

Replica-Exchange Monte Carlo Scheme for Bayesian Data Analysis

Michael Habeck, Michael Nilges,* and Wolfgang Rieping

Unité de Bioinformatique Structurale, Institut Pasteur 25-28, rue du docteur Roux, 75015 Paris

(Received 4 August 2004; published 11 January 2005)

We develop a sampling algorithm to explore the probability densities arising in Bayesian data analysis problems. Our algorithm is a multiparameter generalization of a replica-exchange Monte Carlo scheme. The strategy relies on gradual weighing of experimental data and on Tsallis generalized statistics. We demonstrate the effectiveness of the method on nuclear magnetic resonance data for a folded protein.

DOI: 10.1103/PhysRevLett.94.018105

PACS numbers: 87.15.Aa

Bayesian methods [1] are becoming increasingly important in physical data analysis problems [2]. The essence of a Bayesian analysis is to base all inferences on unknown parameters θ on a *posterior probability* density function $p(\theta)$. Practically, parameter estimation amounts to the generation of random samples from $p(\theta)$ [3,4]. The sampled states are then used to estimate posterior quantities as sample averages. This procedure is analogous to stochastic simulation of a physical system with configuration θ and potential energy $-\log p(\theta)$ in order to calculate its thermodynamic properties.

According to Bayes' theorem [1] the posterior density is proportional to the product of two non-negative functions: $p(\theta) \propto L(\theta)\pi(\theta)$. The *likelihood function* $L(\theta)$ is the probability for measuring the data if the value of the unknown parameter is θ . The *prior density* $\pi(\theta)$ expresses our knowledge about the parameter before carrying out the experiment. When analyzing data for a thermodynamical system with configuration θ and potential energy function $E(\theta)$ at constant temperature β^{-1} , the canonical ensemble is a natural choice for the prior density $\pi(\theta) \propto \exp\{-\beta E(\theta)\}$. The principle of maximum entropy [5] favors this distribution as being least informative and therefore least biased.

When simulating complex physical systems, such as biological macromolecules, one encounters major difficulties: The canonical ensemble exhibits a complicated topology with jagged modes. The problem is to visit all modes, each with correct population weight, and to not get trapped in a single maximum of the distribution (quasi-ergodicity). Thus, in a Bayesian analysis sampling of the prior $\pi(\theta)$ already poses a formidable problem. But computational difficulties are further aggravated by the additional likelihood factor.

Replica Monte Carlo (MC) simulation [6] circumvents the problem of nonergodic sampling in canonical ensemble simulations. Several noninteracting copies of the system, so-called replicas, are simulated in parallel at different temperatures. Exchanges of configurations between neighboring replicas are accepted according to the Metropolis criterion. By this, configurations diffuse between the high-temperature and the low-temperature heat bath which effectively reduces trapping effects.

In this Letter, we propose an extended replica-exchange Monte Carlo scheme for simulating the posterior densities arising in Bayesian data analysis problems. We first generalize the replica algorithm to families of distributions $f(\theta; \xi)$, where ξ is one or several parameters, analogous to the temperature, which controls the shape of the distribution. We assume that the “low-temperature distribution” ($\xi = \xi_{\max}$) is the one we want to simulate: $f(\theta; \xi_{\max}) \propto p(\theta)$. The family $f(\theta; \xi)$ should be chosen such that it is easy to sample configurations from the “high-temperature distribution” $f(\theta; \xi_{\min})$.

We introduce two parameters $\xi = (\lambda, q)$ to independently control the two factors which make up the posterior density:

$$f(\theta; \lambda, q) = [L(\theta)]^\lambda \pi(\theta; q). \quad (1)$$

The parameter λ weighs the likelihood function and thus determines the influence of the data. The extreme values are $\lambda = 1$, in which case the data are fully taken into account, and $\lambda = 0$, which amounts to neglecting the data completely. In order to further improve sampling, we replace the canonical ensemble $\pi(\theta)$ by Tsallis generalized ensemble [7–10] $\pi(\theta; q) \propto \exp\{-\beta E(\theta; q)\}$, thereby introducing a number q that parametrizes a non-linear transformation of the potential energy:

$$E(\theta; q) = \frac{q}{\beta(q-1)} \log\{1 + \beta(q-1)[E(\theta) - E_{\min}]\}, \quad (2)$$

where E_{\min} is chosen such that $E(\theta) \geq E_{\min}$ holds for all configurations θ . For $E(\theta) > E_{\min}$ and $q > 1$ the transformed energy becomes smoother, which enhances the mobility of the Markov chain. In the low energy regime $\beta(q-1)[E(\theta) - E_{\min}] \ll 1$ the Tsallis ensemble reduces to the Boltzmann ensemble. In particular, it holds that $E(\theta; 1) = E(\theta) - E_{\min}$. The effect of the two replica parameters on the potential energy and on the likelihood function is illustrated in Fig. 1.

We simulate m replicas $f(\theta; \lambda_k, q_k)$ in parallel. The complete system is

$$p(\theta_1, \dots, \theta_m) \propto \prod_{k=1}^m f(\theta_k; \lambda_k, q_k). \quad (3)$$

The acceptance probability for an exchange of configura-

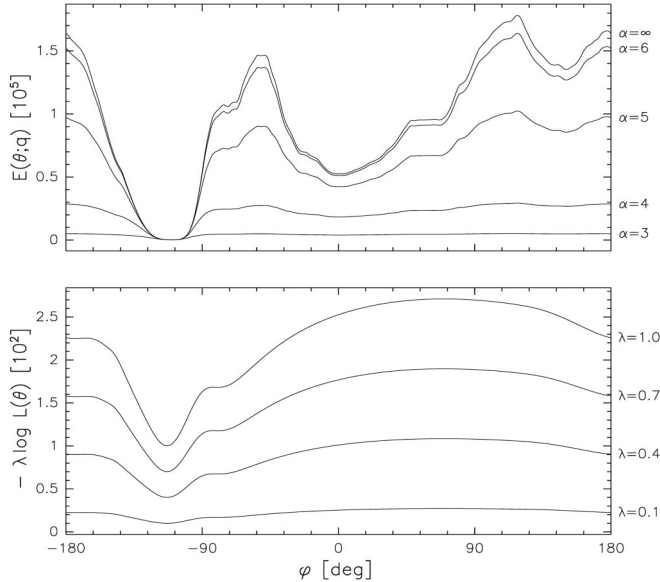


FIG. 1. The effect of the replica parameters q and λ on the prior density and the likelihood function, respectively. In the upper panel, the effective potential energy $E(\theta; q)$ is shown for $q = 1 + 10^{-\alpha}$ varying from 1.0 ($\alpha = \infty$) to 1.001 ($\alpha = 3$). The lower panel shows the effect of weighing the likelihood function by letting λ range from 1.0 to 0.1. The curves are obtained by varying the torsion angle φ in the amino acid Alanine-122 of a Fyn SH3 domain in a folded configuration.

tions between replica k and replica l ,

$$\min \left[1, \frac{f(\theta_k; \lambda_l, q_l) f(\theta_l; \lambda_k, q_k)}{f(\theta_k; \lambda_k, q_k) f(\theta_l; \lambda_l, q_l)} \right], \quad (4)$$

ensures detailed balance. As an advantage to other extended ensemble methods, such as simulated tempering [11] or multicanonical Monte Carlo algorithm [12], the normalization constant $\int d\theta f(\theta; \lambda, q)$ does not need to be known since it cancels out.

We arrange the replicas in a chain, where in the first half q is held constant at value $q_{\min} = 1$, while λ is gradually switched off. In the second half of the chain, λ is held constant at λ_{\min} , while q is increased to its maximum q_{\max} . In practice, $q_{\max} < 2$ suffice to smooth out $E(\theta; q)$ almost completely.

We applied the outlined algorithm to an involved data analysis problem occurring in structural biology. A folded protein constitutes a highly complicated thermodynamical system: the various competing pairwise interactions lead to a rough energy landscape with multiple minima separated by high free energy barriers. Therefore, simulation of the canonical ensemble $\pi(\theta)$ at realistic temperatures already poses a challenging problem [13]. When considering experimental data, the situation becomes more complicated. The likelihood function imposes new “interactions” which severely reduce the mobility of the system.

We parametrize the conformation of the macromolecule in torsion angles; i.e., the covalent topology of the poly-

peptide chain is held fixed. We use the potential energy

$$E(\theta) = \sum_{i < j} \Theta[d_{ij,0} - d_{ij}(\theta)] [d_{ij,0} - d_{ij}(\theta)]^4 \quad (5)$$

as an approximation of the Lennard-Jones potential to describe the pairwise nonbonded interactions [$\Theta(\cdot)$ is the Heaviside step function]. Here, $d_{ij}(\theta)$ denotes the distance between atoms i and j as found in the macromolecule with configuration θ . Covalent parameters and minimum interatomic distances $d_{ij,0}$ are taken from the empirical conformational energy program for peptides (ECEPP/2) [14,15] and the PROLSQ [16] force field, respectively; the temperature was set to 300 K.

In order to determine the molecule’s conformational ground state, we analyzed a set of experimental interproton distances measured by nuclear magnetic resonance (NMR) [17]. The data set consists of n distance measurements $\{d_1, \dots, d_n\}$ which we model as independent. We use a log-normal distribution to describe the observation of a single distance d_i [18,19]. The unknown shape parameter which quantifies the error of the measurements can be eliminated by marginalization [1] yielding the likelihood function

$$L(\theta) = \left(\sum_{i=1}^n \log^2 [d_i / d_i(\theta)] \right)^{-n/2}. \quad (6)$$

Here, $d_i(\theta)$ is the distance corresponding to the i th measurement. The NMR-based distances had been derived from several spectra measured on a deuterated sample of the SH3 domain [20]. Only 154 distances between exchangeable amide protons had been observed. The system comprises 59 amino acids (921 atoms); conformations are parametrized by 275 torsion angles.

We used the algorithm to draw conformational samples from the resulting posterior distribution. Fifty replicas were simulated in parallel. In the first 23 replicas, λ was reduced from $\lambda_{\max} = 1.0$ to $\lambda_{\min} = 0.1$, while q was set to $q_{\min} = 1.0$ (i.e., the prior factor is the Boltzmann ensemble). The values for λ were chosen according to $\lambda_k = [\lambda_{\max} - 0.043(k-1)]^{0.8}$. In the remaining 27 replicas, λ stayed at λ_{\min} , while q was increased to $q_{\max} = 1.1$ following the exponential law $q_k = 0.993 + 0.007 \exp[0.1(k-23)]$. Calculations were performed using a self-written software package [18,19].

Each replica was simulated using a hybrid Monte Carlo (HMC) algorithm [21]. The negative logarithm of the replica ensemble, $-\log f(\theta; \lambda, q) = \beta E(\theta; q) - \lambda \log L(\theta)$, serves as a potential energy. The respective dynamical transition within the HMC algorithm was generated by calculating a dynamics trajectory of 250 integration steps using the leapfrog discretization scheme [22]. Each replica transition was made up of 30 hybrid Monte Carlo steps; 11 500 replica supertransitions were sampled in total.

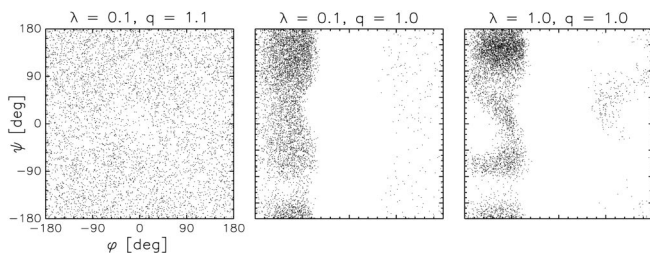


FIG. 2. Samples of the backbone torsion angles φ and ψ of three different replicas: high-temperature heat bath with $\lambda = 0.1$, $q = 1.1$ (left), Boltzmann ensemble with $\lambda = 0.1$, $q = 1.0$ (middle), and posterior distribution with $\lambda = 1.0$, $q = 1.0$ (right).

Figure 2 shows the protein backbone torsion angle samples for the target distribution $f(\theta; 1.0, 1.0)$, the replica at the chain's changeover $f(\theta; 0.1, 1.0)$, and the high-temperature distribution $f(\theta; 0.1, 1.1)$. In the limiting case, $f(\theta; 0, \infty)$ is uniform over the hypercube $[0, 2\pi]^N$ (where N is the number of torsion angles), as the data are practically switched off and the potential energy is nearly flat. In this case, the hybrid Monte Carlo algorithm samples the conformational space ergodically. By decreasing q , noncovalent interactions between the atoms are gradually switched on. The distribution at the changeover is close to the Boltzmann ensemble. Accordingly, the sampled backbone torsion angles closely match the expected distribution for a folded protein. Increasing λ switches on the data and results in the generation of compact conformations.

Figure 3 shows the average radius of gyration for the replicas. In the upper half of the replica chain, conformations are increasingly stretched with decreasing q since the repulsive term of the force field pushes the atoms away from each other. In the lower half of the replica chain, the data pull the conformations together, yielding globular conformational samples that concentrate more and more around the ground state as determined by x-ray crystallography [23]. This can be quantified by comparing the conformational samples with the x-ray structure. With increasing λ and constant q , the distribution of the root mean square deviation (rmsd) between backbone (N, C $^\alpha$, C) coordinates of the conformational samples and the x-ray structure is shifted towards smaller values. We obtain a backbone rmsd between the conformational samples and the ground state of 1.54 ± 0.14 Å. This is a considerable improvement to standard techniques: Molecular dynamics calculations [20] lead to a backbone rmsd of 2.86 ± 0.33 Å.

Figure 4 shows the 20 most probable conformations superimposed onto the x-ray structure. Conformations at the chain's changeover are unordered and distributed according to the canonical ensemble. Because the potential energy in Eq. (5) is purely repulsive and solvent effects are neglected, the most probable conformations are noncompact due to their higher multiplicity. Only with the use of

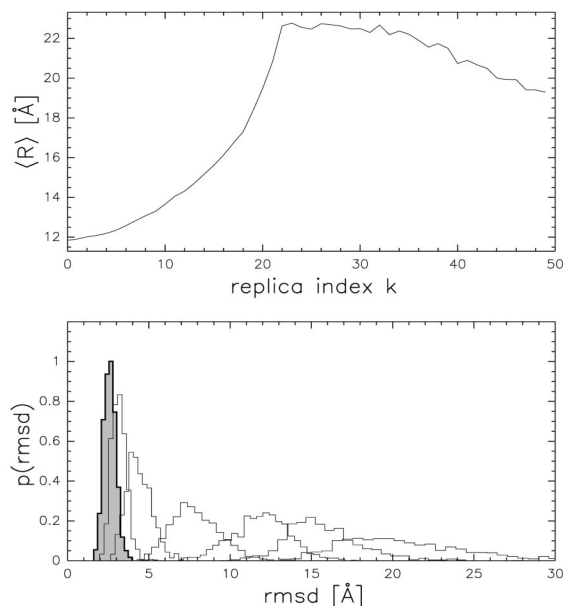


FIG. 3. Upper panel: Dependency of the average radius of gyration $\langle R \rangle$ on the position in the replica chain. The radius of gyration of the x-ray structure is 10.5 Å. Lower panel: Distributions of backbone rmsd values between conformations in the replica ensembles and the x-ray structure. Only replicas with fixed q have been considered. With increasing λ the rmsd distributions are shifted to smaller values. The rmsd histogram for the target distribution is shown as a shaded region.

experimental data can globular conformations get a high probability and exhibit the same fold as the x-ray structure.

One could choose the replica chain differently. The proposed arrangement has the advantage that “evidence” values [1] (corresponding to free energies) can be calculated by thermodynamic integration [3]. Furthermore, our

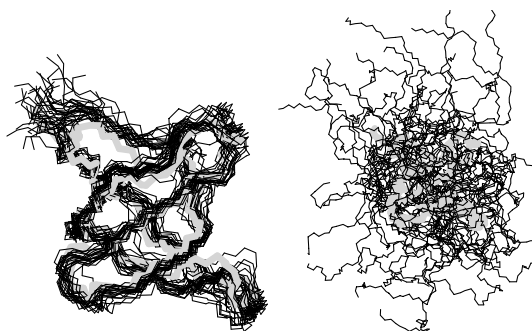


FIG. 4. C $^\alpha$ traces of the 20 most probable conformations, superimposed [24] onto the x-ray structure. Conformations are drawn in black, the backbone trace of the x-ray structure in gray. On the left-hand side, conformational samples of the posterior distribution ($\lambda = 1.0$, $q = 1.0$) are shown. The right-hand side shows structures that are distributed according to the Boltzmann ensemble ($\lambda = 0.1$, $q = 1.0$). Structures on the right-hand side are scaled in size compared to the ensemble on the left.

choice seems natural: The high-temperature part of the chain (q is varied and λ set to its smallest value) produces conformations that show no overlap in van der Waals radii. The low-temperature part (q is set to 1 and λ is increased) selects from the physically meaningful structures those reproducing the data best while maintaining a low potential energy. Instead of using Tsallis' ensemble, one could also vary β in the Boltzmann ensemble. The Tsallis ensemble, however, has the advantage that the overlap between neighboring distributions is higher (only power law instead of exponential decay). Therefore, fewer replicas are needed to span the relevant temperature range.

The scheme is also applicable to data analysis problems with no physical background. In this case, we define the "energy" $E(\theta) = -\log\pi(\theta)/\beta$ with some appropriate "temperature" β^{-1} and apply the same transformations [Eq. (2)] to the pseudoenergy $E(\theta)$. Typically, the prior probability has a finite maximum from which a minimum for the pseudoenergy can be derived.

We outlined a replica-exchange strategy that overcomes the problem of nonergodic sampling in a wide range of data analysis problems. The method is especially suited to systems with a complicated prior structure. Other stochastic sampling algorithms [4], such as HMC, Gibbs sampling, Metropolis MC, or standard replica MC [6], fail in these situations. Molecular dynamics is the most commonly used technique in biomolecular structure determination; however, its poor sampling properties lead to inaccurate and biased results. Our algorithm is capable of sampling the entire parameter space, and by this improves data analyses.

This work was supported by EU Grants No. QLG2-CT-2000-01313 and No. QLG2-CT-2002-00988.

*Electronic address: nilges@pasteur.fr

- [1] E. T. Jaynes, *Probability Theory: The Logic of Science* (Cambridge University Press, Cambridge, UK, 2003).
- [2] G. D'Agostini, Rep. Prog. Phys. **66**, 1383 (2003).
- [3] R. M. Neal, Department of Computer Science, University of Toronto Technical Report No. CRG-TR-93-1, 1993.

- [4] M. H. Chen, Q. M. Shao, and J. G. Ibrahim, *Monte Carlo Methods in Bayesian Computation* (Springer Verlag, Inc., New York, 2002).
- [5] E. T. Jaynes, Phys. Rev. **106**, 620 (1957).
- [6] R. H. Swendsen and J.-S. Wang, Phys. Rev. Lett. **57**, 2607 (1986).
- [7] C. Tsallis, J. Stat. Phys. **52**, 479 (1988).
- [8] U. H. E. Hansmann and Y. Okamoto, Phys. Rev. E **56**, 2228 (1997).
- [9] U. H. E. Hansmann, Chem. Phys. Lett. **281**, 140 (1997).
- [10] T. W. Whitfield, L. Bu, and J. E. Straub, Physica (Amsterdam) **305A**, 157 (2002).
- [11] E. Marinari and G. Parisi, Europhys. Lett. **19**, 451 (1992).
- [12] B. A. Berg and T. Neuhaus, Phys. Rev. Lett. **68**, 9 (1992).
- [13] U. H. E. Hansmann and Y. Okamoto, Curr. Opin. Struct. Biol. **9**, 177 (1999).
- [14] F. A. Momany, R. F. McGuire, A. W. Burgess, and H. A. Scheraga, J. Phys. Chem. **79**, 2361 (1975).
- [15] G. Nemethy, M. A. Pottle, and H. A. Scheraga, J. Phys. Chem. **87**, 1883 (1983).
- [16] W. A. Hendrickson, Methods Enzymol. **115**, 252 (1985).
- [17] K. Wüthrich, *NMR of Proteins and Nucleic Acids* (John Wiley, New York, 1986).
- [18] W. Rieping, M. Habeck, and M. Nilges, in "NMR Analysis of Protein Structure," edited by M. Sattler, M. Nilges, and H. Oschkinat (Springer-Verlag, Heidelberg, to be published)
- [19] M. Habeck, W. Rieping, and M. Nilges, in *Proceedings of the 23rd International Workshop on Bayesian Inference and Maximum Entropy Methods in Science and Engineering*, edited by G. Erickson and Y. Zhai (American Institute of Physics, Melville, NY, 2004), pp. 157–166.
- [20] T. K. Mal, S. J. Matthews, H. Kovacs, I. D. Campbell, and J. Boyd, J. Biomol. NMR **12**, 259 (1998).
- [21] S. Duane, A. D. Kennedy, B. Pendleton, and D. Roweth, Phys. Lett. B **195**, 216 (1987).
- [22] M. P. Allen and D. J. Tildesley, *Computer Simulation of Liquids* (Clarendon Press, Oxford, 1987).
- [23] M. E. Noble, A. Musacchio, M. Saraste, S. A. Courtneidge, and R. K. Wierenga, EMBO J. **12**, 2617 (1993).
- [24] R. Koradi, M. Billeter, and K. Wüthrich, J. Mol. Graphics **14**, 51 (1996).