PROTEINS: Structure, Function, and Bioinformatics 66:61–68 (2007)

Monte Carlo Sampling of Near-Native Structures of Proteins With Applications

Jinfeng Zhang,[1] Ming Lin,[2] Rong Chen,[2,3,4]* Jie Liang,[3]* and Jun S. Liu[1]*
[1]*Department of Statistics, Harvard University, Cambridge, Massachusetts*
[2]*Department of Information and Decision Science, University of Illinois, Chicago, Illinois*
[3]*Department of Bioengineering, University of Illinois, Chicago, Illinois*
[4]*Department of Business Statistics and Econometrics, Peking University, Beijing, People's Republic of China*

ABSTRACT **Since a protein's dynamic fluctuation inside cells affects the protein's biological properties, we present a novel method to study the ensemble of near-native structures (NNS) of proteins, namely, the conformations that are very similar to the experimentally determined native structure. We show that this method enables us to (i) quantify the difficulty of predicting a protein's structure, (ii) choose appropriate simplified representations of protein structures, and (iii) assess the effectiveness of knowledge-based potential functions. We found that well-designed simple representations of protein structures are likely as accurate as those more complex ones for certain potential functions. We also found that the widely used contact potential functions stabilize NNS poorly, whereas potential functions incorporating local structure information significantly increase the stability of NNS. Proteins 2007;66:61–68.** © 2006 Wiley-Liss, Inc.

Key words: near-native structures; sequential Monte Carlo; protein structure simulation; protein structure prediction; structural representation; potential function

INTRODUCTION

Conformational fluctuations of a protein affect its functional roles.[1,2] Treating a protein as a single rigid object can be misleading in tasks such as structure-based drug design[3] and prediction of protein–protein interactions.[4] Instead of using a single structural snapshot as captured by X-ray crystallography, we must consider an ensemble of conformations that collectively describe the native state of a protein.[5,6] A popular approach to capture this intuition is to consider the ensemble of near-native structures (NNS), which is defined as the set of conformations with C_α root-mean-square-deviation (RMSD) less than a threshold value to the native structure (3 Å in this study, unless otherwise stated). Other definitions of NNS will be discussed later.

Several methods have been used for studying NNS, including NMR spectroscopy, molecular dynamics (MD) simulations, Metropolis Monte Carlo,[7] the Gaussian network or elastic network models,[8–10] and chain-growth-based heuristic method.[11] A recent study combines NMR measurements with molecular dynamics simulation to determine the ensemble of protein conformations and associated dynamics.[6] Because of computational limitations, however, NNS as an ensemble of conformations characterizing the native state of a protein is still poorly understood. In this article, we show how to approximate NNS using a sequential Monte Carlo (SMC) approach[12] and how to use these approximations to address the following problems: (i) quantifying the difficulty in predicting proteins of different topology, (ii) choosing an appropriate representation of protein structures for improved efficiency with minimum loss in accuracy, and (iii) comparing different knowledge-based potential functions.

A structural representation is usually evaluated by the proximity of its best-fitted structures to native ones, and a potential function is evaluated by its ability to discern the native structure from decoys.[13–15] However, since a single structure, even if it is the "best," typically cannot represent well the ensemble property of the protein, and a protein's conformational entropy plays an important role in its stability,[16] it is more desirable to evaluate representations and potential functions according to how well they describe the ensemble of NNS.

Let m denote a protein model, which consists of a structural representation and a potential function. The stability of NNS under model m is determined by the change of free energy of the system from the unfolded state to NNS:

$$\Delta G_{\mathrm{NNS},m} = k_{\mathrm{B}}T\ \ln(Q_{\mathrm{U},m}) - k_{\mathrm{B}}T\ \ln(Q_{\mathrm{NNS},m}),$$

Grant sponsor: National Institute of Health; Grant number: GM68958; Grant sponsor: National Science Foundation; Grant numbers: DBI-0133856, DMS-0244541, DMS-0244638; Grant sponsor: NSF China; Grant number: 10228102; Grant sponsor: Whitaker Foundation; Grant number: TF-04-0023.

Jinfeng Zhang and Ming Lin contributed equally to this work.

*Correspondence to: Jun S Liu, Department of Statistics, Harvard University, Cambridge, MA 02143. E-mail: jliu@stat.harvard.edu or Jie Liang, Department of Bioengineering, University of Illinois at Chicago, Chicago, IL 60607. E-mail: jliang@uic.edu or Rong Chen, Department of Information and Decision Science, University of Illinois, Chicago, IL 60607. E-mail: rongchen@uic.edu

Received 28 April 2006; Revised 18 July 2006; Accepted 8 August 2006

Published online 12 October 2006 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/prot.21203

where k_B is the Boltzmann constant, T denotes the temperature, $Q_{NNS,m} = \Sigma_{x \in NNS} e^{(-E_x/k_B T)}$ and $Q_{U,m}$ are the partition functions of the NNS and the unfolded structures, respectively, where E_x is the internal energy of structure x. Since $Q_{NNS,m} + Q_{U,m} = Z = \Sigma_x e^{(-E_x/k_B T)}$, we have:

$$\Delta G_{NNS,m} = -k_B T \ln\left(\frac{P_{NNS,m}}{1 - P_{NNS,m}}\right),$$

where $P_{NNS,m} = Q_{NNS,m}/Z$ is the Boltzmann probability of the NNS. A higher $P_{NNS,m}$ corresponds to a potentially better model. However, having a good estimate of $P_{NNS,m}$ requires good estimates of the partition functions, which are generally difficult to get.[17] Additionally, since the subspace of NNS is a very small portion of the overall space and is also very irregular due to the compact shape of the native structures, Monte Carlo sampling is particularly difficult.[18]

We here take a SMC approach,[12,19] which can be viewed as an extension of the chain growth method explained in Ref. 20. The method generates properly weighted structural configurations of a polypeptide chain by sequentially adding one or a few monomers at a time, just like "growing" a crystal. This general methodology has been followed by several groups in studying protein structures.[21–25] Compared with iterative methods such as the Metropolis algorithm, chain growth methods are especially suitable for generating structural configurations under complicated constraints.

STRUCTURAL MODELS

Discrete *k*-State Off-Lattice Representation of Protein Structures

We use an off-lattice discrete k-state representation for protein structures, in which backbone of each residue is modeled by a pseudo-C_α atom[14,26,27] (Fig. 1). The side-chain of a residue (except glycine) is modeled by one additional atom attached to the backbone C_α atom, which, together with two adjacent C_α atoms, determines the side-chain atom's coordinate. Distances between backbone C_α atoms and their side-chain atoms depend on residue types.[14] There are two radii associated with each residue type: the contact radius and the self-avoiding radius. Contact radius taken from Ref. 14 defines the formation of contacts if the distance of two atoms i and j, $d_{i,j}$, is smaller than the sum of their contact radii. Self-avoiding radius models the excluded volume effect: for C_α atoms it is set as 1.5 Å and for side-chain atoms it is 1.0 Å. In this study, we consider only conformational spaces with self-avoiding structures. We denote this representation as R^k_{SC}, where k indicates the number of discrete states each residue can be positioned. We use R^k_{CA} to denote another representation, in which only the excluded volume effect of C_α atoms is considered (i.e., self-avoiding radii of side-chain atoms are set to zero).

In discrete k-state representations, the backbone position of each residue (excluding the two ends) can take one of k discrete pairs of (α,τ) angles, where α is the pseudo-bond angle formed by three consecutive C_α atoms, and τ is the pseudotorsion angle formed by four consecutive C_α atoms. A study on the choice of k and the associated angle values of (α,τ) was reported in Ref. 27. For the four-state model, the average RMSD between the native protein and its optimal fit (for 978 proteins) is 2.3 Å. For 5, 6, and 8-state models, the average RMSD between the native and the best fit is 1.9, 1.6, and 1.4 Å, respectively.

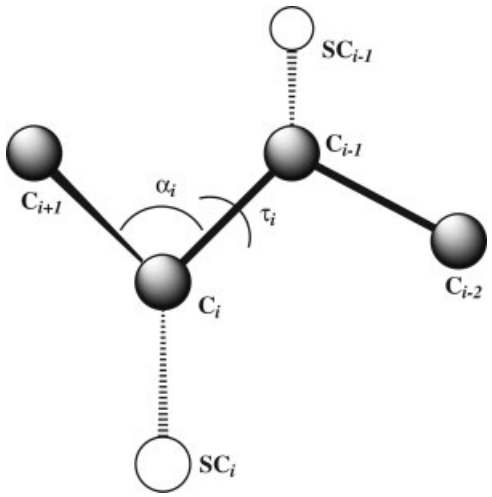

Fig. 1. Discrete state model for protein structures. Bond angle α_i at position i is formed by C_{i-1}, C_i, and C_{i+1}. Torsion angle T_i is the dihedral angle of the two planes formed by atoms (C_{i-2}, C_{i-1}, C_i) and (C_{i-1}, C_i, C_{i+1}).

Potential Functions

Following potential functions will be considered in this study to demonstrate the usefulness of the concept of NNS for evaluating protein models.

i. *Uniform potential (UP):* It takes the form $E_{UP} = 0$. The corresponding Boltzmann distribution is simply the uniform distribution and the partition function is the total number of self-avoiding conformations.

ii. *Contact potential (CP):* This is perhaps the most widely used knowledge-based potential function,[28–33] and takes the general form:

$$E_{CP} = \sum_i^N \sum_j^N e_{i,j} I(i,j),$$

where i and j are atom (or residue) index, $e_{i,j}$ is the contact energy of atoms i and j, and $I(i,j)$ indicates whether i and j are in contact. A recent survey of 29 such CPs concluded that they can be organized into two classes so that the potential functions within each class are highly similar to each other.[34] There is also a strong correlation between these two classes. The potential function used in this study, referred to as HLPL in Ref. 34 belongs to the larger class of the two.

iii. *Local structure potential (LSP):* Its functional form is:

$$E_{\mathrm{LSP}} = \sum_{i=2}^{N-1} e'_{i,s_i},$$

where i is the residue number from 2 to $N-1$, s_i is the actual state of that residue, and $e'_{i,s}$ is a function reflecting the local structural "preference" of that residue. To demonstrate that our NNS criterion can indeed help evaluate potential functions, we here choose an artificial function expected to be helpful. We let

$$e'_{i,s} = -\log(p_{i,s}),$$

where $p_{i,s}$ is the probability of state s at position i, which is estimated from the ensemble of uniformly sampled NNS. Since this probability is estimated from the true native structure, the LSP we adopted here should reflect the information of the true local structure. To compute our LSP for a protein with a given native structure, we estimate the values of p_{i,s_i} from 10,000 NNS samples generated using SMC under potential UP.

iv. *Contact and local structure potential (CALSP):* It is simply the sum of CP and LSP[27]:

$$E_{\mathrm{CALSP}} = E_{\mathrm{CP}} + E_{\mathrm{LSP}},$$

which aims to study how a combination of global and local interactions in the potential function can improve the probability of NNS.

SMC for Generating NNS

Our goal here is to sample conformations that both follow the Boltzmann distribution defined by an energy function and are also close to a given native structure. This gives rise to two constraints: conformational constraint and energetic constraint. Previous applications of SMC deal mainly with one of the two constraints.

In our formulation, a protein's structure is determined by its residues' states (the bond and torsion angles), denoted as $S_n = (s_2, \ldots, s_{n-1})$, where n is the length of the sequence, and s_i takes values in $\{1, \ldots, k\}$, representing one of the k possible states of residue i. We let $\Omega_t = \{S_t\}$ denote a set of self-avoiding conformations of length t in our discrete model space. Here we are interested in estimating

$$Z(\Omega_n) = \sum_{S_n \in \Omega_n} h(S_n), \qquad (1)$$

where $h(\cdot)$ is an arbitrary function of interest. Our SMC strategy for approximating $Z(\Omega_n)$ is a combination of sequential importance sampling and optimal resampling.[12,23,35] In importance sampling, one generates a set of configurations according to a distribution $g(\cdot)$ and then estimates $Z(\Omega_n)$ by weighted average of all samples. A key challenge in using importance sampling is to design a good sampling distribution. An effective approach as first explored by Rosenbluth and Rosenbluth [20] is to construct $g(\cdot)$ sequentially. That is, one "grows" the molecule by sampling one residue a time conditional on the configuration of the previously grown residues; and then weighs the final configuration according to the importance sampling principle. However, this strategy is usually not enough to produce a good estimate of $Z(\Omega_n)$ since the "attrition" becomes very serious when the chain becomes moderately long. To overcome the attrition problem, researchers introduced the pruning and enrichment ideas.[21,36] Recently, researchers in statistics and engineering introduced an alternative strategy-resampling,[23,37] which is more flexible and efficient than the pruning-enrichment approach in many cases.[12,23,37] The following scheme illustrates our algorithm for sampling NNS.

1. We set the initial sample size to $m_1 = 1$, with weight $w_1^{(1)} = 1$. The starting configuration contains the first two residues. Suppose that at step $t-1$, we have m_{t-1} partial configurations with corresponding weights, denoted as $\{(S_{t-1}^{(j)}, w_{t-1}^{(j)}), j = 1, \ldots, m_{t-1}\}$.
2. Chain growth. For each partially grown configuration $S_{(t-1)}^{(j)}$, we exhaustively test all possible attachments of the protein's next residue (a total of k different states), which will generate no greater than k different partial configurations of length t, $\overline{S}_t^{(j,l)} = (S_{(t-1)}^{(j)}, s_t)$, with temporary weights $\overline{w}_t^{(j,l)} = w_{(t-1)}^{(j)}$ (the chain will be removed from further consideration if $\overline{S}_t^{(j,l)} \notin \Omega_t$). We denote all the samples such generated as $\{(\overline{S}_t^{(l)}, \overline{w}_t^{(l)}), l = 1, \ldots, L\}$ (clearly $L \leq km_{t-1}$).
3. Resampling. If $L \leq m$, the upper bound of Monte Carlo sample size, we keep all of the samples and their corresponding weights and set $m_t = L$. If $L > m$, we use the optimal resampling procedure of Fearnhead and Clifford[35,50] to choose $m_t = m$ distinct samples from them with marginal probabilities proportional to a set of priority scores $\beta_t^{(l)}$. The steps of this resampling procedure are as follows:
   a. Solve the constant c such that $\Sigma_{l=1}^{L} \min\{1, c\beta_t^{(l)}\} = m$.
   b. Choose a subset of distinctive members $J_1, J_2, \ldots, J_m$ from the set $\{1, \ldots, L\}$ so that the marginal probability for the l to be selected is equal to $p_l = \min\{c\beta_t^{(l)}, 1\}$. One way to achieve this is to (i) draw $U_0 \sim \mathrm{Unif}[0,1]$, and let $U_j = j - U_0$, for $j = 1, \ldots, m$; and (ii) choose $J_j = l$ if $p_0 + \cdots + p_{l-1} < U_j \leq p_1 + \cdots + p_l$, for $l = 1, \ldots, L$ and $P_0 = 0$.
   c. Let $S_t^{(j)} = \overline{S}_t^{(J_j)}$, and update the new weight as
$$w_t^{(j)} = \bar{w}_t^{(J_j)} / \min\{c\beta_t^{(J_j)}, 1\}.$$
4. When the target length n is reached, $Z(\Omega_n)$ is estimated by $\Sigma_{j=1}^{m_n} w_n^{(j)} h(S_n^{(j)})$, where m_n is the number of samples at length n, and $w_n^{(j)}$ is the importance weight of sample $S_n^{(j)}$.

An advantage of the above resampling method over the previous SMC method[23] and the pruning-enrichment approach[21] is that it guarantees to generate distinctive configurations, yet without losing information.[35]

The priority score $\beta_t(S_t)$ can be understood intuitively as a measure of the chain's "growth perspective", and is used to encourage the growth of chain S_t to certain direc-

tions. We can design different $\beta_t(S_t)$ for different $h(\cdot)$. For example, to estimate the partition function of NNS, we have

$$h(S_n) = I\{S_n : \mathrm{RMSD}(S_n, NS) < 3\}e^{-E(S_n)/T},$$

where RMSD(S_n, NS) denotes the RMSD between conformation S_n and the native structure and $E(S_n)$ is the energy of conformation S_n defined by the potential function and T is set to 1. In this case, we use the priority score

$$\beta_t^l(S_t^{(l)}) = I(S_t^{(l)} \in \Omega_t)e^{-(E(S_t^{(l)}) + \mathrm{RMSD}(S_t^{(l)}, NS_t))/T_t},$$

where $S_t^{(l)}$ is one of L sampled structures of first t residues, NS_t is the corresponding partial native structure, and $E(S_t^{(l)})$ is the energy of the conformation $S_t^{(l)}$. This priority score is based on both the energy of the conformation $S_t^{(l)}$ and its RMSD to the partial native structure NS_t. It encourages conformations that are both of low energy and similar to the native structure. The auxiliary temperature parameter T_t is used to control our reliance on the priority score function, and is set as

$$T_t = \rho\sqrt{n/t}$$

where ρ is set to 0.05 in all cases. Hence, we have a higher temperature at the beginning of growth to induce flexibility, and a cooling down as t increases to enforce the constraint.

RESULTS

Validation of the SMC Method

To evaluate the performance of our SMC method, we compared SMC estimates of several structural properties of a few polypeptide sequences to exact answers obtained by exhaustive enumerations. Subchains of length L of the following eight small proteins are used in these comparisons: **1ail**, α; **1nkd**, α; **1fna**, β; **1mjc**, β; **1vie**, β; **1pft**, coils; **1vcc**, α and β; **2igd**, α and β. These are chosen from the set of 70 nonhomologous proteins with different folds selected in Ref. 38. The running time for generating $m =$ 10,000 distinctive conformations using SMC for a protein of length 100 is about 20 min on a 0.5 GHz Linux machine.

Figure 2(a,b) present results for approximating the partition functions of NNS for subchains of lengths $11 \leq J \leq 15$ for proteins **1mjc** (under the uniform potential) and **1nkd** (under the contact potential). We can see that the SMC estimates with $m = 10{,}000$ are indistinguishable from the exact answers. Figure 2(c) shows estimates of partition functions of all conformations (unconstrained) under both potentials for subchains of **1ail**. Figure 2(d) displays the estimates of probabilities of NNS under the Boltzmann distribution for three proteins.

Finally, Figure 2(e,f) demonstrate our SMC estimates of probabilities of native contacts for two sets of NNS (defined by different RMSD ranges) of **1nkd** at length 15. Contacts are formed by two atoms when their distance is smaller than the sum of their contact radii. The probability of a native contact in NNS is defined as the number of NNS with that kind of contacts over the total number of NNS. Estimating the probability for a certain contact is challenging since it requires us to estimate both the total number of NNS conformations and the size of the subset of NNS containing that contact.

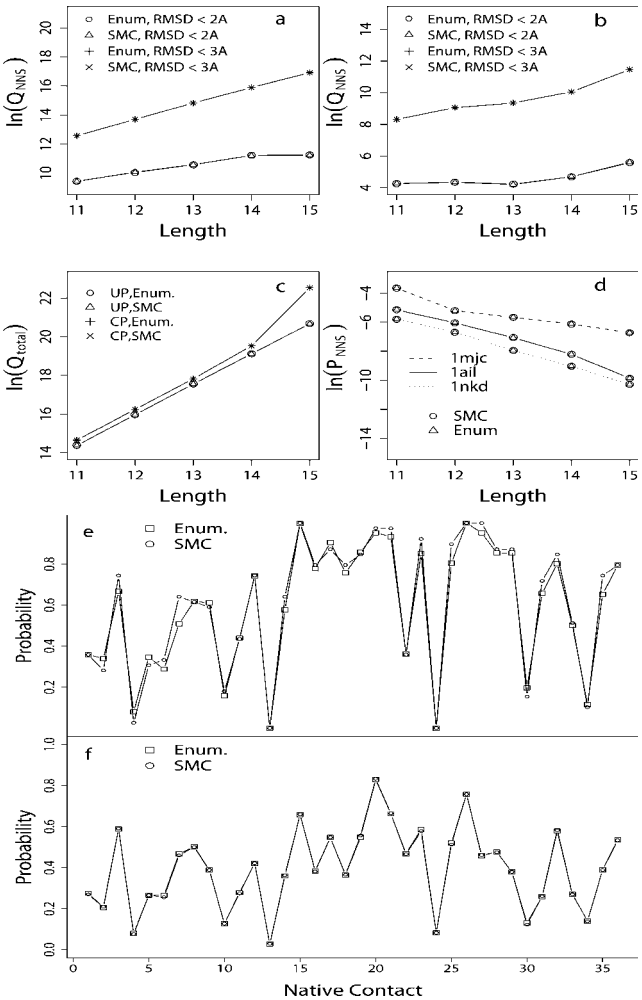

Fig. 2. Comparison of SMC estimations and exact values for: (**a**) Partition functions of NNS (Q_{NNS}) for two sets of NNS defined by RMSD cutoffs of 2 and 3 Å, respectively, under UP for protein **1mjc**; (**b**) Same as above under a contact potential (CP) for protein **1nkd**; (**c**) Partition functions of all conformations under both UP and CP for protein **1ail** from length 11 to 15; (**d**) The probability of NNS under CP with 3 Å RMSD for these proteins (**1ail**, **1mjc**, and **1nkd**) at lengths from 11 to 15; (**e**) Probabilities of native contacts of NNS in the range of 1.0 Å $<$ RMSD $<$ 2 Å of protein **1nkd**; (**f**) Same as above in the range of 2.0 Å $<$ RMSD $<$ 3 Å. For (e) and (f), we define a pair of residue to have a "native contact" if they contact with each other in the native structure, and the *x*-axis in these plots indexes these contacts.

Quantifying the Difficulty of a Protein Structure Prediction Problem

We can measure the level of difficulty of a protein structure prediction problem by the probability that a randomly chosen structure lies in the NNS of this pro-

tein. Here, we consider three sets of NNS determined by RMSD cutoff values: 3 Å (NNS3), 4 Å (NNS4), and 5 Å (NNS5), respectively. The size of the entire conformational space, and the size of NNS are estimated separately, both under the uniform potential function for representation R^5_{CA}. Figure 3 displays average probabilities of NNS for a set of 46 proteins with lengths ranging from 50 to 150. The 46 proteins are grouped according to their lengths with a 5-residue interval, and averages of each group are plotted. We can see that the log-probability of NNS decreases linearly with the chain length at a slope near -1. For example, for protein **1amx** with 150 residues, the log-probability of random sampling a structure within 3 Å RMSD to native structure is -157.7 ± 0.7.

Comparing Effectiveness of Structural Representations

A challenge in protein structure modeling is to find a quantitative structural representation that allows for both accurate modeling of important physical interactions and efficient computation. Discrete state representations of proteins, which are the focus of this article, provide a significant advantage in computation,[25,27,39,40] albeit at a cost of reduced accuracy. In these models, the number of discrete states, the values of α and τ angles, the contact radius, and self-avoiding radius of atoms, and so forth, are all adjustable parameters, of which some can be obtained from studying databases of native protein structures, but others cannot be easily determined. Using two discrete state representations, R^k_{SC} and R^k_{CA}, as an example, we show here how one can compare the effectiveness of different representations based on their abilities to stabilize NNS (i.e., the probabilities of NNS).

We first estimate P_{NNS} under the uniform potential for representations R^k_{CA} with $k = 4$, 5, 6, and 8 for the same eight small proteins used previously. The estimated log-partition functions for fragments of the proteins (starting from residue one with lengths ranging from 13 to 50) for R^k_{CA} are shown in Figure 4(a). The estimated $\log(P_{NNS})$ for segments of lengths ranging from 13 to the full length of these proteins are shown in Figure 4(b). It is interesting to observe that although the total number of conformations increases significantly with the increase of k, the estimated probability of NNS is nearly invariant under different representations. Similar results were observed under contact potentials. It has two implications: one is that using a simpler representation is perhaps desirable and justifiable, and the other is that the difficulty of the protein prediction task for models with a larger number of states (or even continuous model) may not be very different from what is shown here.

We then compared the estimated P_{NNS} under two representations, R^5_{SC} and R^5_{CA}, for understanding the effect of modeling side-chains. Figure 5(a) shows that R^5_{SC} and R^5_{CA} have very similar P_{NNS} under UP for 23 proteins with fewer than 100 residues. But for CP, R^5_{CA} has a much lower P_{NNS} than R^5_{SC}, implying that the side-chain representation of R^5_{SC} is more desirable for protein modeling.

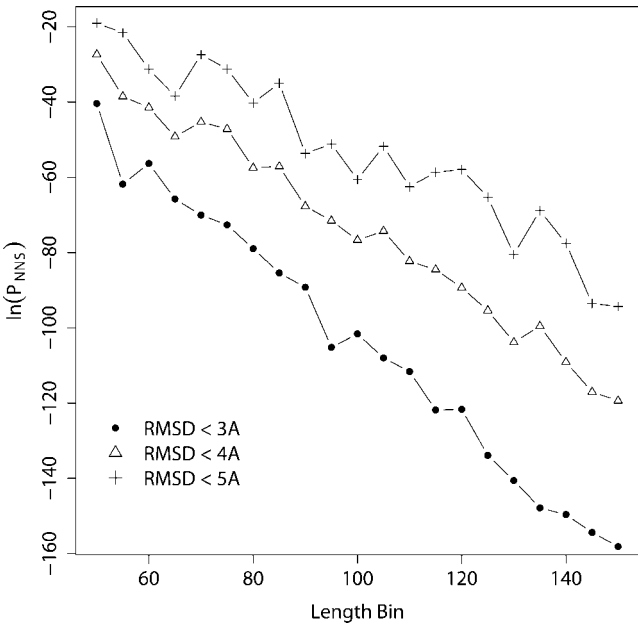

Fig. 3. Estimated probabilities of NNS defined by three RMSD cutoff values (3, 4, and 5 Å, respectively) for 61 proteins with lengths ranging from 31 to 150 residues. Error bars are too small to show on the figure. An example with the standard deviation is given in the text.

Comparing Potential Functions for a Given Representation

Potential functions for successful structure predictions must meet a basic requirement that the defined energy landscape stabilizes the ensemble of NNS among all other competing structures. Thus, we use P_{NNS} again as a criterion to evaluate potential functions. To illustrate this approach, we compare potential functions UP, CP, LSP, and CALSP under the same representation R^5_{SC}.

We estimated P_{NNS} for a set of 23 proteins that are no larger than 100 residues. Figure 5(b) displays the comparison of P_{NNS} under four potential functions. We can see that the P_{NNS} values are similar under CP and UP, and are significantly lower than those obtained under LSP and CALSP. This indicates that NNS are stabilized poorly by CP for structural representation model R^5_{SC}. Potential function CALSP, which combines both local structure and contact information, results in a slightly smaller P_{NNS} than LSP for small proteins, but have slightly higher P_{NNS} for larger proteins, which is probably due to the fact that long proteins have more core residues forming contacts. It can also be seen that P_{NNS} decreases much slower with protein length under LSP and CALSP than under UP and CP.

DISCUSSION

We have developed a new Monte Carlo technique for estimating quantitative properties of NNS. The effectiveness of this method was validated by comparisons with exact answers for small polypeptide chains and also by a systematic study of properties of NNS. To the best of our

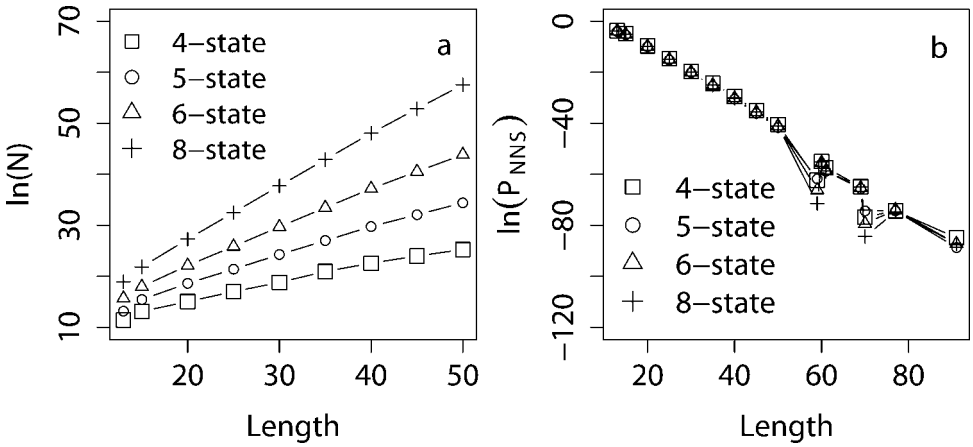

Fig. 4. Average properties estimated for eight small proteins. (**a**) Estimated numbers of conformations (*N*) for models with 4, 5, 6, and 8 discrete states, respectively, with partial chain lengths ranging from 13 to 50; (**b**) Estimated probabilities of NNS for models with 4, 5, 6, and 8 states, respectively, with chain lengths ranging from 13 to the full length of each protein.

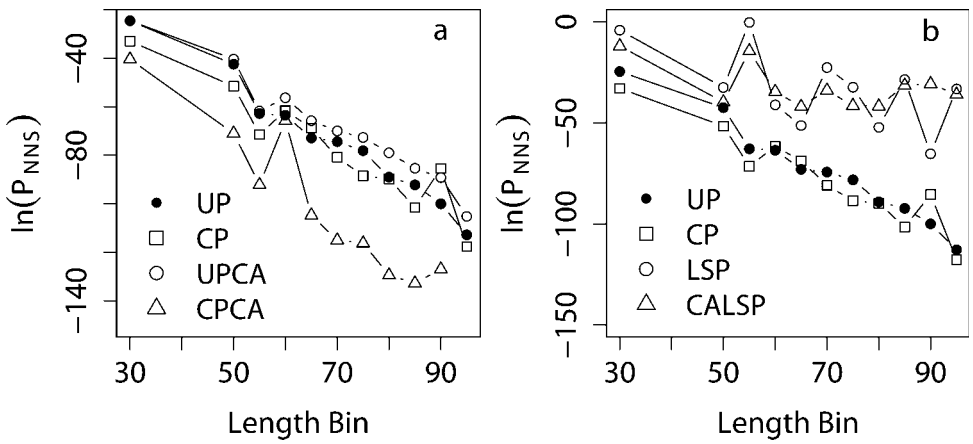

Fig. 5. Comparison of different representations and potential functions for 23 proteins with lengths ranging from 31 to 99 residues using 5-state representations. (**a**) Estimated probability of NNS for UP and CP using two different representations, R^5_{SC} (denoted by UP and CP, respectively, in the figure) and R^5_{CA} (denoted by UPCA and CPCA); (**b**) Estimated probability of NNS for different potential functions using R^5_s. UP, uniform potential; CP, contact potential; LSP, local sequence-structure potential; CALSP, contact and local sequence-structure potential.

knowledge, this level of accuracy has not been achieved in previous studies, especially for free energy and entropy estimations.[17,41]

The problem of assessing the difficulty of protein structure prediction has been examined previously using known protein structures or through curve fitting and extrapolation of simulated data.[42–45] Using the probability of NNS (P_{NNS}) to quantify a protein's "difficulty", we observed [Fig. 4(b)], to our surprise, that this "difficulty" measure is not affected by the number of states used in a discrete-state structural representation, but is significantly affected by side-chain representations: P_{NNS} is much higher for R^5_{SC} than for R^5_{CA} (i.e., favoring the side-chain representation of R^5_{SC}) when a contact potential was employed. Although structures represented by R^k_{CA} may have lower RMSD to native structure than those of R^k_{SC} due to the side-chain excluded volume effect, the excluded volume effect of side-chains may play a positive role in favoring NNS since it also prevents some nonnative conformations being packed too tightly, with a lower energy than NNS.

Using the same criterion, we compared four potential functions, and observed that the pair-wise contact potential employed is not suitable for studying protein folding, a conclusion similar to that given in Refs. 46 and 47. We found that samples generated under such contact potential function are often non-native, even though they have lower energy values than NNS. Our method can also help improve potential functions, since an analysis of structural features of the Monte Carlo samples we generated both in NNS and in general space can provide an informative diagnosis of the problematic areas of a potential function. For example, from the samples of low energy non-native structures, we find that most of them,

although very compact, lack protein-like local structures, suggesting the need for incorporating more accurate descriptions of local structures. As a proof of principle, we created two artificial potential functions, LSP and CALSP, and found that they indeed improved the stability of NNS dramatically.

Finally, we note some limitations and possible extensions of the current method and models. Despite their importance, NNS ensembles are not as well-defined experimentally as native structures. NNS defined in this study are those within certain RMSD range to a native structure. Different similarity measures can be used.[48] For example, in cases where a protein experiences large movement, its alternative active structures may have large RMSD to the native structure, but most of the local structures remain similar. In such cases, a local similarity measure can be used to define the set of NNS. In some studies, NNS needs to be generated only for an interested part of a native structure while keeping other parts of the structure fixed. Algorithms have been developed to sample internal sections of a chain molecule, such as the Internal configurational biased Monte Carlo (ICB) algorithm,[49] which can be incorporated into the SMC framework. Although a simplified model of protein structures is used in this study, our method can be applied to all-atom models. While the conformational space for large proteins become inhibitive, NNS can be generated using all-atom representations for certain functionally important regions of the target protein, which may provide us more information than the single native structure for studying structure-based drug design and protein–protein interactions.

REFERENCES

1. Frauenfelder H, Sligar S, Wolynes P. The energy landscapes and motions of proteins. Science 1991;254:1598–1603.
2. Dill K, Chan H. From levinthal to pathways to funnels. Nat Struct Biol 1997;4:10–19.
3. Carlson H. Protein flexibility and drug design: how to hit a moving target. Curr Opin Chem Biol 2002;6:447–452.
4. Goh C, Milburn D, Gerstein M. Conformational changes associated with protein–protein interactions. Curr Opin Struct Biol 2004; 14:104–109.
5. Gerstein M, Krebs W. A database of macromolecular motions. Nucleic Acids Res 1998;26:4280–4290.
6. Larsen K, Best R, Depristo M, Dobson C, Vendruscolo M. Simultaneous determination of protein structure and dynamics. Nature 2005;433:128–132.
7. Shimada J, Kussell E, Shakhnovich E. The folding thermodynamics and kinetics of crambin using an all-atom Monte Carlo simulation. J Mol Biol 2001;308:79–95.
8. Yang L, Liu X, Jursa C, Holliman M, Rader A, Karimi H, Bahar I. iGNM: A database of protein functional motions based on Gaussian network model. Bioinformatics 2005;21:2978–2987.
9. Tirion M. Large amplitude elastic motions in proteins from a single-parameter, atomic analysis. Phys Rev Lett 1996;77:1905–1908.
10. Bahar I, Atilgan A, Erman B. Direct evaluation of thermal fluctuations in proteins using a single-parameter harmonic potential. Fold Des 1997;2:173–181.
11. DePristo M, de Bakker PIW, Lovell S, Blundell T. Ab initio construction of polypeptide fragments: efficient generation of accurate, representative ensembles. Proteins 2003;51:41–55.
12. Liu J, Chen R. Sequential Monte Carlo methods for dynamic systems. J Am Stat Assoc 1998;93:1032–1044.
13. Park B, Levitt M. The complexity and accuracy of discrete state models of protein structure. J Mol Biol 1995;249:493–507.
14. Park B, Levitt M. Energy functions that discriminate X-ray and near native folds from well-constructed decoys. J Mol Biol 1996; 258:367–392.
15. Tobi D, Shafran G, Linial N, Elber R. On the design and analysis of protein folding potentials. Proteins 2000;40:71–85.
16. Shortle D, Simons K, Baker D. Clustering of low-energy conformations near the native structures of small proteins. Proc Natl Acad Sci USA 1998;95:11158–11162.
17. Reinhardt W, Miller M, Amon L. Why is it so difficult to simulate entropies, free energies, and their differences? Acc Chem Res 2001; 34:607–614.
18. Hansmann U, Okamoto Y. New Monte Carlo algorithms for protein folding. Curr Opin Struct Biol 1999;9:177–183.
19. Chen R, Liu J. Mixture Kalman filters. J R Stat Soc B 2000;62: 493–509.
20. Rosenbluth M, Rosenbluth AW. Monte Carlo calculation of the average extension of molecular chains. J Chem Phys 1955;23:356–359.
21. Grassberger P. Pruned-enriched rosenbluth method: simulations of theta polymers of chain length up to 1 000 000. Phys Rev E 1997; 56:3682–3693.
22. Gan H, Tropsha A, Schlick T. Generating folded protein structures with a lattice growth algorithm. J Chem Phys 2000;113: 5511–5524.
23. Zhang J, Liu J. A new sequential importance sampling method and its application to the two-dimensional hydrophobic–hydrophilic model. J Chem Phys 2002;117:3492–3498.
24. Liang J, Zhang J, Chen R. Statistical geometry of packing defects of lattice chain polymer from enumeration and sequential Monte Carlo method. J Chem Phys 2002;117:3511–3521.
25. Zhang J, Chen Y, Chen R, Liang J. Importance of chirality and reduced flexibility of protein side chains: a study with square and tetrahedral lattice models. J Chem Phys 2004;121:592–603.
26. Zhang J, Chen R, Tang C, Liang J. Origin of scaling behavior of protein packing density: a sequential Monte Carlo study of compact long chain polymers. J Chem Phys 2003;118:6102–6109.
27. Zhang J, Chen R, Liang J.Empirical potential function for simplified protein models: combining contact and local sequence-structure descriptors. Proteins 2006;63:949–960.
28. Hendlich M, Lackner P, Weitckus S, Floeckner H, Froschauer R, Gottsbacher K, Casari G, Sippl M. Identification of native protein folds amongst a large number of incorrect models. The calculation of low energy conformations from potentials of mean force. J Mol Biol 1990;216:167–180.
29. Miyazawa S, Jernigan R. Residue–residue potentials with a favorable contact pair term and an unfavorable high packing density term, for simulation and threading. J Mol Biol 1996;256: 623–644.
30. Zhou H, Zhou Y. Distance-scaled, finite ideal-gas reference state improves structure-derived potentials of mean force for structure selection and stability prediction. Protein Sci 2002;11:2714–2726.
31. Li X, Hu C, Liang J. Simplicial edge representation of protein structures and alpha contact potential with confidence measure. Proteins 2003;53:792–805.
32. Hao M, Scheraga H. Designing potential energy functions for protein folding. Curr Opin Struct Biol 1999;9:184–188.
33. Bastolla U, Vendruscolo M, Knapp E. A statistical mechanical method to optimize energy functions for protein folding. Proc Natl Acad Sci USA 2000;97:3977–3981.
34. Pokarowski P, Kloczkowski A, Jernigan R, Kothari N, Pokarowska M, Kolinski A. Inferring ideal amino acid interaction forms from statistical protein contact potentials. Proteins 2005;59: 49–57.
35. Fearnhead P, Clifford P. On-line inference for hidden Markov models via particle filters. J R Stat Soc B 2003;65:887–899.
36. Wall F, Erpenbeck J. New method for the statistical computation of polymer dimensions. J Chem Phys 1959;30:634–637.
37. Liu JS. Monte Carlo strategis in scientific computing. New York: Springer; 2001.
38. Fain B, Xia Y, Levitt M. Design of an optimal Chebyshev-expanded discrimination function for globular proteins. Protein Sci 2002;11:2010–2021.
39. Gordon T, Brown S. Minimalist models for protein folding and design. Curr Opin Struct Biol 2003;13:160–167.
40. Kolinski A, Skolnick J. Reduced models of proteins and their applications. Polymer 2004;45:511–524.

41. Kollman P. Free energy calculations: applications to chemical and biochemical phenomena. Chem Rev 1993;93:2395–2417.
42. Reva B, Finkelstein A, Skolnick J. What is the probability of a chance prediction of a protein structure with an rmsd of 6 Å? Fold Des 1998;3:141–147.
43. Sullivan D, Kuntz I. Distributions in protein conformation space: implications for structure prediction and entropy. Biophys J 2004; 87:113–120.
44. Cohen F, Sternberg M. On the prediction of protein structure: the significance of the root-mean-square deviation. J Mol Biol 1980; 138:321–333.
45. Maiorov V, Crippen G. Significance of root-mean-square deviation in comparing three-dimensional structures of globular proteins. J Mol Biol 1994;235:625–634.
46. Thomas P, Dill K. Statistical potentials extracted from protein structures: how accurate are they? J Mol Biol 1996;257:457–469.
47. Hu C, Li X, Liang J. Developing optimal non-linear scoring function for protein design. Bioinformatics 2004;20:3080–3098.
48. Wallin S, Farwer J, Bastolla U. Testing similarity measures with continuous and discrete protein models. Proteins 2003;50:144–157.
49. Uhlherr A. Monte Carlo conformational sampling of the internal degrees of freedom of chain molecules. Macromolecules 2000;33: 1351–1360.
50. Carpenter J, Clifford P, Fearnhead P. An improved particle for non-linear problems. IEEE Proc Radar Sonar Navign 1999;146: 2–7.