **237**, 37 (1952).
8. I. Prigogine, G. Nicolis, *J. Chem. Phys.* **46**, 3542 (1967).
9. T. Misteli, *J. Cell Biol.* **155**, 181 (2001).
10. F. J. Nédélec, T. Surrey, A. C. Maggs, S. Leibler, *Nature* **389**, 305 (1997).
11. K. Zahn, G. Maret, C. Russ, H. H. von Grünberg, *Phys. Rev. Lett.* **91**, 115502 (2003).
12. K. Zahn, R. Lenke, G. Maret, *Phys. Rev. Lett.* **82**, 2721 (1999).
13. J. Gray, *Ciliary Movement* (Cambridge Univ. Press, New York, 1928).
14. J. Gray, G. J. Hancock, *J. Exp. Biol.* **32**, 802 (1955).
15. G. I. Taylor, *Proc. R. Soc. London Ser. A* **209**, 447 (1951).
16. K. I. Okamoto, Y. Nakaoka, *J. Exp. Biol.* **192**, 61 (1994).
17. S. Gueron, K. Levit-Gurevich, *Biophys. J.* **74**, 1658 (1998).
18. E. Nielsen, F. Severin, J. M. Backer, A. A. Hyman, M. Zerial, *Nat. Cell Biol.* **1**, 376 (1999).
19. B. Hoellndobler, E. O. Wilson, *The Ants* (Springer, Berlin, 1990).
20. L. J. Fauci, A. McDonald, *Bull. Math. Biol.* **57**, 679 (1995).
21. P. Lenz, J. F. Joanny, F. Jülicher, J. Prost, *Phys. Rev. Lett.* **91**, 108104 (2003).
22. L. D. Landau, E. M. Lifshitz, *Fluid Mechanics, Course of Theoretical Physics* (Pergamon Press, Oxford, 1987).
23. S. H. Strogatz, *Nonlinear Dynamics and Chaos* (Westview Press, Cambridge, MA, 2000).
24. M. A. Sleight, Ed., *Cilia and Flagella* (Academic Press, London, 1974).
25. J. Howard, *Mechanics of Motor Proteins and the Cytoskeleton* (Sinauer Associates, Sunderland, MA, 2001).
26. H. Moore, K. Dvorakova, N. Jenkins, W. Breed, *Nature* **418**, 174 (2002).
27. C. Dombrowski, L. Cisneros, S. Chatkaew, R. E. Goldstein, J. O. Kessler, *Phys. Rev. Lett.* **93**, 098103 (2004).
28. We thank D. Babcock, C. Brokaw, R. Goldstein, F. Jülicher, H. Machemer, K. Müller, F. Nédélec, E. Schäffer, and members of the Howard lab and Jülicher lab for discussions and comments on the manuscript. All authors contributed ideas and discussion, and I.H.R. carried out experiments, programming, and data analysis.

#### Supporting Online Material

[www.sciencemag.org/cgi/content/full/309/5732/300/DC1](http://www.sciencemag.org/cgi/content/full/309/5732/300/DC1)

Materials and Methods  
SOM Text  
Figs. S1 to S6  
References  
Movies S1 to S5

27 January 2005; accepted 24 May 2005  
10.1126/science.1110329

## Inferential Structure Determination

Wolfgang Rieping,\* Michael Habeck,\* Michael Nilges†

Macromolecular structures calculated from nuclear magnetic resonance data are not fully determined by experimental data but depend on subjective choices in data treatment and parameter settings. This makes it difficult to objectively judge the precision of the structures. We used Bayesian inference to derive a probability distribution that represents the unknown structure and its precision. This probability distribution also determines additional unknowns, such as theory parameters, that previously had to be chosen empirically. We implemented this approach by using Markov chain Monte Carlo techniques. Our method provides an objective figure of merit and improves structural quality.

A major difficulty in the determination of three-dimensional macromolecular structures is that experimental data are indirect. We observe

physical effects that depend on the atomic geometry and use a forward model to relate the observed data to the atomic coordinates. For example in nuclear magnetic resonance (NMR), the intensity  $I_i$  of peaks in nuclear Overhauser effect spectroscopy (NOESY) data is proportional to the inverse sixth power of the distance  $d_i$  of two spins:  $I_i = \gamma d_i^{-6}$  (1). This isolated spin pair approximation (ISPA) involves an unknown scaling factor  $\gamma$ . It seems

straightforward to obtain the structure in the example: simply use the observed intensities to calculate sufficient distances to define the structure.

In realistic applications, this approach runs into difficulties. One problem is that the forward model is usually inherently degenerate, meaning that different conformations can lead to the same observations and therefore cannot be distinguished experimentally, and even a formally invertible forward model is practically degenerate if the data are incomplete. A further complication is that there are uncertainties in both the data and the forward model: Data are subject to experimental errors, and theories rest on approximations. Moreover, the forward model typically involves parameters that are not measurable. Algorithms for structure calculation from x-ray reflections, NMR spectra, or homology-derived restraints should account for these fundamental difficulties in some way.

Structure determination in general is an ill-posed inverse problem, meaning that going from the data to a unique structure is impossible. However, the current paradigm in structure calculation is to attempt an inversion of the forward model. Most algorithms minimize a hybrid energy  $E_{\text{hybrid}} = E_{\text{phys}} +$

Unité de Bioinformatique Structurale, Institut Pasteur, CNRS URA 2185, 25-28 rue du Docteur Roux, 75724 Paris CEDEX 15, France.

\*These authors contributed equally to this work.

†To whom correspondence should be addressed. E-mail: nilges@pasteur.fr

$w_{\text{data}} E_{\text{data}}$  (2), where a nonphysical energy  $E_{\text{data}}$  uses the forward model and a restraining function to assess the agreement between data and structure. A force field  $E_{\text{phys}}$  describes the physical properties of the macromolecule, such as bonded and nonbonded interactions between the atoms, and partially removes the degeneracy of the problem. The rationale is that minimization of the hybrid energy effectively inverts the forward model, yielding the “true” structure.

This strategy works in the case of many data of good quality. In less favorable situations, the ill-posed nature of the inverse problem becomes apparent. Specifically, it remains unclear how to choose auxiliary parameters like the weight  $w_{\text{data}}$  or theory parameters such as the scaling factor  $\gamma$  in the ISPA. Because the hybrid energy minimization paradigm offers no principle to settle these issues, such parameters need to be determined heuristically.

The principal difficulty in structure determination by NMR is the lack of information that is indispensable to reconstruct the structure unambiguously. By formulating an optimization problem (“search for the minimum of  $E_{\text{hybrid}}$ ”), one however implicitly assumes that there is a unique answer. Repeating the optimization procedure multiple times to obtain several “unique” solutions hides but does not solve the ambiguity and makes it difficult to judge the validity and precision of NMR structures in an objective way.

We suggest that it is a misconception to use structure calculation methods that are only appropriate if the objective is to obtain a unique structure. Instead, we view structure determination as an inference problem, requiring reasoning from incomplete and uncertain information. We consider the entire conformational space and use the data only to rank the molecule’s possible conformations. We assign a number  $P_i$  to every conformation  $X_i$ . If  $P_i > P_j$ , conformation  $X_i$  is more supported by the data than  $X_j$ . Cox (3) proved that such rankings are equivalent to a probability and that probability theory is the only consistent calculus to solve inference problems. The distribution of the probabilities  $P_i$  reflects the information content of the data. If all but one  $P_i$  vanish, the data determine the structure uniquely. If  $P_i$  are uniform throughout conformational space, the data are completely uninformative with respect to the structure.

Any inferential structure determination (ISD) is solved by calculating the probabilities  $P_i$ . We demand the probabilities to be objective in the sense that they only depend on data  $D$  and on relevant prior information  $I$  (such as the forward model or knowledge about physical interactions). Thus,  $P_i$  is a conditional probability,  $P_i = P(X_i|D, I)$ ; it is not a frequency of occurrence but a quantitative representation of our state of knowledge. In the case of a continuous parametrization of con-

formations, such as Cartesian coordinates,  $P_i$  is a density  $p(X|D, I)$ . A direct consequence of probability calculus is Bayes’ theorem (4), which formally solves our inference problem. The posterior density

$$p(X|D, I) \propto p(D|X, I) p(X|I) \quad (1)$$

factorizes into two natural components: The likelihood function  $p(D|X, I)$  combines a forward model and an error distribution and quantifies the likelihood of observing data  $D$  given a molecular structure  $X$ . Because we model deviations between measurements and predictions explicitly, the precision of the coordinates depends on the quality of the data and on the accuracy of the forward model. In the ideal case of a uniquely invertible model, the likelihood function is only peaked at the structure that satisfies the data (i.e., the conventional approach is contained as limiting case). The prior density  $p(X|I)$  takes prior knowledge about biomolecular structures into account and is determined by the physical energy and the temperature of the system (5).

The error distribution and the forward model typically contain auxiliary parameters  $\xi$  that are unavailable from the data but necessary in order to describe the problem adequately. In Bayesian theory, such nuisance parameters are treated in the same way as the coordinates: They are estimated from the experimental data by replacing  $X$  with  $(X, \xi)$  in Eq. 1. Assuming independence of  $X$  and  $\xi$ , the

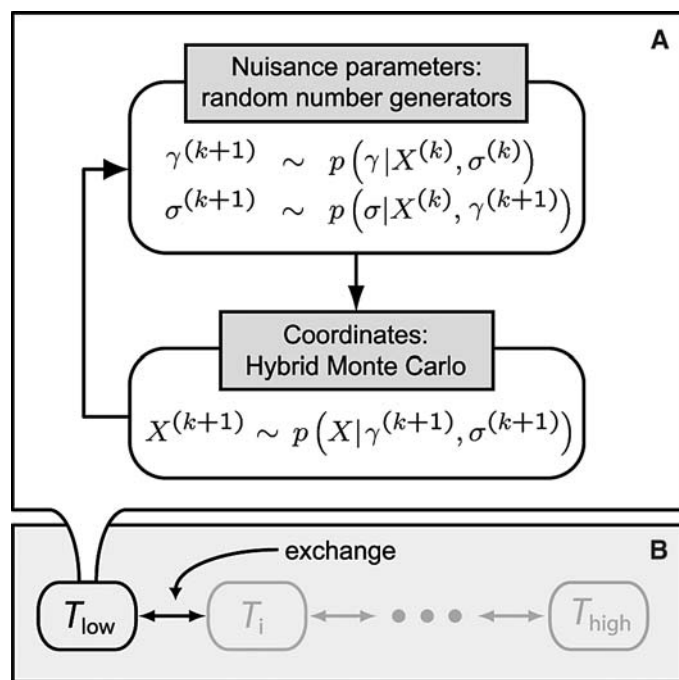
joint posterior density for all unknown parameters is

$$p(X, \xi|D, I) \propto p(D|X, \xi, I) p(X|I) p(\xi|I) \quad (2)$$

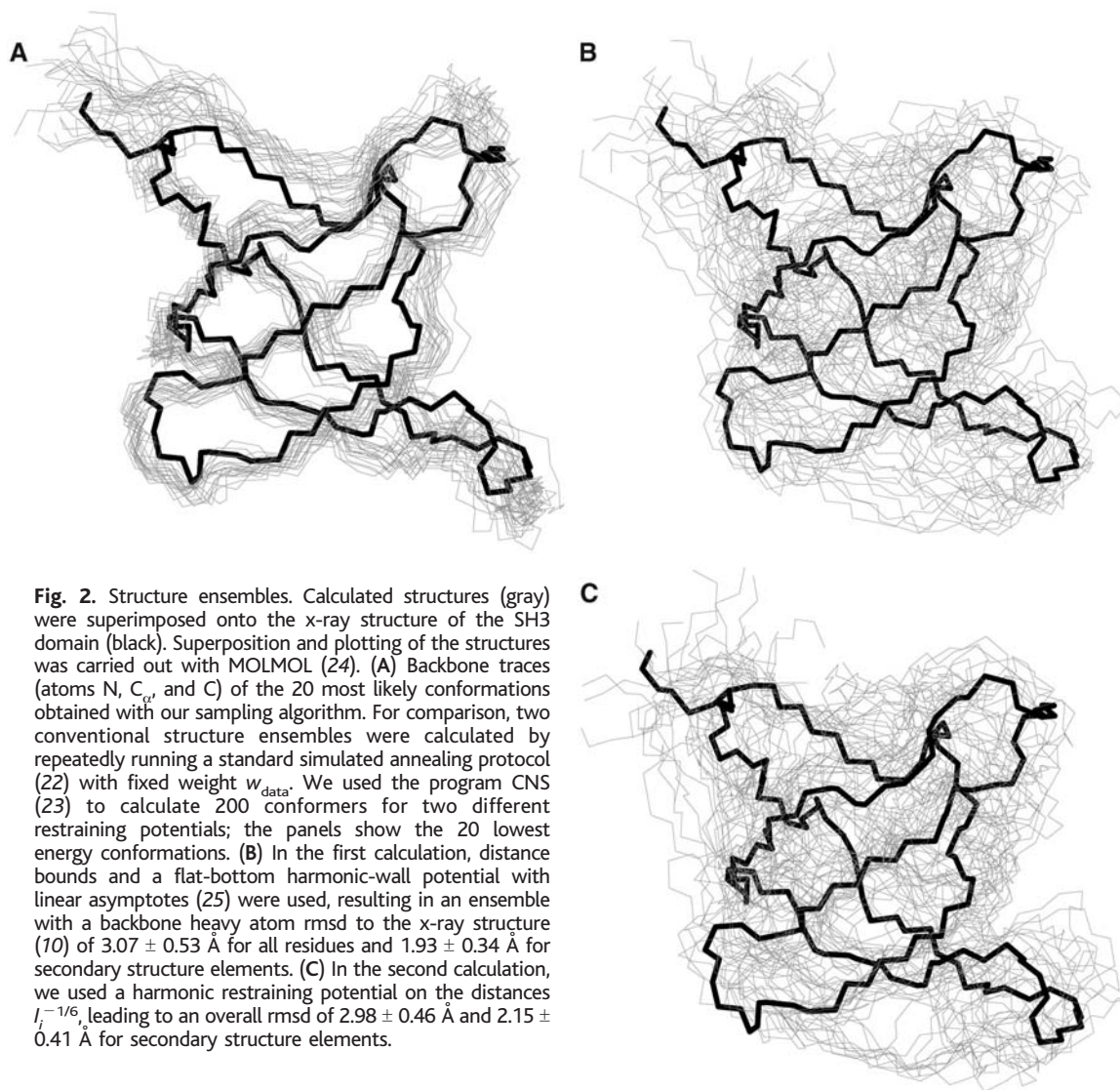
Equation 2 provides a unique rule to determine any quantity that is not accessible by experiment.

To demonstrate the practical feasibility of the ISD approach, we infer the molecular structure of the Fyn SH3 domain (59 amino acids length). Experimental distances between amide protons were derived from a series of NOESY spectra on a {15N, 2H} enriched protein (6). The data set is sparse: It comprises 154 measurements, of which on average only one per amino acid provides long-range structural information. The forward model  $I_i = \gamma d_i^{-6}(X)$  defined by the ISPA does not account for experimental errors and systematic effects like spin diffusion (7) and internal dynamics (7); hence, observed intensities will deviate from theoretical predictions. A log normal distribution (5) describes these deviations and introduces a second nuisance parameter  $\sigma$  that quantifies their magnitude. Thus, we have two nuisance parameters,  $\xi = (\gamma, \sigma)$ .

Although given in analytically closed form (5), it is practically impossible to evaluate the posterior density  $p(X, \gamma, \sigma|D, I)$  over all conformational space. Therefore, in our view, structure calculation comprises posterior simulation, which samples only regions that carry a considerable amount of probability mass. We have



To overcome energy barriers, we embed this scheme in a replica-exchange strategy, which simulates a sequence of heated copies of the system. Samples of the target distribution are generated in the low-temperature copy ( $T_{\text{low}}$ ) and propagate via stochastic exchanges between intermediate copies ( $T_{\text{low}} < T_i < T_{\text{high}}$ ) to the high-temperature system ( $T_{\text{high}}$ ). The temperature  $T_{\text{high}}$  is chosen such that the polypeptide chain can move freely in order to escape local modes of the probability density.



developed a Markov chain Monte Carlo (MCMC) algorithm based on the replica-exchange method (8) to simulate the joint posterior density of a structure determination problem (5, 9) (Fig. 1 and fig. S1).

The most pronounced features of the posterior density can be represented in a set of conformational samples. Although this looks at first glance like a conventional structure ensemble, the rationale behind our approach to obtain conformational samples is very different. The uncertainty of atomic positions is directly influenced by the uncertainty of nuisance parameters and by the quality of the data. Effects not described in the ISPA, such as protein dynamics, tend to increase the deviations between predicted and measured peak intensities. This is reflected in an increase of the error  $\sigma$  and consequently leads to a loss in structural precision. However, unless the forward model incorporates experimental information on protein dynamics, we cannot discriminate motion from imprecisions due to experimental errors or lack of data.

Compared with conventional structure ensembles, our conformational samples are much better defined and systematically closer to the structure obtained with x-ray crystallography (10) (Fig. 2). A comparison of the 20 most probable conformations with the x-ray structure yields a backbone heavy atom rmsd (root mean square deviation) of  $1.84 \pm 0.20$  Å for all residues and  $1.36 \pm 0.19$  Å for the secondary structural elements. This is a considerable improvement over conventional techniques used in (6), where an ensemble with an overall rmsd of  $2.86 \pm 0.33$  Å and an rmsd of  $2.01 \pm 0.28$  Å for secondary structure elements was obtained. This improvement originates in the calculation of structures by random sampling, which searches conformational space more exhaustively and suppresses topologically unlikely conformations. Misfolds such as mirror images can only be realized in a small number of ways; thus, they are entropically suppressed and do not show up in a statistical ensemble. Discriminating such conformations

on the basis of the hybrid energy is more difficult, in particular if the data are sparse.

A probabilistic structure ensemble is exclusively determined by the data and the working hypotheses that enter the analysis (which are in the presented example the ISPA, the log-normal error distribution, and our choice of the force field). Modifications will, of course, lead to changes in the structures. The atom positions, for example, are sensitive to the parameters and the functional form of the force field used in the conformational prior density. This also holds for conventional approaches, which are based on analogous assumptions. However, in addition, conventional methods require empirical rules to treat nuisance parameters, because they cannot be determined from the hybrid energy alone. Cross-validation (11, 12) and maximum likelihood methods (13), for example, have successfully been applied in NMR and crystallographic refinement to determine certain nuisance parameters such as the weight  $w_{\text{data}}$ . The ISD approach goes

beyond these techniques. Once the working hypotheses are made, Eq. 2 provides definite rules to determine any nuisance parameter, including its uncertainty, directly from the data (Fig. 3). Therefore heuristics and other subjective elements are superfluous.

Because conventional structure ensembles depend on user-specific parameter settings and on the minimization protocol, it is difficult if not impossible to assign statistically meaningful error bars to atomic coordinates. In contrast, stochastic samples drawn from the joint posterior density  $p(X, \gamma, \sigma|D, I)$  are statistically well defined and can directly be used to calculate estimates of mean values and standard deviations (14). As a special case, we can derive an average structure with atom-

wise error bars and are thus able to define an objective figure of merit for NMR structures (Fig. 4).

Bayesian and maximum likelihood approaches have already proven useful for data analysis and partial aspects of structure refinement in NMR spectroscopy and x-ray crystallography (15, 16, 13, 17). Our results suggest that structure determination can be solved entirely in a probabilistic framework.

It is straightforward to apply our approach to other NMR parameters. In case of three-bond scalar coupling constants, for example, an appropriate forward model is the Karplus curve (18) involving three coefficients that are treated as nuisance parameters. However, our method is not restricted to NMR data and can

be applied to other structure determination problems. Besides theoretical coherence, a rigorous probabilistic approach has decisive practical advantages. It has no free parameter and is stable for many more than the two nuisance parameters used in the example (19). Hence, tedious and time-consuming searches for optimal values are no longer necessary. Once the forward model to describe the data has been chosen, probability calculus uniquely determines the posterior distribution for all unknowns. It is then only a computational issue to generate posterior samples. Further intervention is not required, and structure determination attains objectivity.

#### References and Notes

1. S. Macura, R. R. Ernst, *Mol. Phys.* **41**, 95 (1980).
2. A. T. Brünger, M. Nilges, *Q. Rev. Biophys.* **26**, 49 (1993).
3. R. T. Cox, *Am. J. Phys.* **14**, 1 (1946).
4. E. T. Jaynes, *Probability Theory: The Logic of Science* (Cambridge Univ. Press, Cambridge, 2003).
5. Materials and methods are available as supporting material on Science Online.
6. T. K. Mal, S. J. Matthews, H. Kovacs, I. D. Campbell, J. Boyd, *J. Biomol. NMR* **12**, 259 (1998).
7. G. Lipari, A. Szabo, *J. Am. Chem. Soc.* **104**, 4546 (1982).
8. R. H. Swendsen, J.-S. Wang, *Phys. Rev. Lett.* **57**, 2607 (1986).
9. M. Habeck, M. Nilges, W. Rieping, *Phys. Rev. Lett.* **94**, 0181051 (2005).
10. M. E. Noble, A. Musacchio, M. Saraste, S. A. Courtneidge, R. K. Wierenga, *EMBO J.* **12**, 2617 (1993).
11. A. T. Brünger, *Nature* **355**, 472 (1992).
12. A. T. Brünger, G. M. Clore, A. M. Gronenborn, R. Saffrich, M. Nilges, *Science* **261**, 328 (1993).
13. P. D. Adams, N. S. Pannu, R. J. Read, A. T. Brünger, *Proc. Natl. Acad. Sci. U.S.A.* **94**, 5018 (1997).
14. M. H. Chen, Q. M. Shao, J. G. Ibrahim, *Monte Carlo Methods in Bayesian Computation* (Springer Verlag, New York, 2002).
15. R. B. Altman, O. Jardetzky, *Methods Enzymol.* **177**, 218 (1989).
16. M. Andrec, G. T. Montelione, R. M. Levy, *J. Magn. Reson.* **139**, 408 (1999).
17. G. Bricogne, *Methods Enzymol.* **276**, 361 (1997).
18. M. Karplus, *J. Am. Chem. Soc.* **85**, 2870 (1963).
19. W. Rieping, M. Habeck, M. Nilges, data not shown.
20. S. Geman, D. Geman, *IEEE Trans. Pattern Anal. Mach. Intell.* **6**, 721 (1984).
21. S. Duane, A. D. Kennedy, B. Pendleton, D. Roweth, *Phys. Lett. B* **195**, 216 (1987).
22. M. Nilges, M. J. Macias, S. I. O'Donoghue, H. Oschkinat, *J. Mol. Biol.* **269**, 408 (1997).
23. A. T. Brünger et al., *Acta Crystallogr. D* **54**, 905 (1998).
24. R. Koradi, M. Billeter, K. Wüthrich, *J. Mol. Graph.* **14**, 51 (1996).
25. M. Nilges, S. I. O'Donoghue, *Prog. Nucl. Magn. Reson. Spectrosc.* **32**, 107 (1998).
26. The authors thank I. D. Campbell for kindly providing the experimental SH3 NMR data. This work was supported by European Union grants QL2-CT-2000-01313 and QL2-CT-2002-00988. The 20 most likely structures and the restraint list used in the calculation have been deposited in the Protein Data Bank under accession code 1ZBJ. The structure determination program is available from the authors on request to M.N.

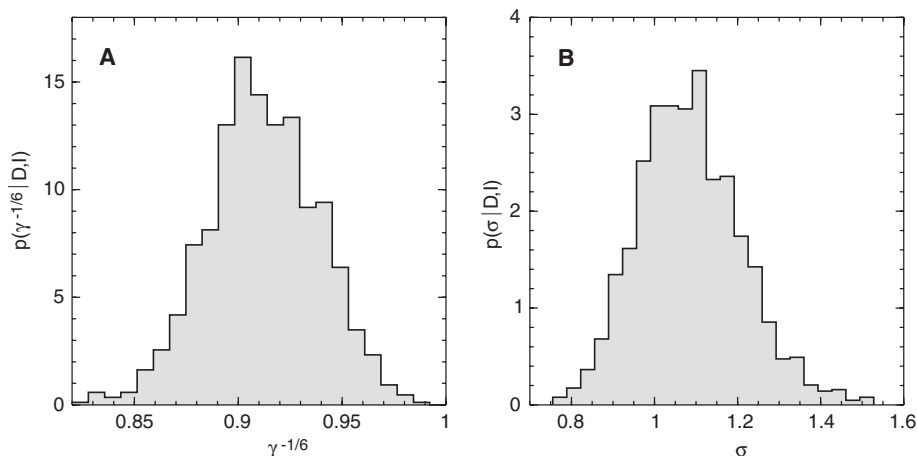
#### Supporting Online Material

[www.sciencemag.org/cgi/content/full/309/5732/303/DC1](http://www.sciencemag.org/cgi/content/full/309/5732/303/DC1)

Materials and Methods

Fig. S1

31 January 2005; accepted 16 May 2005  
10.1126/science.1110428



**Fig. 3.** Estimation of nuisance parameters. Posterior histograms compiled from MCMC samples for the scaling factor  $\gamma$  in the ISPA and for the width  $\sigma$  of the log normal error distribution. (A) Posterior histogram  $p(\gamma^{-1/6}|D, I)$  for the inverse sixth power of  $\gamma$ . This factor corrects interproton distances to match the experimental distances best. (B) Posterior histogram  $p(\sigma|D, I)$  for the error  $\sigma$ . In conventional approaches, this analog to the weight ( $w_{\text{data}} \propto \sigma^{-2}$ ) can only be estimated via cross-validation or must be set empirically.

**Fig. 4.** Conformational uncertainty. MOLMOL "sausage" plot of the mean structure with atom-wise error bars indicated by the thickness of the sausage. The 20 most probable conformations (also shown in Fig. 2A) from the simulation of the joint posterior distribution  $p(X, \gamma, \sigma|D, I)$  were used to calculate the average structure and its precision. The local precision ranges from 0.6 Å for secondary structure elements to 4.6 Å for loop regions (bottom and right-hand side) and termini (top). The average precision is 1.07 Å. The average precision of the structure ensembles calculated with CNS is 4.93 Å for the flat-bottom harmonic-wall potential and 5.04 Å for the harmonic potential.

