Bayesian inference applied to macromolecular structure determination

Michael Habeck, Michael Nilges,* and Wolfgang Rieping
Unité de Bio-Informatique Structurale, Institut Pasteur 25-28, Rue du Docteur Roux, 75015 Paris, France
(Received 17 March 2005; published 20 September 2005)

The determination of macromolecular structures from experimental data is an ill-posed inverse problem. Nevertheless, conventional techniques to structure determination attempt an inversion of the data by minimization of a target function. This approach leads to problems if the data are sparse, noisy, heterogeneous, or difficult to describe theoretically. We propose here to view biomolecular structure determination as an inference rather than an inversion problem. Probability theory then offers a consistent formalism to solve any structure determination problem: We use Bayes' theorem to derive a probability distribution for the atomic coordinates and all additional unknowns. This distribution represents the complete information contained in the data and can be analyzed numerically by Markov chain Monte Carlo sampling techniques. We apply our method to data obtained from a nuclear magnetic resonance experiment and discuss the estimation of theory parameters.

DOI: 10.1103/PhysRevE.72.031912 PACS number(s): 87.15.Aa

I. INTRODUCTION

Biomolecules such as proteins fold into thermodynamically stable structures that are crucial for their biological function. To date, x-ray crystallography and nuclear magnetic resonance (NMR) spectroscopy are the only experimental methods that permit the determination of macromolecular structures with atomic resolution routinely. Both techniques measure the molecule's response to an externally controlled perturbation, such as the intensities in a x-ray diffraction pattern or the cross peaks in a nuclear Overhauser effect spectroscopy (NOESY) experiment [1]. Analysis of a measurement y requires a *forward model* f that predicts the response from a given conformation θ: $y=f(\theta)$. The structure determination problem is to determine the unknown structure from measurements $\{y_i\}$; i.e., to go in the inverse direction.

Ideally, it holds that $y_i=f_i(\theta)$ for the true structure, and we would have to solve a system of nonlinear equations in order to determine its unknown coordinates. This is commonly attempted [2] by minimizing a target function of the form

$$G(\theta) = E(\theta) + \lambda F(\theta), \tag{1}$$

where F assesses the match between predicted and observed data, and λ is an unknown positive weighting constant. In case of noiseless data, the penalty term F is minimal for conformations that exactly satisfy the nonlinear equations. The potential energy E is a regularizer that accounts for the underdeterminedness of the inverse problem.

The analysis of realistic data, however, is more involved due to several reasons.

First, experimental data are incomplete. In NMR spectroscopy, for instance, the most informative measurands are dipolar relaxation rates [3]. These are observable for protons close in space and thus constitute only a small subset of all interatomic distances. Thus, the experiment provides incomplete information in the sense that the data can be explained by multiple conformations.

Second, the forward model almost always involves additional quantities α that are not measurable; i.e., $y_i=f_i(\theta;\alpha)$. The parametrization α of the forward model might not be of primary interest, yet it is essential to establish the relation between atomic coordinates and the measurands. Examples are the phases in x-ray crystallography or the scale of cross-relaxation rates in NMR.

Third, the data are subject to errors. Typically, errors are caused by quantities that vary uncontrollably during experimentation, and by the forward model which is often based on approximations. Prior to analysis, data are often preprocessed, which can introduce additional uncertainty. In practice, the various factors leading to deviations of the observed from the ideal data are unknown. However, they affect our choice of the weighting constant λ in the hybrid energy (1), since the quality of the data determines how much we can trust them.

Ad hoc solutions to overcome these difficulties have been proposed. Heuristics like cross validation [4,5], for example, are utilized to judge the quality of structures and to determine the unknown weighting λ. In case of heterogeneous or noisy data, however, cross validation becomes time consuming and unstable. Empirical methods are also employed to determine the parametrization of the forward model (the calibration of NOESY signals [6], for example, determines the unknown scale of cross-relaxation rates). A further difficulty is to assess the reliability of the reconstructed structure. Since the optimization framework lacks a means to quantify uncertainty, it is not possible to define statistically meaningful error bars. Usually, the reliability of NMR structures is estimated by running the optimization algorithm several times, each time from random initial conditions. The resulting "ensemble" is utilized like a statistical sample to estimate the precision of the coordinates. However, since such an ensemble mainly reflects properties of the minimization protocol, this approach is problematic.

II. INFERENTIAL STRUCTURE DETERMINATION

The difficulties rendering structure determination by inversion an ill-posed problem are due to the uncertainty and

*Electronic address: nilges@pasteur.fr

incompleteness of the information provided by the experiment. Structure determination requires reasoning from partial knowledge and is therefore an inference problem. Since incomplete information entails uncertain conclusions, the derivation of a single and unique structure is, as a matter of principle, impossible. However, this is what one tries when minimizing the hybrid energy: One attempts to invert the equations $y_i = f_i(\theta)$ numerically.

The optimization approach is inadequate because it neglects the ambiguity inherent in the data analysis or at best treats it in an ad hoc fashion. Instead, we are seeking a means to represent this ambiguity in the most honest way. Cox proved [7] that the only way to quantify uncertainty systematically and consistently is through probabilities. In this view, probability theory is an extension to deductive logic as it provides rules for inductive reasoning [8]. Probabilities become degrees of belief that measure the validity of a conclusion. They span a continuum of truth values thereby extending deductive logic. Probabilities always refer to what we already know. Therefore, only a conditional probability $P(H|I)$ is well-defined: It quantifies the plausiblity of a hypothesis H in the context of information I. This is radically opposed to the frequentist interpretation where probabilities are defined through the long-run behavior of "random variables."

A hypothesis can be any statement which is compliant with the Boolean algebra of propositions. The demand for consistency with the rules of propositional logic imposes a structure on the plausibility measure: the Algebra of Probable Inference [7]. Its fundamental relationships are the sum rule: $P(A|I)+P(\bar{A}|I)=1$ and the product rule: $P(A,B|I) = P(A|B,I)P(B|I)$.

In structure determination we are concerned with propositions: "data D were recorded" and "during the experiment the molecular structure was θ" symbolized through D and θ, respectively. What do the data tell us about the unknown structure? The complete answer to this question is the conditional probability $P(\theta|D,I)$. The interpretation of the data always needs to be based on some model $P(D|\theta,I)$, the *likelihood*, which establishes the connection between theory and experiment. Bayes' theorem [8], a direct consequence of the product rule, inverts this probability

$$P(\theta|D,I) = P(\theta|I)\frac{P(D|\theta,I)}{P(D|I)}. \tag{2}$$

Here, the *prior* probability $P(\theta|I)$ occurs naturally. It represents our background knowledge about the molecular conformation before experimentation. The probability $P(D|I)$ serves here as a normalization constant which will not be important for what follows. Bayes' theorem reads: The likelihood for observing D has to be weighted with the prior probability for θ to yield the *posterior* probability $P(\theta|D,I)$ which describes our knowledge about θ after the experiment. Thus, probability theory solves any inverse problem in biomolecular structure determination in a well-defined way by ranking all possible conformations according to their posterior probabilities. The case of exactly invertible data is contained in this formulation as a limiting case.

We propose to solve any structure determination problem by calculating the posterior probability $P(\theta|D,I)$ and by using this probability to quantify any hypothesis about the unknown structure. In order to distinguish our principle from the conventional minimization approach, we coin it inferential structure determination (ISD) [9].

III. MODELING MACROMOLECULAR STRUCTURAL DATA

In order to infer the structure of a biomolecule from given data, we first need to set up the likelihood and the prior probability. Interactions between the atoms restrict the possible conformations of a molecule. We incorporate this prior knowledge by means of a force field $E(\theta)$ which quantifies intramolecular interactions. Provided the experimental data set is sufficiently complete, solvent interactions have only a minor influence on the overall quality of NMR structures [10] and will therefore be neglected. Assuming that experiments are carried out at a constant temperature β^{-1} and following the principle of maximum entropy [11], the canonical ensemble

$$\pi(\theta) = \frac{1}{Z(\beta)}\exp\{-\beta E(\theta)\} \tag{3}$$

represents our prior knowledge: $\mathrm{d}P(\theta|I) = \pi(\theta)\mathrm{d}\theta$.

The specific form of the likelihood depends on the nature of the data and, as outlined above, generally consists of two constituents: A *forward model* $\hat{y}_i = f_i(\theta;\alpha)$ relates the conformational degrees of freedom θ to the expected observations $\hat{y}_i$ and possibly involves nonmeasurable parameters α. An *error model* $g(y_i;\hat{y}_i,\sigma)$ accounts for deviations of the observed from the expected data. The parametrization of the error model introduces a second kind of unknowns, the "errors" σ. Given n independent measurements $D=\{y_1,\ldots,y_n\}$, we are therefore dealing with an extended likelihood

$$P(D|\theta,\alpha,\sigma,I) = \prod_{i=1}^{n} g(y_i;f_i(\theta,\alpha),\sigma) \tag{4}$$

instead of $P(D|\theta,I)$. In order to evaluate the likelihood, the parametrization of the forward model and the error model (α and σ, respectively) need to be known. Hence, both quantities appear on the conditioning side in Eq. (4). For given data D, we use the notation $L(\theta,\alpha,\sigma) \propto P(D|\theta,\alpha,\sigma,I)$ to indicate, that we view the density (4) as a function of the hypothesis parameters.

The auxiliary parameters α and σ need to be introduced to describe the data appropriately; in statistical parlance such quantities are called *nuisance parameters* [8]. Since it is often unclear how to set such unknowns, the treatment of nuisance parameters poses a great problem to optimization algorithms. Even if these parameters were to be optimized during structure calculation, empirical rules are still needed since an equivalent of the extended likelihood is missing: The hybrid energy (1) is a target function for the coordinates only.

Probability theory, in contrast, permits the determination of any unknown quantity. All we need is experimental evi-

dence that logically depends on the respective parameter. Application of Bayes' theorem then immediately yields an extended posterior density which allows us to do inferences on θ, α, and σ

$$p(\theta,\alpha,\sigma) \propto L(\theta,\alpha,\sigma)\pi(\theta,\alpha,\sigma). \tag{5}$$

The posterior density (5) is a joint probability which simultaneously determines all unknowns. Here, $\pi(\theta,\alpha,\sigma)$ denotes the most general prior distribution defined on the joint hypothesis space spanned by the parameters θ, α, and σ. In many practical problems one can assume mutual independence of the nuisance parameters and the coordinates. In this case the prior distribution factorizes: $\pi(\theta,\alpha,\sigma) = \pi(\theta)\pi(\alpha)\pi(\sigma)$. However, in some applications this assumption is inadequate, as, for example, in x-ray crystallography where the phases depend on the structure.

Formally, we can eliminate uninteresting parameters via integration. We use this so-called *marginalization* rule [8] to arrive at a posterior distribution which depends on the atomic coordinates only

$$p(\theta) \propto \int \mathrm{d}\alpha\, \mathrm{d}\sigma L(\theta,\alpha,\sigma)\pi(\theta,\alpha,\sigma). \tag{6}$$

We can just as well reduce our hypothesis space to the macromolecular configuration space by defining the *integrated* likelihood function

$$L(\theta) = \int \mathrm{d}\alpha\, \mathrm{d}\sigma L(\theta,\alpha,\sigma)\pi(\alpha,\sigma|\theta) \tag{7}$$

with the conditional prior density being $\pi(\alpha,\sigma|\theta) = \pi(\theta,\alpha,\sigma)/\pi(\theta)$. The function $L(\theta)$ is a sort of "effective likelihood" that one obtains when considering the nuisance parameters α and σ as "hidden variables." Using Eq. (7), we can apply Bayes' theorem in conformational space and arrive at the same result as through elimination of the nuisance parameters in the posterior distribution [Eq. (6)].

IV. APPLICATION TO DISTANCE DATA MEASURED BY NMR

As an application of ISD, we analyse dipolar relaxation rates measured in the NOESY experiment [1]. NOESY is a multidimensional pulsed NMR experiment that measures the exchange of magnetization due to dipolar relaxation. Each resonance in a NOESY spectrum can be assigned to a pair of interacting spins k and l. In a first order approximation, the volume V_{kl} of a resonance peak is proportional to the inverse sixth power of the distance r_{kl} of the two spins [1]. This is due to the r^{-3} dependence of dipolar interactions and to the fact that relaxation is a second order effect. Thus, our forward model to describe the observed volume V_{kl} is

$$V_{kl}(\theta;\gamma) = \gamma r_{kl}^{-6}(\theta), \tag{8}$$

where γ denotes a positive scaling factor. This model is called the isolated spin-pair approximation (ISPA) [12] because it reduces the magnitization transfer in a multispin system to the transfer between pairs of spins.

The ISPA is an approximation. Besides spin diffusion [1], it neglects the internal flexibility of the macromolecule [13]. We therefore expect observed volumes to deviate from those predicted by relation (8) not so much because of experimental errors but due to our imprecise forward model. If one neglects the dynamics of the molecule, peak volumes are positive. It is therefore convenient to describe these deviations using a lognormal distribution [14]

$$g(V_{kl};\hat{V}_{kl},\sigma) = \frac{1}{\sqrt{2\pi\sigma^2 V_{kl}}} \exp\left\{-\frac{1}{2\sigma^2}\ln^2(\hat{V}_{kl}/V_{kl})\right\}, \tag{9}$$

defined on the positive axis ($V_{kl}>0$) and involving an error parameter $\sigma>0$. Our error model is a conservative choice and can be motivated by the principle of maximum entropy. Since peak volumes are non-negative their errors are multiplicative. Assuming no systematic error and knowledge of the variance of the logarithmic errors, the lognormal distribution results as the least biased error model.

Our data comprise n peak volumes V_{kl}, where each peak has been assigned to a pair of interacting spins k and l. The complete likelihood function (4) for NOESY measurements is

$$L(\theta,\gamma,\sigma) \propto \sigma^{-n} \exp\left[-\frac{n}{2\sigma^2}\{\ln^2[v(\theta)/\gamma] + s^2(\theta)\}\right], \tag{10}$$

where we introduced the statistics

$$v(\theta) = \left(\prod_{(k,l)} V_{kl} r_{kl}^6(\theta)\right)^{1/n}, \tag{11a}$$

$$s(\theta) = \left(\frac{1}{n}\sum_{(k,l)} \ln^2[V_{kl} r_{kl}^6(\theta)/v(\theta)]\right)^{1/2}; \tag{11b}$$

$v(\theta)$ is the geometric average of the ratios of the measured and the uncalibrated predicted volumes, $s(\theta)$ is the standard deviation of the normalized logarithmic ratios. The sum involves those n pairs of protons for which a NOESY cross peak was observed.

This model requires two nuisance parameters, γ and σ. Since knowledge of the molecular conformation tells us nothing about the calibration factor γ or the error σ, our prior distribution factorizes

$$\pi(\theta,\gamma,\sigma) = \pi(\gamma)\pi(\sigma)\pi(\theta) = \frac{1}{\gamma\sigma}\frac{1}{Z(\beta)} \exp\{-\beta E(\theta)\}. \tag{12}$$

The canonical ensemble makes up the conformational part of the prior. We choose Jeffreys' prior [15] to describe our knowledge of both nuisance parameters γ and σ. Jeffreys' priors represent the fact that we know nothing about γ and σ except that both are scale parameters; i.e., the likelihood would not change, if γ and σ were measured in different units.

At this stage, it is instructive to compare our probabilistic approach with conventional, minimization-based techniques. The negative logarithm of the posterior distribution $p(\theta,\gamma,\sigma)$ may serve as a joint target function for the most probable structure and the most probable values of γ and σ.

In contrast, the hybrid energy (1) is restricted to conformational space and therefore serves as target function for the coordinates only. A comparison of both target functions yields the following analogies: The counterparts of the physical energy and the penalty term in Eq. (1) are the negative logarithms of the prior probability and the likelihood function: $E \propto -\ln \pi$, $F \propto -\ln L$. The weight λ is proportional to the inverse squared error: $\lambda \propto \sigma^{-2}$. The hybrid energy, however, lacks a term analoguous to the prior probabilities for γ and σ and the normalization constant σ^{-n} stemming from the extended likelihood function (10). In probabilistic modeling these terms serve as regularizers and allow the determination of γ and σ along with the conformationial degrees of freedom. Thus, hybrid energy minimization turns out to be a special case of the Bayesian formulation. It is only valid in cases where we dispose of prior knowledge about unknowns of the forward model and about experimental errors.

For model (7) we can calculate the integrated likelihood analytically:

$$L(\theta) = [s(\theta)]^{-(n-1)}, \tag{13}$$

which depends on the molecular conformation only through the statistic $s(\theta)$. We can obtain this likelihood function when setting $\gamma = v(\theta)$ and $\sigma = s(\theta)$ in the joint likelihood function (10) and reducing the number of data by one. This shows that estimation of the nuisance parameters correlates all measurements at a cost of one measurement. Based on the integrated likelihood (13), we can define a target function which depends on the coordinates only

$$-\ln p(\theta) = (n-1)\ln s(\theta) + \beta E(\theta). \tag{14}$$

Minimization of $-\ln p(\theta)$ allows us to determine the most probable structure without assuming knowledge of γ and σ. This shows that inferential structure determination does not contain free parameters but eliminates auxiliary quantities by using the rules of probability calculus. In other words, we can either use the extended posterior $p(\theta, \gamma, \sigma)$ or the marginalized posterior $p(\theta)$ and obtain the same results.

V. STRUCTURE CALCULATION

If our aim is the structure of biological macromolecules such as proteins, analytical investigations of the posterior density become very complex, and we have to resort to numerical methods. Any analysis based on the posterior density boils down to the calculation of integrals of the form [8]

$$I(h) = \int \mathrm{d}\theta\, \mathrm{d}\alpha\, \mathrm{d}\sigma h(\theta, \alpha, \sigma) p(\theta, \alpha, \sigma). \tag{15}$$

If the hypothesis function h is independent of the nuisance parameters, we can directly integrate over the marginalized posterior $p(\theta)$. At present, the most powerful methods for computing quantities of the form (15) are based on Markov chain Monte Carlo (MCMC) algorithms [16]. These methods construct a first order Markov process by sucessively applying a stochastic transition kernel. After a certain convergence period, the Markov chain generates random samples from the posterior distribution.

In our view, structure calculation amounts to the generation of random samples from the joint posterior density $p(\theta, \alpha, \sigma)$. This differs fundamentally from conventional structure calculation algorithms based on nonlinear optimization. Posterior sampling not only identifies high-probability modes in due proportion to each other but also yields direct estimates of the precision of all hypothesis parameters.

Our algorithm for generating posterior samples $\{\alpha^{(t)}, \sigma^{(t)}, \theta^{(t)}\}$ uses a combination of three MCMC strategies. The Gibbs sampling procedure [17] facilitates a split up of the sampling scheme into three steps. Each parameter is sampled sequentially conditioned on the current values of the other parameters

$$\alpha^{(t+1)} \sim p(\alpha | \theta^{(t)}, \sigma^{(t)}),$$

$$\sigma^{(t+1)} \sim p(\sigma | \theta^{(t)}, \alpha^{(t+1)}),$$

$$\theta^{(t+1)} \sim p(\theta | \alpha^{(t+1)}, \sigma^{(t+1)}). \tag{16}$$

In order to apply the Gibbs sampling scheme, we must thus be able to simulate the conditional posterior densities for the nuisance parameters and the coordinates. For simple distributions this can be done by using random number generators. However, for highly correlated parameters such as the conformational degrees of freedom, more powerful methods need to be employed.

Already the conformational prior $\pi(\theta)$ exhibits a complicated topography with ridges and isolated peaks. Noncovalent interactions, for instance, penalize van der Waals overlaps and correlate all parameters. We employ the hybrid Monte Carlo (HMC) algorithm [18] to simulate the conditional posterior in conformation space. The HMC method uses molecular dynamics (MD) [19] to generate a candidate conformation which is accepted according to the Metropolis criterion [20]. The dynamics is defined by using the negative logarithm of the conditional conformational posterior distribution

$$-\ln p(\theta | \alpha, \sigma) = -\ln L(\theta, \alpha, \sigma) + \beta E(\theta) \tag{17}$$

as potential energy.

For realisitic biomolecular systems, the Gibbs sampler (16) is likely to get stuck in high-probability modes and thus fails to explore the entire parameter space. These modes correspond to different configurations of the macromolecule that fulfill the data comparably well. However, missing a high-probability fold would bias our analysis.

A physical system that is trapped in a metastable state can be melted by increasing the temperature. If the kinetic energy is sufficiently high the system easily explores all regions of the configuration space. The Replica-exchange Monte Carlo method [21] exploits this observation: It considers a composite Markov chain comprising several noninteracting copies of the system. Each of these "heat baths" is simulated at a different temperature. By exchanging configurations between neighboring copies, the heat baths are coupled, which significantly enhances the mobility of the individual Markov chains.

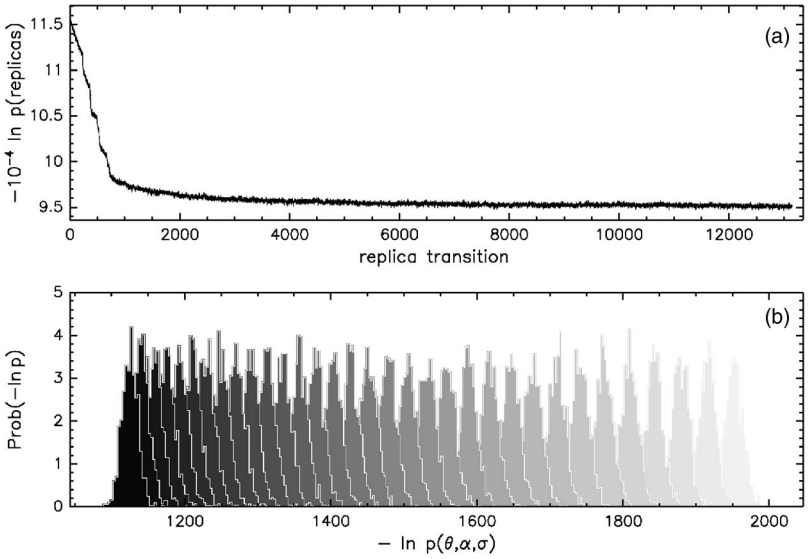

FIG. 1. Upper panel: trace of the negative logarithm of the posterior distribution of all heat baths. Lower panel: distributions of the negative logarithms of the individual posterior densities. The darkness of a histogram corresponds to the position of its heat bath in the replica sequence. Darker histograms correspond to lower "temperatures."

We introduce two "temperatures" [22]: A parameter λ weighs the likelihood, thereby controlling the contribution of the data. A second parameter q determines the shape of the conformational prior. Our heat baths are

$$f(\theta,\alpha,\sigma;\lambda,q) \propto [L(\theta,\alpha,\sigma)]^{\lambda}\pi(\theta;q)\pi(\alpha)\pi(\sigma). \quad (18)$$

The weight λ varies between 0 and 1. Following Hansmann and Okamoto [23], we use Tsallis' extension of the canonical ensemble as conformational prior

$$\pi(\theta;q) \propto \{1+\beta(q-1)[E(\theta)-E_{\min}]\}^{-q/(q-1)}. \quad (19)$$

The parameter $q \geqslant 1$ controls the degree of deformation; $q \to 1$ restores the Boltzmann ensemble. The Tsallis ensemble no longer suppresses high-energy configurations exponentially. This allows atoms to pass through each other, which facilitates large conformational changes. Our target distribution corresponds to $\lambda=q=1$; i.e., $p(\theta,\alpha,\sigma)=f(\theta,\alpha,\sigma;1,1)$. We introduce Tsallis' ensemble merely for computational reasons. The thermodynamic properties of the molecule are still described by a Boltzmann ensemble.

VI. ANALYSIS OF A NOESY EXPERIMENT

We used the outlined approach to analyze NMR measurements for the Tudor domain of the human Survival of Motor Neuron (SMN) protein [24]. The Tudor domain consists of 56 amino acids (comprising 859 atoms) and exhibits a β-barrel fold. At room temperature, bond lengths, bond angles, and ring planarities show little variance. In good approximation we keep these parameters fixed and use the force field of the empirical conformational energy program for peptides (ECEPP/2) [25,26] to describe the covalent geometry of the polypeptide chain. In this case, 254 torsion angles $\theta=\{\theta_j\}$ are the only degrees of freedom [27]. A repulsive potential approximates the Lennard-Jones potential and describes non-bonded forces acting between atom k and l

$$E_{kl}(\theta) = \frac{c_{kl}}{2}\begin{cases}[d_{kl}-d_{kl}(\theta)]^4, & d_{kl}(\theta) < d_{kl} \\ 0, & d_{kl}(\theta) \geqslant d_{kl}\end{cases}. \quad (20)$$

Values for force constants and minimum distances c_{kl} and d_{kl}, respectively, were taken from the PROLSQ [28] x-ray refinement program.

Two data sets were derived from NOESY spectra recorded for ^{13}C and ^{15}N edited protein samples [24]. The ^{13}C

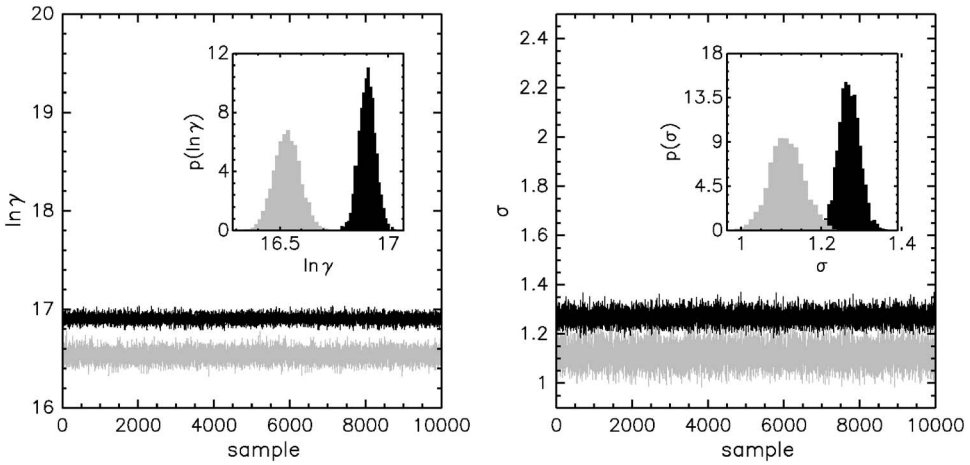

FIG. 2. Posterior samples of spectral scales and errors of the two data sets. Left panel: samples of the spectral scales γ_C (black) and γ_N (grey) drawn from the joint posterior. Right panel: Posterior samples of the error parameters σ_C (black) and σ_N (grey). The insets show the corresponding histograms.

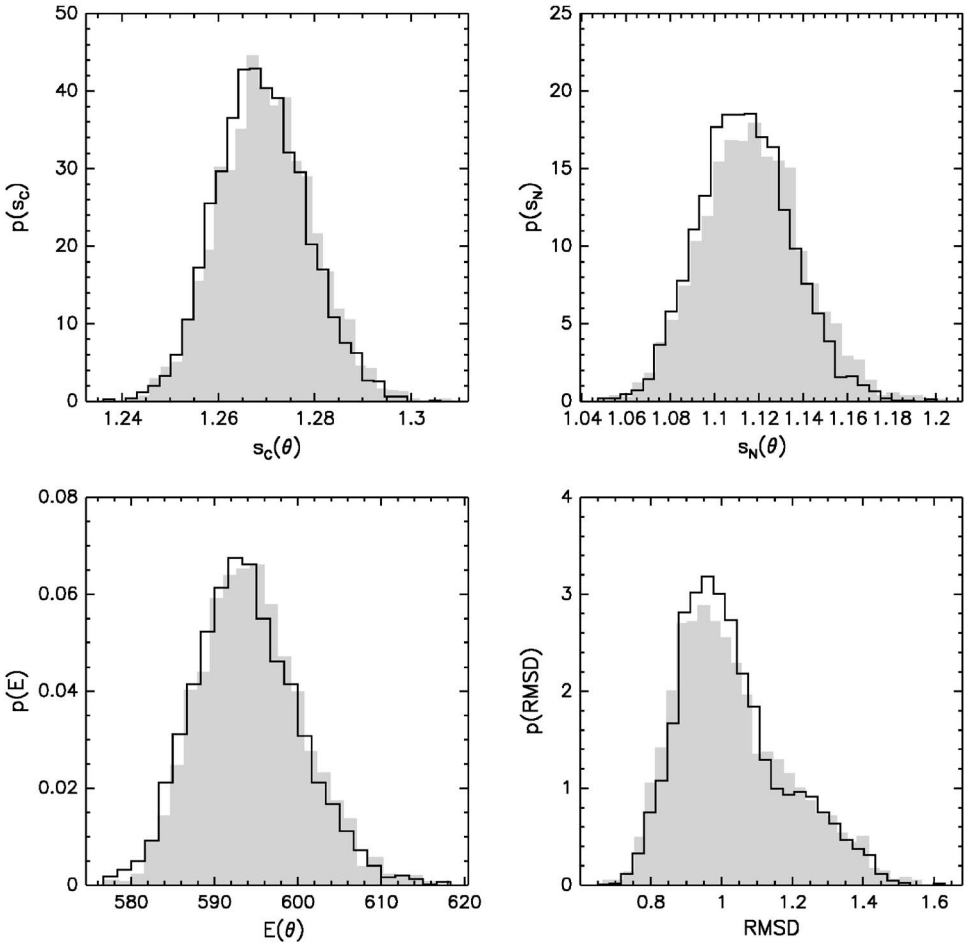

FIG. 3. Posterior histograms of conformational macrovariables that demonstrate the equivalence of the two simulations. Results from the simulation of the joint posterior density are shown as solid curves, histograms for the simulation of the marginal posterior are shown as the gray shaded area. Upper panels: distribution of the statistic $s(\theta)$ [Eq. (11b)] for the ^{13}C data set (left) and for the ^{15}N data set (right). Lower panels: histogram of the energy $E(\theta)$ (left) and of the RMSD to the x-ray structure (right).

data comprise $n_C=1444$, the ^{15}N data $n_N=431$ assigned cross peaks. In order to describe the data, we have to introduce four nuisance parameters: two spectral scales $\alpha=\{\gamma_C,\gamma_N\}$ and two errors $\sigma=\{\sigma_C,\sigma_N\}$.

The conditional posterior densities for the nuisance parameters are the lognormal distribution and the gamma distribution [14]

$$\gamma_j \sim \mathrm{LN}[\ln v_j(\theta),\sigma_j^2/n_j], \tag{21a}$$

$$\sigma_j^{-2} \sim G[n_j/2,\chi_j^2(\theta,\gamma_j)/2], \tag{21b}$$

with the goodness of fit for a single data set being $\chi_j^2(\theta,\gamma_j)=n_j[\ln^2[v_j(\theta)/\gamma_j]+s_j^2(\theta)]$. Random number generators for these distributions exist and can readily be applied in the Gibbs sampling scheme (16). Torsion angle samples are obtained by applying the HMC method to the conditional conformational posterior

$$p(\theta|\gamma,\sigma) \propto \exp\left\{-\frac{1}{2}\sum_{j=\mathrm{C,N}}\chi_j^2(\theta,\gamma_j)/\sigma_j^2-\beta E(\theta)\right\}. \tag{22}$$

Hamilton equations of motion were derived from $-\ln p(\theta|\gamma,\sigma)$ and integrated with the leapfrog algorithm [29]. Our replica arrangement consists of 50 copies, simulated at a temperature of 300 K with q ranging from 1.001 to 1.1 and λ from 0.1 to 1.0. We ordered the heat baths such that, in the first half of the arrangement, λ was succesively turned off. In the second half also nonbonded interactions were switched off by increasing q [22]. In total, we performed 13 000 replica transitions, each consisting of 25 HMC steps; the MD trajectories of HMC moves had a length of 250 steps. In each heat bath, we use an extended conformation as the initial state. The algorithm is parallelized such that every heat bath is simulated on a separate processor. A replica simulation demands more computational resources than standard minimization techniques. We used a PC cluster with 50 nodes; the simulation converged after two days. After convergence, the time required to calculate a single structure is comparable to that needed by conventional methods.

Figure 1(a) shows the negative logarithm of the joint posterior probability of all copies. After an initial convergence phase, the composite Markov chain reaches its invariant distribution, which is indicated by a plateau. The mixing efficiency of the Markov chain depends on the number of heat baths as well as on the exchange rate between neighboring copies. Therefore, the parameters λ and q need to be chosen carefully as they control the overlap between neighboring posterior distributions and hence the exchange rate. Our choice of both parameters resulted in an average exchange rate of 70% and sufficiently large overlaps [see Fig. 1(b)].

Simulation of the full posterior $p(\theta,\gamma_N,\gamma_C,\sigma_N,\sigma_C)$ estimates the unknown spectral scales and errors along with the torsion angles. Figure 2 shows traces of the posterior samples of the four nuisance parameters. In every instance,

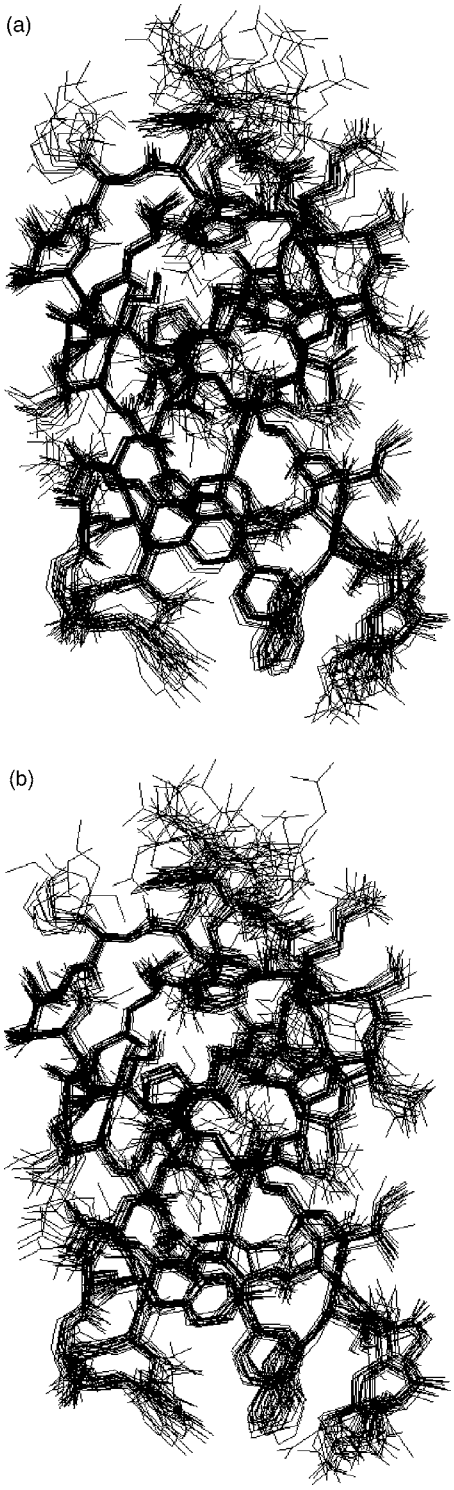

FIG. 4. Bundles of the 20 most likely conformations. (a) structures from the simulation of the joint posterior. (b) structure from the simulation of the marginal posterior. The superposition and plotting was carried out with the program MOLMOL [31].

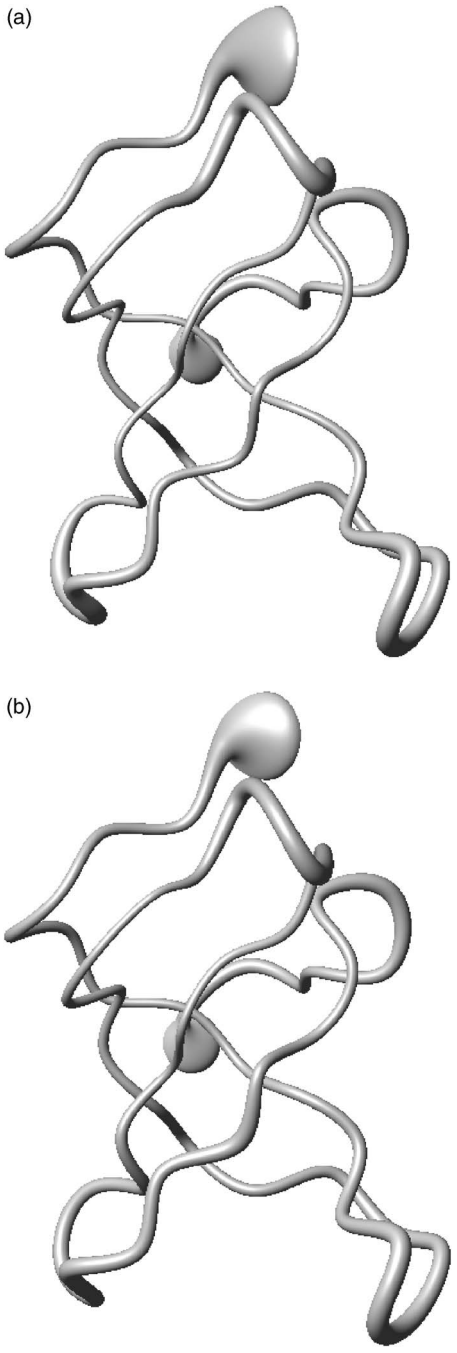

FIG. 5. Mean structures shown as MOLMOL "sausage plots." The thickness of the sausage indicates the precision of the coordinates. (a) mean structure from the simulation of the full posterior. (b) mean structure from the simulation of the marginalized posterior.

the samples vary around their most probable value. Representing the posterior samples as histograms, we obtain unimodal distributions with a certain spread. The spread is a measure for the precision of the parameter estimate and shows that the nuisance parameters cannot be determined from the data with absolute certainty. In both cases, the parameters used to model the ^{15}N data have a larger spread, which reflects the smaller size of the ^{15}N data set compared to the ^{13}C data. The error σ_C is larger than σ_N, which is due to relaxation mechanisms that affect the ^{13}C data more than the ^{15}N measurements.

Figure 2 is one of our main results. It demonstrates that the information contained in the experimental data suffices to determine the unknowns of both the forward model and the

error model. They no longer need to be determined by empirical rules or by the user. Hence, our conformational samples are not dependent on personal beliefs, but are objective in the sense, that they exclusively reflect the information content of the experimental data.

Simulation of the marginalized posterior $p(\theta)$ (13) is equivalent to simulation of the joint posterior $p(\theta, \gamma_N, \gamma_C, \sigma_N, \sigma_C)$. In the first case, integration over γ_j and σ_j is done analytically, in the latter case, by Monte Carlo integration. This mathematical equivalence holds for any likelihood function and prior distribution [see Eq. (6)] and is not a consequence of the factorization of the prior distribution. Figure 3 demonstrates this equivalence. Both simulations yield equivalent structures with respect to the macrovariables s_C, s_N, and E and are also in good agreement in terms of conformational accuracy. In case of the full posterior we obtain an expected carbon alpha root mean square distance (RMSD) to the x-ray structure [30] of 1.01 Å with an uncertainty of 0.15 Å, and for the marginal posterior a value of 1.02 ± 0.14 Å.

From our posterior samples, we selected a set of most likely conformations (which are members of a multidimensional confidence region in conformational space). The resulting conformational bundles are shown in Fig. 4. These bundles look similar to structure ensembles that are conventionally used to represent a molecular structure determined by NMR. However, the way we generate our structure ensemble is fundamentally different from the standard procedure. First, our ensemble is based on a closed mathematical expression for the conformational probability distribution. In conventional approaches structural variability expressed by an ensemble is a result of varitions in the initial conditions and also depends on the minimization protocol. This "operational variability" must not be confused with our statistically rigorous definition, which is independent of the structure calculation procedure. Thus, we could have used any other Markov chain Monte Carlo algorithm to calculate the structural uncertainty, provided the algorithm is ergodic. Second, based on the mathematical expression for the structure ensemble, we can rigorously define what we mean by "sampling" conformational space. Here, we sample conformational space by generating a sequence of random conformations that are distributed according to $p(\theta)$. Arbitrary "selection criteria" [6] that minimization-based techniques require to define the ensemble are thus superfluous.

In terms of structural quality, the most probable conformation of the Tudor domain obtained by our method is comparable to the structure calculated by a standard minimization method [24]. Our simulations of the full and the marginalized posterior yield both identical mean structures and an identical structural uncertainty (Fig. 5), which again illustrates the equivalence of the two procedures. We can directly estimate the conformational precision by the standard deviation of the posterior samples, as we would do for any parameter that is determined from experimental data. The atom-wise error bars are exclusively determined by the experimental data and by the assumptions required for data analysis (which are basically the choice of the forward model and of the error model). Since nuisance parameters are estimated along with the atomic coordinates, our estimate of structural precision is unbiased and objective.

VII. CONCLUSIONS

Bayesian probability theory is well-suited to formalize and solve macromolecular structure determination problems. We demonstrated that a full Bayesian analysis of NMR data is feasible by means of Markov chain Monte Carlo sampling, which enables reconstruction of the molecular structure of medium-sized proteins. We devised a prior and a likelihood for experimental data obtained from NOESY experiments. However, the approach is completely general and extensible to more complex data and theories. Bayes' theorem combines the prior and the likelihood into the conformational posterior distribution which provides an unbiased representation of our uncertainty about the true molecular structure. Hence, we calculate the uncertainty of a structure on the basis of a mathematically closed expression and therefore strictly separate algorithmic issues from data modeling. In contrast, the precision of the atomic coordinates calculated by conventional methods largely depends on the properties of the minimization protocol used to generate the structure ensemble and on choices in data treatment prior to structure calculation.

A major advantage of the Bayesian approach over optimization-based techniques is its ability to cope with nuisance parameters. Standard techniques do not provide efficient methods for obtaining optimal values, in particular not in the case of multiple nuisance parameters. In our approach auxiliary quantities need not be chosen empirically but can be estimated along with the atomic coordinates. Moreover, the precision of all hypothesis parameters is obtained. That way, it is possible to provide an objective measure of precision for NMR structures.

In case of large data sets of good quality, the accuracy of structures calculated with our method is comparable to those calculated with standard techniques. However, test calculations with sparse data sets show that our method outperforms standard techniques [9].

It is straightforward to extend our models to other NMR observables such as scalar or dipolar coupling constants. But also other kinds of structural information could be included in the likelihood function as, for example, diffraction data from x-ray crystallography or information on evolutionary relatedness of proteins. A refinement of the conformational prior density would make use of a more realistic force field but should also exploit the knowledge deposited in the structure data bases.

Twenty years ago, Jaynes [32] imagined the following scenario: "Bayesian methods … apply also to a mass of new problems that cannot be formulated at all in orthodox terms; and computers are now … performing very nontrivial data

analysis in such diverse fields as spectrum estimation, medical instrumentation, …, and what will probably become the largest area of application, biological macromolecular structure determination." We hope that our work heads into this direction.

ACKNOWLEDGMENTS

We thank Michael Sattler and Remko Sprangers for kindly providing their NOESY data on the Tudor domain. This work was supported by EU Grant Nos. QLG2-CT-2000-01313 and QLG2-CT-2002-00988.

[1] S. Macura and R. R. Ernst, Mol. Phys. **41**, 95 (1980).
[2] A. T. Brünger and M. Nilges, Q. Rev. Biophys. **26**, 49 (1993).
[3] K. Wüthrich, *NMR of Proteins and Nucleic Acids* (Wiley, New York, 1986).
[4] A. T. Brünger, Nature (London) **355**, 472 (1992).
[5] A. T. Brünger, G. M. Clore, A. M. Gronenborn, R. Saffrich, and M. Nilges, Science **261**, 328 (1993).
[6] P. Güntert, Q. Rev. Biophys. **31**, 145 (1998).
[7] R. T. Cox, *The Algebra of Probable Inference* (John Hopkins University Press, Baltimore, Maryland, 1961).
[8] E. T. Jaynes, *Probability Theory: The Logic of Science* (Cambridge University Press, Cambridge, UK, 2003).
[9] W. Rieping, M. Habeck, and M. Nilges, Science **309**, 303 (2005).
[10] B. Xia, V. Tsui, D. A. Case, H. J. Dyson, and P. E. Wright, J. Biomol. NMR **22**, 317 (2002).
[11] E. T. Jaynes, Phys. Rev. **106**, 620 (1957).
[12] D. Neuhaus and M. P. Williamson, *The Nuclear Overhauser Effect in Structural and Conformational Analysis* (VCH Publishers Inc., New York, 1989).
[13] G. Lipari and A. Szabo, J. Am. Chem. Soc. **104**, 4546 (1982).
[14] The lognormal distribution is defined as

$$\mathrm{LN}(x;\mu,\sigma) = \frac{1}{\sqrt{2\pi\sigma^2}x} e^{-1/2\sigma^2(\log x - \log \mu)^2}$$

where x, μ, and σ are strictly positive quantities. The gamma distribution is defined as

$$\mathrm{G}(x;\alpha,\beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x},$$

where x, α, and β are strictly positive quantities and Γ is the gamma function.
[15] H. Jeffreys, Proc. R. Soc. London, Ser. A **186**, 453 (1946).
[16] M. H. Chen, Q. M. Shao, and J. G. Ibrahim, *Monte Carlo Methods in Bayesian Computation* (Springer-Verlag, New York, 2002).
[17] S. Geman and D. Geman, IEEE Trans. Pattern Anal. Mach. Intell. **6**, 721 (1984).
[18] S. Duane, A. D. Kennedy, B. Pendleton, and D. Roweth, Phys. Lett. B **195**, 216 (1987).
[19] B. Alder and T. Wainwright, J. Chem. Phys. **31**, 459 (1959).
[20] N. Metropolis, M. Rosenbluth, A. Rosenbluth, A. Teller, and E. Teller, J. Chem. Phys. **21**, 1087 (1957).
[21] R. H. Swendsen and J.-S. Wang, Phys. Rev. Lett. **57**, 2607 (1986).
[22] M. Habeck, M. Nilges, and W. Rieping, Phys. Rev. Lett. **94**, 018105 (2005).
[23] U. H. E. Hansmann and Y. Okamoto, Phys. Rev. E **56**, 2228 (1997).
[24] P. Selenko, R. Sprangers, G. Stier, D. Buehler, U. Fischer, and M. Sattler, Nat. Struct. Biol. **8**, 27 (2001).
[25] F. A. Momany, R. F. McGuire, A. W. Burgess, and H. A. Scheraga, J. Phys. Chem. **79**, 2361 (1975).
[26] G. Nemethy, M. A. Pottle, and H. A. Scheraga, J. Phys. Chem. **87**, 1883 (1983).
[27] P. J. Flory, *Statistical Mechanics of Chain Molecules* (Carl Hanser Verlag, Munich, 1969).
[28] W. A. Hendrickson, Methods Enzymol. **115**, 252 (1985).
[29] M. P. Allen and D. J. Tildesley, *Computer Simulation of Liquids* (Clarendon Press, Oxford, 1987).
[30] R. Sprangers, M. Groves, I. Sinning, and M. Sattler, J. Mol. Biol. **327**, 507 (2003).
[31] R. Koradi, M. Billeter, and K. Wüthrich, J. Mol. Graphics **14**, 51 (1996).
[32] E. T. Jaynes (1984), unpublished manuscript: What's wrong with Bayesian methods?, URL http://bayes.wustl.edu/~etj/articles/whatswrong.pdf.