Generating properly weighted ensemble of conformations of proteins from sparse or indirect distance constraints

Ming Lin,[1] Hsiao-Mei Lu,[2] Rong Chen,[3] and Jie Liang[2,a)]
[1]*Department of Information and Decision Science, University of Illinois at Chicago, 845 S. Morgan St., Chicago, Illinois 60607, USA*
[2]*Department of Bioengineering, University of Illinois at Chicago, 845 S. Morgan St., Chicago, Illinois 60607, USA*
[3]*Department of Statistics, Rutgers University, 110 Frelinghuysen Rd., Piscataway, New Jersey 08854-8019, USA*

(Received 30 April 2008; accepted 16 July 2008; published online 2 September 2008)

Inferring three-dimensional structural information of biomacromolecules such as proteins from limited experimental data is an important and challenging task. Nuclear Overhauser effect measurements based on nucleic magnetic resonance, disulfide linking, and electron paramagnetic resonance labeling studies can all provide useful partial distance constraint characteristic of the conformations of proteins. In this study, we describe a general approach for reconstructing conformations of biomolecules that are consistent with given distance constraints. Such constraints can be in the form of upper bounds and lower bounds of distances between residue pairs, contact maps based on specific contact distance cutoff values, or indirect distance constraints such as experimental ϕ-value measurement. Our approach is based on the framework of sequential Monte Carlo method, a chain growth-based method. We have developed a novel growth potential function to guide the generation of conformations that satisfy given distance constraints. This potential function incorporates not only the distance information of current residue during growth but also the distance information of future residues by introducing global distance upper bounds between residue pairs and the placement of reference points. To obtain protein conformations from indirect distance constraints in the form of experimental ϕ-values, we first generate properly weighted contact maps satisfying ϕ-value constraints, we then generate conformations from these contact maps. We show that our approach can faithfully generate conformations that satisfy the given constraints, which approach the native structures when distance constraints for all residue pairs are given.
 [DOI: 10.1063/1.2968605]

I. INTRODUCTION

Three-dimensional structures of biomacromolecules (such as protein, DNA, and RNA molecules) are essential for understanding their biological functions. The primary sources of structural information of biomacromolecules are x-ray diffractions[1] and nuclear magnetic resonance (NMR) experiments.[2,3] In NMR studies, the assessment of chemical shifts of atomic nuclei with spins can provide information about distances between specific pairs of atoms. In addition, a number of biochemical techniques such as disulfide linking[4,5] and electron paramagnetic resonance labeling[6,7] can also provide useful partial information about distances between residues. When accurate values of distances between residue pairs are available, they can be represented by a *distance matrix*, where an entry (i,j) of the matrix denotes the distance between the corresponding two residues i and j. In other cases, the distance information is inexact, but can be represented by a *contact map*, which denotes whether the distance of each residue pair is below or above a specified distance threshold.

With distance information of various levels of details, a challenging problem is to integrate all the distance constraints into global information about the properties of ensemble structures of the biomacromolecule.[2,8,9] When the distance constraints are complete and accurate, the positions of all residues can be obtained by carrying out a singular value decomposition calculation.[10,11] When the distance constraints are incomplete or inaccurate, one needs to solve an optimization problem and find the structures that best reproduces these distance constraints.[9,12,13] If the size of the molecule is large, this is a very challenging problem.

Several previous studies address the problem of generating protein conformations from contact maps.[14,15] These approaches can be expanded to generate conformations when only indirect or implicit distance constraints are available. In Ref. 16, Vendruscolo *et al.* generated the conformations of the transition state ensemble (TSE) important in protein folding studies. In this case, no explicit distance constraints between residue pairs are given. Rather, indirect information in the form of ϕ-value constraints is known for a subset of residues. Here the ϕ-value at a residue measured experimentally is interpreted as the ratio of the average number of

a)Author to whom correspondence should be abbressed. Tel: (312)355–1789. FAX: (312)996–5921. Electronic mail: jliang@uic.edu.

native contacts formed by the residue in the transition state conformations to the number of contacts formed in the native structure of ground state.

In this paper, we focus on protein structures and develop a general method to obtain ensemble of conformations that satisfy distance constraints given either in the form of an incomplete set of distance bounds, a set of binary conditions whether the distance is below or above a specific threshold, or in the indirect form of experimentally measured ϕ-values. Our approach is based on the framework of sequential Monte Carlo (SMC) method, a growth-based method in which residues are added to an existing partial chain one by one until a conformation of full length is obtained.[17–19] In addition to generating structures, we can also estimate important physical properties of molecular ensembles. As the probabilities of growing viable conformation samples become exceedingly small because of strong distance constraints and the self-avoiding requirement, an efficient sampling strategy becomes critical in order to obtain full chain conformations consistent with all distance constraints.

This paper is organized as follows. In Sec. II, we introduce a cubic lattice model and the incomplete and indirect distance constraints we work with. We then discuss the general approach of SMC method and a new growth potential function. Results are presented in Sec. III.

II. MODEL AND METHOD

A. Protein model and distance constraints

1. Three-dimensional model for proteins

We use a three-dimensional cubic lattice model to represent a protein conformation. The sides of a cubic cell have unit length $\gamma=1.3$ Å. A length-n protein conformation is represented by a connected chain $\boldsymbol{x}_n=(x_1,x_2,\ldots,x_n)$, where the ith C_α atom of the conformation is located at site $x_i=(x_{i1},x_{i2},x_{i3})$ on the cubic lattice.[18,20,21]

For proteins molecules, the locations of C_α atoms satisfy certain constraints. We assume the C_α atoms in our model can only be placed on the lattice sites with the following constraints. First, the Euclidean distance between neighboring residues x_{i-1} and x_i is between 3.5 and 4.1 Å. Second, the direction of the vector x_i-x_{i-1} must be within 30° of four canonical directional vectors, which are specifically determined by the residue type of residue i and the locations x_{i-1}, x_{i-2}, and x_{i-3}. These canonical vectors are derived from a discrete four-state off-lattice model of proteins, which gives four possible locations of x_i for monomer i given the locations of x_{i-1}, x_{i-2}, and x_{i-3}, and the type of residue $i-1$. Details of obtaining optimal canonical directional vectors are given in Ref. 22. Third, we enforce the self-avoiding constraint. Specifically, non-neighboring residues are not permitted to be closer than 3.5 Å, which is the smallest distance of non-neighboring residues observed in 646 representative proteins from the Protein Data Bank (PDB).

On average, there are about 23 candidate positions for placing x_i in our model, although the exact number depends on the different relative positions of x_{i-3}, x_{i-2}, and x_{i-1}, as well as the type of the $(i-1)$th residue. Figure 1 provides an illustration of this lattice model.

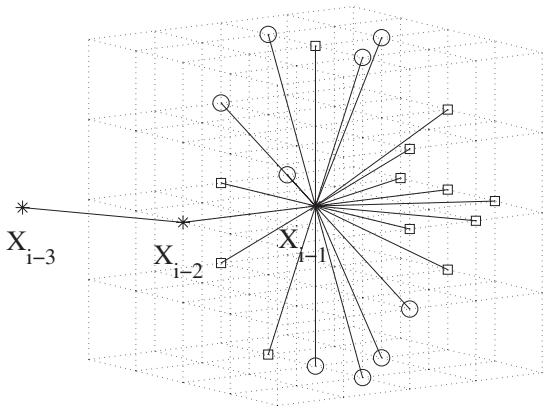

FIG. 1. Illustration of the cubic lattice model. We have set the cell unit length to 1.8 Å instead of 1.3 Å here for clarity. Given the locations of x_{i-3}, x_{i-2}, and x_{i-1}, there are 21 positions (marked by "□" and "◯") satisfying the first distance condition. Only nine positions (marked by "◯") among them also satisfy the second direction condition. These positions all satisfy the third self-avoidance condition.

2. Direct distance constraints

Distance constraints for protein chain $\boldsymbol{x}_n=(x_1,\ldots,x_n)\in\mathbb{R}^3$ are written in the following general form as

$$l_{ij}\le\|x_i-x_j\|\le u_{ij}\quad\text{for all }(i,j)\in\mathcal{D},$$

where $\|x_i-x_j\|$ is the Euclidean distance between residues i and j; l_{ij} and u_{ij} are the lower and upper bounds of the distance between residues i and j. Here l_{ij} can be 0 and u_{ij} can be $+\infty$ if only upper bound or lower bound is available, respectively. $\mathcal{D}$ is a given set of (i,j) residue pairs, in which such constraints are known. $\mathcal{D}$ is often a much smaller subset of the complete set of all residue pairs. The problem of determining the conformation $\boldsymbol{x}_n=(x_1,\ldots,x_n)$ according to such distance constraints has been studied before.[8,23] In this paper, we focus on constraints of distance between C_α atoms, and we only consider the structure of C_α chain. The general principle can be applied to other types of distance constraints, and side chain repacking methods can be used to generate more detailed protein structures.[24,25]

3. Indirect distance constraints and experimentally measured ϕ-values

An important class of studies on protein folding is to characterize the properties of the TSE. TSE represents the structures around the saddle point of the potential energy surface, and these structures are often followed by a large structural change in protein unfolding process.[26–29]

It is challenging to characterize TSE due to the complexity of the folding and unfolding processes. Experimental

research on this problem focuses on the measurement of ϕ-value at individual residue position, defined as the ratio of change in stability to the transition state upon mutation versus the change to the native folded state.[27–31]

ϕ-value provides information about the native likeness of TSE.[32,33] For example, if ϕ_i, the ϕ-value at residue i, is close to 1, the transition state is thought to have almost the same structure at residue i as the native state. If ϕ_i is close to 0, the transition state is likely to be in the denatured state in this region. An important question therefore is to translate ϕ-value measurements into explicit conformational information of protein structures in the TSE.[16,31,34]

Let $\phi_i^{\exp}$ be the experimentally measured ϕ-value at residue i. Based on experimental observations, it is reasonable to assume that changes in protein stability are proportional to the change in the number of contacts in a protein structure.[35] Based on this assumption, the *calculated* ϕ-value ϕ_i^{calc}, which relates to the protein structure, is defined as $\phi_i^{\text{calc}} = C_i^{\text{TSE}}/N_i^N$, where C_i^{TSE} is the average number of contacts formed by residue i in the TSE and N_i^N is the number of contacts formed by residue i in the native structure. In studies based on molecular dynamics simulations, Li and Daggett[31] showed that ϕ_i^{calc} is in good agreement with $\phi_i^{\exp}$. Vendruscolo *et al.*[16] further introduced a different definition of ϕ_i^{calc},

$$\phi_i^{\text{calc}} = \frac{N_i^{\text{TSE}}}{N_i^N}, \tag{1}$$

where N_i^{TSE} is taken as the average number of *native* contacts instead of all contacts of residue i in the TSE. In this case, the TSE is defined as the set of conformations with ϕ_i^{calc} very close to the corresponding experimental measured $\phi_i^{\exp}$ at all positions. An important question is therefore how to obtain explicit conformations of TSE that satisfy these indirect distance constraints of ϕ-values. A number of studies have shown promising results.[16,34]

B. Generating conformations with various distance constraints using SMC

In general, one can aim to obtain conformations that are at the global minimum of an error function measuring deviation in distance from the lower bounds and upper bounds of the distance constraints,

$$\mathcal{E}(\boldsymbol{x}_n) = \sum_{i,j} [\max^2\{l_{ij} - \|x_i - x_j\|, 0\} + \max^2\{\|x_i - x_j\| - u_{ij}, 0\}] \quad \text{for all} \quad (i,j) \in \mathcal{D}, \tag{2}$$

in which the distance constraints are incomplete and inaccurate.[9,12] Our goal is to generate a set of conformations satisfying all distance constraints and following certain target distribution $\pi(\boldsymbol{x}_n)$, for example, the uniform distribution of all feasible conformations satisfying the distance constraints, or the Boltzmann distribution associated with an energy function. If the true energy function was known, it could be used to estimate the thermodynamics properties of the ensemble of protein conformations following the Boltzmann distribution. In reality, one can approximate the unknown true energy function with various empirically derived energy functions, such as the Miyazaw–Jernighan potential function,[36] the geometric poetntial,[37,38] and many other potential functions as reviewed in Ref. 39.

Since our goal here is to minimize the error function $\mathcal{E}(\boldsymbol{x}_n)$, instead of approximating the true energy function, we can set the energy function to be proportional to the error function [Eq. (2)]. More details of this target distributions we use are described in Sec. III. Let $\boldsymbol{x}_t = (x_1, \ldots, x_t)$ be the vector for the positions of residues from 1 up to t. We recursively place residue t at position x_t following a trial distribution $g_t(x_t|\boldsymbol{x}_{t-1})$. The trial distribution proposes possible positions with different probabilities for residue t to be placed under the condition that positions $x_1, \ldots, x_{t-1}$ for residues 1 to $t-1$ are given. The joint trial distribution for a chain with t residues at positions $x_1, \ldots, x_t$ is given by

$$g_t(\boldsymbol{x}_t) = g_1(\boldsymbol{x}_1) g_2(x_2|\boldsymbol{x}_1) \cdots g_t(x_t|\boldsymbol{x}_{t-1}).$$

Following the principle of importance sampling,[40–42] the design of the trial distribution can accommodate different types of bias, which allows great flexibility for improving sampling efficiency. However, each final sample of full length chain $\boldsymbol{x}_n$ needs to be weighted to remove the bias so the original target distribution $\pi(\boldsymbol{x}_n)$ can be recovered. Specifically, we assign a weight

$$w(\boldsymbol{x}_n^{(j)}) = \pi(\boldsymbol{x}_n^{(j)})/g_n(\boldsymbol{x}_n^{(j)})$$

to each conformation sample $\boldsymbol{x}_n^{(j)}$, $j = 1, \ldots, m$, where $g_n(\boldsymbol{x}_n)$ is the trial distribution of the full chain. Then the expected mean value of physical property represented by a function $h(\boldsymbol{x}_n)$ of conformation $\boldsymbol{x}_n$ following the target distribution $\pi(\boldsymbol{x}_n)$ can be estimated by

$$\mathbb{E}_\pi(h(\boldsymbol{x}_n)) \simeq \frac{\sum_{j=1}^m w(\boldsymbol{x}_n^{(j)}) h(\boldsymbol{x}_n^{(j)})}{\sum_{j=1}^m w(\boldsymbol{x}_n^{(j)})}.$$

We adopt the framework developed in Ref. 43 to generate sample conformations which minimizes the loss introduced in the resampling step when choosing a number of distinct samples from a larger sample set. It helps to maintain the diversity of the samples. Let m_t be the number of samples

we retain in the tth iteration, $m_{\max}$ be the maximum value of m_t, the algorithm for generating samples are described in Algorithm 1.

Algorithm 1 Generating conformation

Set $m_1=1$, $w_1^{(1)}=1.0$ and place the first residue at fixed $x_1^{(1)}$.
for $t=2$ to n do
$L_t=0$;
{L_t: number of length t chains that can be obtained from samples obtained at step $t-1$.}
for sample $j=1:m_{t-1}$ do
Find all of the valid sites $x_t^{(i,j)}$ $i=1,\cdots,l_t^{(j)}$ for placing x_t next to partial chain $x_{t-1}^{(j)}$.
{$l_t^{(j)}$=number of available sites to place x_t next to partial chain $x_{t-1}^{(j)}$.}
Generate $l_t^{(j)}$ number of t-length chain $\tilde{x}_t^{(L_t+i)}=(x_{t-1}^{(j)},x_t^{(i,j)})$
$\tilde{w}_t^{(L_t+i)}=w_{t-1}^{(j)}$. {Temporary weights for uniform distribution.}
$L_t=L_t+l_t^{(j)}$.
end for
if $L_t\leq m_{\max}$ then
Let $m_t=L_t$ and $\{(x_t^{(j)},w_t^{(j)})\}_{j=1}^{m_t}=\{(\tilde{x}_t^{(l)},\tilde{w}_t^{(l)})\}_{l=1}^{L_t}$.
else
Let $m_t=m_{\max}$.
for $l=1$ to L_t do
Assign a priority score $\beta_t^{(l)}$ for chain $\tilde{x}_t^{(l)}$ according to the constraints.
end for
Find constant c such that $\Sigma_{l=1}^{L_t}\min\{c\beta_t^{(l)},1\}=m_{\max}$. {e.g. by binary search.}
Draw r from uniform distribution $\mathcal{U}[0,1)$.
for $j=1:m_{\max}$ do
Let $r_j=j-r$.
Find integer J_j such that $\Sigma_{l=1}^{J_j-1}\min\{c\beta_t^{(l)},1\}<r_j\leq\Sigma_{l=1}^{J_j}\min\{c\beta_t^{(l)},1\}$.
Select sample $x_t^{(j)}=\tilde{x}_t^{(J_j)}$.
Set weight $w_t^{(j)}=\tilde{w}_t^{(J_j)}/\min\{c\beta_t^{(J_j)},1\}$.
end for
end if
end for
for $j=1:m_n$ do
Calculate importance weight $w(x_n^{(j)})\propto w_n^{(j)}\pi(x_n^{(j)})$.
end for

The key step in this algorithm is to construct high quality priority scores $\beta_s^{(l)}$, which works as the trial distribution $g_t(x_t|\boldsymbol{x}_{t-1})$ to guide the growth of the partial chains toward more profitable regions.

In this algorithm, it is not necessary to require the growth starts from the first residue x_1. In fact, growth can start from any place, as long as the newly placed monomer is connected to the existing partial chain. For example, growth can start in one direction from a position in the middle of the primary sequence of the chain. After it reaches the end of the chain, the growth process can go back to the starting residue and continue to grow in the other direction of the primary sequence. That is, the order of placing residues can be $(x_k,x_{k-1},\ldots,x_1,x_{k+1},\ldots,x_n)$ or $(x_k,x_{k+1},\ldots,x_n,x_{k-1},\ldots,x_1)$ for any residue k located in the middle of the chain. The steps of adding a new residue to existed partial chain are the same as above. In this study, we choose the order of placing residues so that the fragment of the first 20 residues to be placed has the largest number of distance constraints.

C. Priority score

The choice of a good priority score β_t used in Algorithm 1 is very important. A carefully designed β_t can successfully guide the growth of the conformation so that the full chain will eventually obey all the distance constraints, hence increasing the sampling acceptance rate. A difficulty in the growth-based method is that when adding current residue, the distance information of future residues cannot be directly used. To solve this problem, the priority score we develop consists of three components: growth potential from upper bounds of the distance constraints, growth potential from reference points, and growth potential from lower bounds of the distance constraints. The first two components of the priority score incorporate the distance information of future residues.

1. Growth potential from upper bounds of the distance constraints

Given the upper bounds for the distances between residue pairs in a subset $\mathcal{D}$ of all residue pairs, we first develop distance upper bounds λ_{ij} between all residue pairs (i,j), $i,j=1,\ldots,n$.

Let $q(k)$ be an upper bound of distances between two residues that are k residues away in the protein primary sequence. For constructing the upper bounds $q(k)$ for small sequence separation k, we enumerate self-avoiding chains on the discrete lattice model using the protein sequence of interest. We have

$$q(k)=\max_i \max_{\boldsymbol{x}(i,k)}(\|x_{i+k}-x_i\|),$$

where $\boldsymbol{x}(i,k)$ is a self-avoiding chain of length k starting at residue i. In this study, we enumerate fragments of chains for $k=1,\ldots,5$ at different starting positions i, and take the largest as $q(k)$. When sequence separation k is large, enumeration is infeasible. We approximate $q(k)$ by $k_1q(5)+q(k_2)$ if $k=5k_1+k_2$, where $k_1\in\mathbb{Z}$, $k_2=0,\ldots,4$. This is an upper bound as it assumes the chain is attached at some residues without angle constraint.

Consider a complete graph $\mathcal{G}$ with n vertices, each vertex represents a residue. The length of edge between any two vertices i and j is set to

$$e_{ij}=\begin{cases}\min\{u_{ij},q(|i-j|)\}, & \text{if } (i,j)\in\mathcal{D}\\ q(|i-j|), & \text{otherwise.}\end{cases}$$

We can use the Floyd algorithm[44] to identify the shortest path p_{ij} between any two vertices i and j in this complete

graph $\mathcal{G}$. The distance upper bound λ_{ij} between residues i and j is then set to the total length of the shortest path p_{ij}.

After obtaining the distance upper bound λ_{ij} and the corresponding path p_{ij}, we construct the potential function thatcontributes to the priority score as

$$f_1(\boldsymbol{x}_t) = \sum_{i<j,(i,j)\in\mathcal{P}} h_1(\|x_i - x_j\|, \lambda_{ij}), \tag{3}$$

where $\mathcal{P}$ is a set of (i,j) pairs such that on the shortest pathp$_{ij}$ between i and j, the two ends x_i and x_j are in the partial chain $\boldsymbol{x}_t$, but none of the residues between i and j are in $\boldsymbol{x}_t$. This is to avoid double counting of the distance constraints. The function h_1 is a loss function to measure the violation of constraint $\|x_i - x_j\| \le \lambda_{ij}$. Usually, $h_1(\|x_i - x_j\|, \lambda_{ij})$ is set to zero when $\|x_i - x_j\| \le \lambda_{ij}$, and monotonically nondecreasing as $\|x_i - x_j\| - \lambda_{ij}$ increases. Different types of $h_1(\cdot)$ can be chosen for different considerations, which we will discuss in detail in later sections.

2. Growth potential from reference points

Given a partial chain $\boldsymbol{x}_t$, if the position of a future residue j ($x_j \notin \boldsymbol{x}_t$) is strongly constrained, e.g., there are more than four residues in the existing chain $\boldsymbol{x}_t$ having distance constraints related to residue j, then residue j can only be placed in a small spatial region. We generate candidate position for x_j on lattice sites within this small space. More specifically, if a future residue j has distance constraints,

$$l_{i_k j} \le \|x_{i_k} - x_j\| \le u_{i_k j}, \quad k = 1, \ldots, K,$$

where x_{i_k}, $k=1,\ldots,K$, are in the existing chain $\boldsymbol{x}_t$, and $K \ge 5$, we use Newton's climbing method[45] to find a position z in $\mathbb{R}^3$ such that

$$z = \arg\min_x F(x) = \arg\min_x \sum_{k=1}^{K} (\|x_{i_k} - x\| - u_{i_k j})^2,$$

in which z is obtained by iteratively performing $z = z - (F''(z))^{-1} F'(z)$. We then search the sites on the cubic lattice around position z and choose the site x that minimizes

$$\sum_{k=1}^{K} [\max^2\{l_{i_k j} - \|x_{i_k} - x\|, 0\} + \max^2\{\|x_{i_k} - x\| - u_{i_k j}, 0\}]$$

as the candidate position for residue j. Denote the candidate position as x_j^*, we use it as a reference point to guide the growth of the chain. The following potential function is used to encode this:

$$f_2(\boldsymbol{x}_t) = \sum_{(i,j)\in\mathcal{P}'} h_2(\|x_i - x_j^*\|, \lambda_{ij}), \tag{4}$$

where $\mathcal{P}'$ is a set of (i,j) pairs such that on the shortest path p_{ij} between i and j, x_i is in the partial chain $\boldsymbol{x}_t$ constructed so far, x_j^* is the reference point, and none of the residues between i and j are in $\boldsymbol{x}_t$. As before, h_2 is the loss function to measure the violation of constraint $\|x_i - x_j^*\| \le \lambda_{ij}$.

3. Growth potential from lower bounds of the distance constraints

This potential function penalizes the violation of lower bound constraint,

$$f_3(\boldsymbol{x}_t) = \sum_{(i,j)\in\mathcal{S}\cap\mathcal{D}} h_3(\|x_i - x_j\|, l_{ij}), \tag{5}$$

where $\mathcal{S}$ is the set of residue pair (i,j) in which x_i and x_j exist in the partial chain $\boldsymbol{x}_t$. Here h_3 is the loss function to measure the violation of constraint $\|x_i - x_j\| \ge l_{ij}$. Hence, $h_3(\|x_i - x_j\|, l_{ij}) = 0$, when $\|x_i - x_j\| \ge l_{ij}$, and is monotonically nondecreasing as $\|x_i - x_j\| - l_{ij}$ decreases.

4. Combined priority score

The combined priority score $\beta_t^{(l)}$ for chain $\tilde{\boldsymbol{x}}_t^{(l)}$ is set as

$$\beta_t^{(l)} = \exp\left[- \frac{\rho_1 f_1(\tilde{\boldsymbol{x}}_t^{(l)}) + \rho_2 f_2(\tilde{\boldsymbol{x}}_t^{(l)}) + \rho_3 f_3(\tilde{\boldsymbol{x}}_t^{(l)})}{\tau_t} \right], \tag{6}$$

where ρ_1, ρ_2, and ρ_3 are coefficients of the three growth potential functions, τ_t is a temperaturelike variable. The choice of loss functions h_1, h_2, and h_3 in f_1, f_2, and f_3, and coefficients $\rho_1, \rho_2, \rho_3, \tau_t$ will be described in later sections.

D. Generating conformations from incomplete residue distance constraints

In this section, we discuss how to use Algorithm 1 to generate protein conformations with given constraints in the form of small intervals of distances between a subset of residue pairs. The distance constraints are represented as[12,13]

$$d_{ij} - \epsilon_{ij} \le \|x_i - x_j\| \le d_{ij} + \epsilon_{ij} \quad \text{for all } (i,j) \in \mathcal{D},$$

where d_{ij} is the distance of residues i and j in the native structure. The set $\mathcal{D}$ is assigned as follows: each nonneighboring residue pair within short range distance (SRD) in the native structure is selected in $\mathcal{D}$ with a certain probability, e.g., (20%, 40%,…, 100%) independently. The SRD is selected as 10 Å for residue level structure following

Ref. 13. All residue pairs with distance $d_{ij}>10$ Å are excluded from $\mathcal{D}$. Variations ϵ_{ij}, ϵ_{ij} of the bounds are randomly selected from uniform distribution $\mathcal{U}[0,1)$ independently, so that the distance variation is under 1 Å, about 10% of the true distance d_{ij} as in Ref. 13.

In this problem, the priority score in Algorithm 1 is set for Eq. (6) with

$$h_1(z,\lambda)=h_2(z,\lambda)=\begin{cases}(z-\lambda)^2, & \text{if } z>\lambda\\ 0, & \text{if } z\le\lambda,\end{cases}$$

and

$$h_3(z,l)=\begin{cases}(z-l)^2, & \text{if } z<l\\ 0, & \text{if } z\ge l,\end{cases}$$

for f_1, f_2, and f_3, and parameters are set as $\rho_1=1$, $\rho_2=1$, $\rho_3=1$, and $\tau_t=0.5$. Here z is the value of the corresponding distance.

The loss functions are chosen so that the distance between any two residues i and j in the conformational sample does not deviate too much from the given constrained interval $[l_{ij},u_{ij}]$, in case that not all constraints can be perfectly satisfied simultaneously. The loss functions h_1, h_2, and h_3 are concave downward functions of the distance $\|x_i-x_j\|$, which increases rapidly as $\|x_i-x_j\|$ departs the constrained interval $[l_{ij},u_{ij}]$.

E. Generating conformations from contact map of distance cutoff

We now describe how to generate conformations based on a given incomplete contact map, where distances between some residue pairs are known to be either above or below a cutoff value in our calculation. We use 8.5 Å as the cutoff value. This value has been used by Vendruscolo *et al.* in Ref. 16.

The contact map of a length n polymer chain is a $n\times n$ symmetric matrix $\mathcal{C}=\{c_{ij}\}_{n\times n}$, where $c_{ij}=1$ if residues i and j are in contact, and $c_{ij}=0$ otherwise. A given contact map is equivalent to a set of distance constraints,

$$\|x_i-x_j\|\le 8.5 \text{ Å} \quad \text{for all } c_{ij}=1,$$

$$\|x_i-x_j\|>8.5 \text{ Å} \quad \text{for all } c_{ij}=0.$$

For this problem, we use

$$h_1(z,\lambda)=h_2(z,\lambda)=\mathbb{I}(z-\lambda>0)$$

and

$$h_3(z,l)=\mathbb{I}(z-l<0),$$

in Eqs. (3)–(5), respectively, to construct the priority score in Eq. (6). Here $\mathbb{I}(\cdot)$ is the indicator function: $\mathbb{I}(\cdot)=1$ if the statement represented by $(\cdot)$ is true, 0 otherwise. Parameters in Eq. (6) are taken as $\rho_1=1$, $\rho_2=1$, $\rho_3=0.8$, and $\tau_t=0.2$.

The loss functions are chosen in order to keep the contact map of the generate conformational samples as close to the given target contact map as possible if not completely satisfied. In particular, if the distance $\|x_i-x_j\|$ violates the distance constraint, the corresponding loss function increases from 0 to 1 instantly.

F. Generating contact maps and conformations from indirect distance constraints by ϕ-values

In this section, we describe how to obtain contact maps based on indirect distance constraints in the form of experimentally measured ϕ-values.

1. Generating contact maps from indirect distance constraints

ϕ-values of TSE and contact maps. For generating conformations of the TSE, our target distribution $\pi(\boldsymbol{x}_n)$ is the uniform distribution of all conformations $\boldsymbol{x}_n$ satisfying the ϕ-value constraints,

$$\phi_i^{\text{calc}}(\boldsymbol{x}_n)=\frac{N_i^{\text{TSE}}}{N_i^N}\approx\phi_i^{\text{exp}},\quad i\in\mathcal{I}=\{I_1,\ldots,I_T\},$$

where $\mathcal{I}$ represents the set of residues whose ϕ-values have been measured experimentally, and $I_1,\ldots,I_T$ are the indexes of these residues. By definition, $\phi_i^{\text{calc}}(\boldsymbol{x}_n)$ can be computed when the conformation $\boldsymbol{x}_n$ of the full chain is known. When only information of partial chains $\boldsymbol{x}_{t-1}$ is available during chain growth, it is difficult to construct an effective conditional trial distribution $g_t(x_t|\boldsymbol{x}_{t-1})$.

Our approach is to translate the ϕ-value constraints into contact maps of equivalent distance constraints. These contact maps provide more direct information on distance constraints for generating conformations. We then sample conformations following the contact map constraints, which will

automatically satisfy all ϕ-value constraints. We describe briefly how to generate conformations from these ϕ-value derived contact maps in Sec. II F 2.

From ϕ-values to contact maps. Because of the intrinsic symmetry of the contact map $\mathcal{C}=\{c_{ij}\}_{n\times n}$, we consider c_{ij} and c_{ji} as the same entry in $\mathcal{C}$. Let $\mathcal{N}$ be the set of residue pairs (i,j) forming native contacts. By definition, the calculated ϕ-value ϕ_i^{calc} for residue i of a conformation only depends on the values of c_{ij} in its contact map that are native contacts formed by residue i. Let

$$\mathcal{C}_i=\{c_{ij}|(i,j)\in\mathcal{N}\}.$$

The size of this set, $|\mathcal{C}_i|$, is the number of contacts formed by residue i in native structure. Note that if $(i,j)\in\mathcal{N}$, both $\mathcal{C}_i$ and $\mathcal{C}_j$ contain c_{ij}.

To generate contact map $\mathcal{C}$ satisfying the ϕ-value constraints, we only need to decide which subset of native contacts to preserve for residue i whose experimental ϕ-value is available. To satisfy the ϕ-value constraint, there needs to be $|\mathcal{C}_i|\cdot\phi_i^{\text{exp}}$ number of native contacts preserved for residue i in the contact map. That is, we need to assign either 0 or 1 to elements in $\mathcal{C}_i, i\in\mathcal{I}$, such that there are exactly $|\mathcal{C}_i|\cdot\phi_i^{\text{exp}}$ number of "1" s in $\mathcal{C}_i$. That is, for each generated contact map we should have $\Sigma_{c_{ij}\in\mathcal{C}_i}c_{ij}=|\mathcal{C}_i|\cdot\phi_i^{\text{exp}}$, $i\in\mathcal{I}$. For simplicity, we denote $\Psi_i=|\mathcal{C}_i|\cdot\phi_i^{\text{exp}}$ in the subsequent discussion.

Now we generate contact map samples properly weighted with respect to the uniform distribution of all contact maps that physically satisfy the ϕ-value constraints. Each contact map sample $\mathcal{C}$ generated by importance sampling via the use of a trial distribution $g(\mathcal{C})$ is weighted by $v=1/g(\mathcal{C})$. Here $g(\mathcal{C})$ is the probability to generate contact map $\mathcal{C}$.

A similar problem has been studied in Ref. 46 for generating 0–1 tables with fixed marginal sums. Although in our problem, the contact map has to be symmetric and only part of it needs to be filled, some techniques in Ref. 46 can be used to improve the sampling efficiency.

Specifically, we proceed by assigning the proper numbers of "1's" and "0's" in the rows for residues with experimental ϕ-value measurement. That is, we fill 0's and 1's in $\mathcal{C}_i$, $i\in\mathcal{I}$, and repeat this position after position, until the rows corresponding to all residues with experimental ϕ-value measurement are assigned. Let m^* denote the number of contact map samples we will generate, $\mathcal{C}^k_{I_1:I_t}$ denote the partially filled kth contact map we have obtained thus far after finishing positions I_1 to I_t, and $v_t^{(k)}$ be the weight of the kth contact map that has been partially filled up to position I_t. The algorithm for generating contact maps from ϕ-value measurement is listed as Algorithm 2.

Algorithm 2 Generating contact map

for $k=1$ to m^* do
  $\mathcal{C}^{(k)}_{I_1:I_0}=\emptyset$, $v_0^{(k)}=1$
end for
for position index $t=1$ to T do
  for sample $k=1$ to m^* do
    for $s=t$ to T do
      Divide $\mathcal{C}_{I_s}$ into disjoint sets $\mathcal{S}^{(k)}_{0,I_s}$, $\mathcal{S}^{(k)}_{1,I_s}$, and $\mathcal{S}^{(k)}_{u,I_s}$ based on partial contact map $\mathcal{C}^{(k)}_{I_1:I_{t-1}}$, where $\mathcal{S}^{(k)}_{1,I_s}=\{\mathcal{C}_{I_s,j}|\text{already filled with 1}\}$, $\mathcal{S}^{(k)}_{0,I_s}=\{\mathcal{C}_{I_s,j}|\text{already filled with 0}\}$, and $\mathcal{S}^{(k)}_{u,I_s}=\{\mathcal{C}_{I_s,j}|\text{unspecified}\}$.
    end for
    repeat
      for $s=t$ to T do
        if $|\mathcal{S}^{(k)}_{1,I_s}|>\Psi_{I_s}$ then
          Remove this sample. {Already too many "1"s.}
        else if $|\mathcal{S}^{(k)}_{1,I_s}|=\Psi_{I_s}$ then
          Fill all elements in $\mathcal{S}^{(k)}_{u,I_s}$ with 0.
          Update $\mathcal{S}^{(k)}_{0,I_j}$, $\mathcal{S}^{(k)}_{1,I_j}$, $\mathcal{S}^{(k)}_{u,I_j}$, $j\in\{t,\cdots,T\}$.
        end if
        if $|\mathcal{S}^{(k)}_{1,I_s}|+|\mathcal{S}^{(k)}_{u,I_s}|<\Psi_{I_s}$ then
          Remove this sample. {Already too many "0"s.}
        else if $|\mathcal{S}^{(k)}_{1,I_s}|+|\mathcal{S}^{(k)}_{u,I_s}|=\Psi_{I_s}$ then
          Fill all elements in $\mathcal{S}^{(k)}_{u,I_s}$ with 1.
          Update $\mathcal{S}^{(k)}_{0,I_j}$, $\mathcal{S}^{(k)}_{1,I_j}$, $\mathcal{S}^{(k)}_{u,I_j}$, $j\in\{t,\cdots,T\}$.
        end if
      end for
    until $\mathcal{S}^{(k)}_{u,I_t}=\emptyset$, or none of $\mathcal{S}^{(k)}_{0,I_s}$, $\mathcal{S}^{(k)}_{1,I_s}$, $\mathcal{S}^{(k)}_{u,I_s}$, $s\in\{t,\cdots,T\}$ changes.
    {This step must converge because the number of unspecified positions decreases monotonically as the iteration proceeds.}
    if $\mathcal{S}^{(k)}_{u,I_t}=\emptyset$ then
      $\mathcal{C}^{(k)}_{I_t}$ is completed and let weight $v_t^{(k)}=v_{t-1}^{(k)}$.
    else
        Fill $\mathcal{S}^{(k)}_{u,I_t}$ with $\Psi_{I_t}-|\mathcal{S}^{(k)}_{1,I_t}|$ "1"s following CP-distribution.
        {When there are unspecified entries in this row.}
        Update weight $v_t^{(k)}$ by Eq. (8).
    end if
  end for
  Optionally resample[38] $\{(\mathcal{C}^{(k)}_{I_1:I_t}, v_t^{(k)})\}_{k=1}^{m^*}$ if many samples were removed.
end for

Constrained Poisson (CP) distribution. The details of CP distribution can be found in Ref. 47. Briefly, we sample 0's and 1's to fill each entry $s_1,\ldots,s_{|\mathcal{S}^{(k)}_{u,I_t}|}$ of $\mathcal{S}^{(k)}_{u,I_t}$ described in Algorithm 2 with probability proportional to

$$g(s_1,\ldots,s_{|\mathcal{S}^{(k)}_{u,I_t}|})\propto\prod_{j=1}^{|\mathcal{S}^{(k)}_{u,I_t}|}p_j^{s_j}(1-p_j)^{1-s_j},$$

and the total number of assigned 1's is $\Sigma_{j=1}^{|\mathcal{S}^{(k)}_{u,I_t}|}s_j=\Psi_{I_t}-|\mathcal{S}^{(k)}_{1,I_t}|$. Here $p_j\in[0,1\text{stretchy='true'}]$ are the chosen parameters to improve the sample survival probability of this distribution.

Parameters for conditional Poisson (CP) distribution. For each entry $s_j\in\mathcal{S}^{(k)}_{u,I_t}$ to be filled (whose corresponding entry in the contact map is c_{I_t,J_j}), we assign the parameter p_j for the CP distribution as

$$p_j = \begin{cases} \dfrac{\Psi_{J_j} - |\mathcal{S}^{(k)}_{1,J_j}|}{|\mathcal{S}^{(k)}_{u,J_j}|}, & \text{if } c_{I_t,J_j} \in \mathcal{S}^{(k)}_{u,I_t} \cap A \\ \max\left\{ \dfrac{\Psi_{I_t} - |\mathcal{S}^{(k)}_{1,I_t}| - \sum_{c_{I_t,i} \in \mathcal{S}^{(k)}_{u,I_t} \cap A} p_i}{|\mathcal{S}^{(k)}_{u,I_t} \setminus A|}, 0.1 \right\}, & \text{if } c_{I_t,J_j} \in \mathcal{S}^{(k)}_{u,I_t} \setminus A, \end{cases}$$

where $A=\{c_{I_t,I_{t+1}}, c_{I_t,I_{t+2}}, \ldots, c_{I_t,I_T}\}$ are entries recording existence of contacts between residue I_t and other future residues with experimental ϕ-values.

If J_j is a position with ϕ-measurement but currently unspecified, we assign p_j as the ratio of the number of 1's to be assigned $\Psi_{J_j} - |\mathcal{S}^{(k)}_{1,J_j}|$ and the number of unspecified positions $|\mathcal{S}^{(k)}_{u,J_j}|$ for residue J_j.

If J_j is a position currently unspecified but not a position with known ϕ-measurement, we assign p_j as the ratio of the number of 1 to be assigned $\Psi_{I_t} - |\mathcal{S}^{(k)}_{1,I_t}|$, minus an expected number $\Sigma_{c_{I_t,i} \in \mathcal{S}^{(k)}_{u,I_t} \cap A} p_i$ of 1's that will be assigned for future positions with ϕ values, and the number $|\mathcal{S}^{(k)}_{u,I_t} \setminus A|$ of unspecified positions without known ϕ-values, or the value of 0.1, which ever is larger. This choice of p_j is expected to fill $\Psi_{J_j} - |\mathcal{S}^{(k)}_{1,J_j}|$ number of 1's in $|\mathcal{S}^{(k)}_{u,j}|$ for $j \in \{I_t, I_{t+1}, \ldots, I_T\}$. Note that in this assignment, p_i is guaranteed to have a value between 0 and 1.

Realization of CP distribution. The overall idea for sampling from the CP distribution is to take out $\Psi_{I_t} - |\mathcal{S}^{(k)}_{1,I_s}|$ number of elements from the set $\mathcal{S}^{(k)}_{u,I_t}$ one by one without following specific probability replacement. These elements will be assigned as 1's, while the remaining ones will be 0's.[46]

Specifically, let $a_j = p_j/(1-p_j)$. Suppose $\overline{\mathcal{S}}^{(k)}_{u,I_t}(i)$ are the remaining elements after taking out i elements ($i=0,1,\ldots,\Psi_{I_t} - |\mathcal{S}^{(k)}_{1,I_s}| - 1$). Each $s_j \in \overline{\mathcal{S}}^{(k)}_{u,I_t}(i)$ will be selected as next element to be taken out and assigned the value of 1 with probability

$$P(s_j, \overline{\mathcal{S}}^{(k)}_{u,I_t}(i)) = \frac{a_j \cdot R(\Psi_{I_t} - |\mathcal{S}^{(k)}_{1,I_s}| - i - 1, \overline{\mathcal{S}}^{(k)}_{u,I_t}(i) \setminus \{s_j\})}{(\Psi_{I_t} - |\mathcal{S}^{(k)}_{1,I_s}| - i) \cdot R(\Psi_{I_t} - |\mathcal{S}^{(k)}_{1,I_s}| - i, \overline{\mathcal{S}}^{(k)}_{u,I_t}(i))},$$

where $R(i,\mathcal{S})$ is

$$R(i,\mathcal{S}) = \sum_{\mathcal{B} \subset \mathcal{S}, |\mathcal{B}|=i} \left(\prod_{j \in \mathcal{B}} a_j \right). \tag{7}$$

It is the summation of $\Pi_{j \in \mathcal{B}} a_j$ of all size i subsets $\mathcal{B}$ in $\mathcal{S}$.

For an integer i and a subset $\mathcal{S} \subset \mathcal{S}^{(k)}_{u,I_t}$, $R(i,\mathcal{S})$ can be calculated using the recursive formula

$$R(i,\mathcal{S}) = R(i, \mathcal{S} \setminus \{s_j\}) + a_i R(i-1, \mathcal{S} \setminus \{s_j\})$$

for any $s_j \in \mathcal{S}$. The initial conditions for the recursion are $R(0,\mathcal{S})=1$ for any $\mathcal{S} \subset \mathcal{S}^{(k)}_{u,I_t}$ and $R(i,\mathcal{S})=0$ for any $|\mathcal{S}| < i$.

Updating sample weight. The weight associated with a sample of contact map is updated as

$$v^{(k)}_t = v^{(k)}_{t-1} \cdot \frac{R(\Psi_{I_t} - |\mathcal{S}^{(k)}_{1,I_s}|, \mathcal{S}^{(k)}_{u,I_t})}{\prod_{j=1}^{|\mathcal{S}^{(k)}_{u,I_t}|} a_j^{s^{(k)}_j}}, \tag{8}$$

where $s^{(k)}_1, \ldots, s^{(k)}_{|\mathcal{S}^{(k)}_{u,I_t}|}$ is a realization of $s_1, \ldots, s_{|\mathcal{S}^{(k)}_{u,I_t}|}$ for the kth contact map and $R(i,\mathcal{S})$ is defined in Eq. (7).

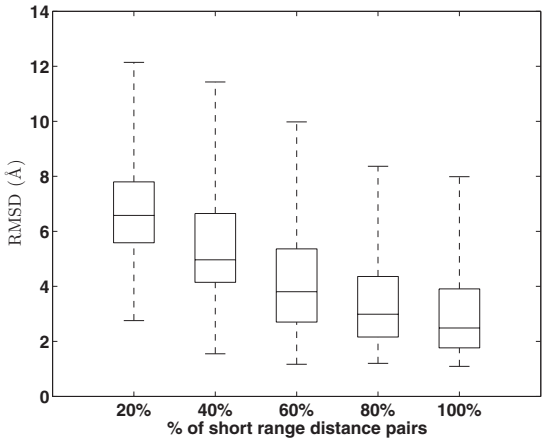

FIG. 2. Box plot of expected to native structures rmsd expectations measured in Å of conformations following Boltzmann distribution of the error function $\pi(\boldsymbol{x}_n) \propto \exp\{-\mathcal{E}(\boldsymbol{x}_n)/\tau|\mathcal{D}|\}$ for 189 proteins with length between 80 and 120. The boxes have lines at the lower quartile, median, and upper quartile values. The lines extending from each end of the boxes are to show the rest of the data. X axis is the percentage of native SRD pairs included in the constraint set $\mathcal{D}$.

2. Generating conformations from contact map samples derived from ϕ-values

With a set of properly weighted samples of contact map $\{(\mathcal{C}^{(k)}, v^{(k)}_T), k=1,\ldots,m^*\}$, we draw a subset of it. The probability for each sample to be drawn is proportional to $v^{(k)}_T$. For each selected contact map, we use it as the target contact map to generate conformations following Algorithm 1, using the priority score described in Sec. II E. The set of the generated conformations form the TSE.

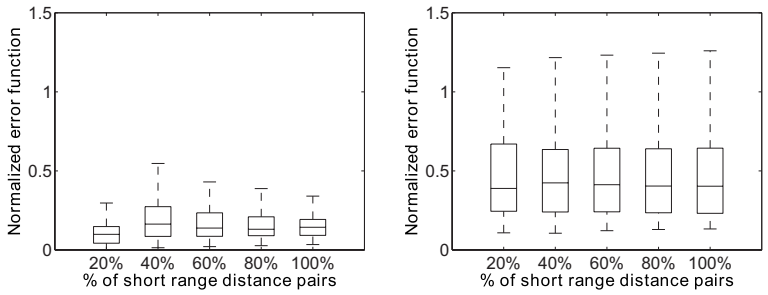

FIG. 3. Normalized error function of the recovered structure and the fittest structure. (a) Box plot of normalized error function $\mathcal{E}(\boldsymbol{x}_n)/|\mathcal{D}|$ of recovered structures of 189 proteins with length of 80–120; (b) box plot of normalized error function $\mathcal{E}(\boldsymbol{x}_n)/|\mathcal{D}|$ of the fittest native structures of 189 proteins with length of 80–120. The boxes have lines at the lower quartile, median, and upper quartile values. The lines extending from each end of the boxes are to show the rest of the data. X axis is the percentage of native SRD pairs included in the constraint set $\mathcal{D}$.

III. RESULTS

A. Result of generating conformations from incomplete residue distance constraints

This section shows the result of generating protein conformations with given constraints in the form of small intervals of distances for a subset of residue pairs as described in Sec. II D.

Consider the Boltzmann distribution $\pi(\boldsymbol{x}_n) \propto \exp\{-\mathcal{E}(\boldsymbol{x}_n)/\tau|\mathcal{D}|\}$, where $\mathcal{E}(\boldsymbol{x}_n)/|\mathcal{D}|$ is the error function defined in Eq. (2) normalized by the number of constraints, τ is a temperaturelike parameter in the Boltzmann function. It reflects deviation from the lower and upper bounds of the distance constraints. Here we set $\tau=0.5$. We use Algorithm 1 to estimate the expected root mean square distance (rmsd) to the native structure of conformations following this Boltzmann distribution. The algorithm is applied to 189 proteins chosen from PDB, whose lengths are between 80 and 120. The distance constraints are constructed for non-neighboring residue pairs whose distances are less than 10 Å (SRD). The percentage of SRD pairs included in the given constraint set $\mathcal{D}$ varies from 20% to 100%.

The growth priority score used in Algorithm 1 is described in Sec. II D. We repeat the algorithm 20 times independently with at most $m_{\max}=1000$ samples being kept during each computation. The corresponding estimated rmsd expectations of distance constraint set $\mathcal{D}$ that includes different percentages of SRDs are plotted in Fig. 2. The boxes in the figure have lines at the lower quartile, median, and upper quartile values of the estimated expectations of the 189 proteins. We can see the corresponding expectation of rmsd becomes smaller as the percentage of the constraints increases. This is expected, as the Boltzmann probabilities $\pi(\boldsymbol{x}_n)$ of conformations close to the native structure tend to be larger as more distance constraints are available.

We can choose the conformation with the smallest error function Eq. (2) from the generated conformation samples as the *recovered structure*. In Fig. 3(a), we plot the values of normalized error function $\mathcal{E}(\boldsymbol{x}_n)/|\mathcal{D}|$ of these recovered structures, compared to the values of normalized error function of the *fittest native structures* [Fig. 3(b)]. The fittest native structure is the conformation in our discrete model, whose rmsd to the native structure is the smallest. It is obtained by a greedy growth method (Ref. 19) with a local minimal rmsd to the native structure. Although the objective of our algorithm is to generate conformations following the Boltzmann distribution $\pi(\boldsymbol{x}_n) \propto \exp\{-\mathcal{E}(\boldsymbol{x}_n)/\tau|\mathcal{D}|\}$, we still can find conformations with smaller error function values in terms of violation of distance donstraints than the fittest native structures.

The rmsd's of the recovered structures to native structures are plotted in Fig. 4. When the distance informations of all SRD are provided, the recovered structures of 160 out of the 189 proteins have rmsd to the native structures less than 3 Å. In general, the recovered structures approach native structures as more distance constraints are incorporated. This shows that the priority score β_t we use introduces larger probability to generate conformations close to the native structure when more distance constraints are available.

We compare the difficulties of recovering structures from distance constraints among different protein classes. The rmsd's of the recovered structures to native structures of ten proteins of different classes are reported in Table I. Compared to alpha helical proteins, the recovered structures from incomplete distance constraints for beta proteins and alpha/beta proteins have larger rmsd's to the native structures. Table II reports the normalized error function of the recovered structures and the fittest native structures (in parentheses). Although the recovered structure and the fittest structure are both fixed, depending on the choice of the constraints at different percentages, values of the error function normalized by the number of constraints will be different. We also report the number of violated distance constraints of the recovered structures and the fitted native structures in Table III. The results show that although the recovered structures violate some of the distance constraints, values of the normalized

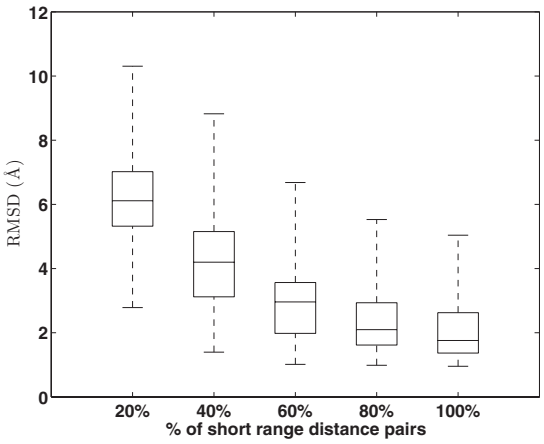

FIG. 4. Box plot of rmsd's measured in Å of recovered structures of 189 proteins with length of 80–120 to native structures. The boxes have lines at the lower quartile, median, and upper quartile values. The lines extending from each end of the boxes are to show the rest of the data. X axis is the percentage of native SRD pairs included in the constraint set $\mathcal{D}$.

TABLE I. rmsd's measured in Å of the recovered structures and the fittest native structures to native structures of ten proteins of different classes. Number of all SRD pairs: the number of all residue pairs with distance less than 10 Å. % of SRD: the percentage of SRDs included in the constraint set $\mathcal{D}$.

RMSD to native structure measured in Å									
					Structure recovered from % of SRD				
PDB ID	Protein class	Protein length	# of all SRD pairs	Fittest native structure	20%	40%	60%	80%	100%
2mhr	All alpha	118	765	0.9	5.3	3.9	1.9	2.3	1.5
256b	All alpha	106	752	1.0	9.6	3.7	1.5	1.2	1.4
1cmc	All alpha	104	619	0.9	6.3	4.9	2.7	2.1	2.3
1btn	All beta	106	749	1.5	6.1	6.5	4.0	4.1	2.3
1f7d	All beta	118	796	1.4	7.9	8.4	8.0	5.1	4.5
1f86	All beta	115	816	1.2	5.5	5.3	3.4	2.6	2.3
2trx	Alpha/beta	108	728	1.1	5.2	3.5	2.1	2.1	1.8
1bkf	Alpha/beta	107	788	1.5	5.6	2.6	2.0	2.1	1.6
1lkk	Alpha/beta	105	719	1.0	6.2	4.3	1.8	2.3	1.6
1puc	Alpha/beta	101	455	0.9	10.6	8.9	7.1	7.5	4.2

error function can be much smaller than the fittest native structures. This is because the loss functions h_1, h_2, and h_3 we use are concave downward functions, which focus on preventing the distance between residues being far away from the given distance constraints.

The relatively large number of constraint violation may be due to certain limitation of our discrete model. There may not exist any conformation on the lattice satisfying all of the distance constraints. To address this issue, we construct a different set of distance constraints using the fittest native structure among the conformations of the discrete model, which is obtained by a greedy method. The new set of distance constraints are

$$\tilde{d}_{ij} - \epsilon_{ij} \leq \|x_i - x_j\| \leq \tilde{d}_{ij} + \epsilon_{ij} \quad \text{for all } (i,j) \in \mathcal{D},$$

where $\tilde{d}_{ij}$ is the distance of residues i and j in the fittest native structure. In this case, there exists at least one conformation, the fittest native structure, in the discrete model satisfying all the distance constraints. Under this setting, the normalized error function $\mathcal{E}(\mathbf{x}_n)/|\mathcal{D}|$ of the recovered structures is plotted in Fig. 5, and the rmsd's of the recovered structures to the fittest native structures are plotted in Fig. 6. Among 189 proteins, the recovered structures of 40 proteins can match the fittest native structures perfectly when all SRD pairs are in the constraint set $\mathcal{D}$.

B. Result of generating conformations from contact map of distance cutoff

This section shows the result of generating conformations based on a given contact map, where the distances between residue pairs are known to be either above or below a cutoff value (8.5 Å).[16]

We choose 20 proteins with length of 50–200 from the Protein Data Bank and generate conformations from their complete native contact map using Algorithm 1. We repeat the computation ten times independently and at most $m_{\max} = 1000$ samples are kept during each computation. The conformation with the smallest numbers of missing contacts (residue pairs that form contact in the native structure but not in the generated conformation) and extraneous contacts (residue pairs that form contact in the generated conformation but not in the native structure) is chosen as the recovered struc-

TABLE II. Value of the normalized error function of the recovered structures and of the fittest native structures (in parentheses) of ten proteins of different classes. % of SRD: the percentage of SRDs included in the constraint set $\mathcal{D}$.

Normalized error function					
	Structure recovered from % of SRD				
PDB ID	20%	40%	60%	80%	100%
2mhr	0.050 (0.108)	0.117 (0.105)	0.085 (0.122)	0.092 (0.136)	0.093 (0.140)
256b	0.060 (0.237)	0.049 (0.199)	0.077 (0.203)	0.085 (0.201)	0.083 (0.202)
1cmc	0.018 (0.211)	0.071 (0.196)	0.071 (0.164)	0.068 (0.165)	0.097 (0.161)
1btn	0.072 (0.842)	0.465 (0.648)	0.478 (0.650)	0.530 (0.678)	0.370 (0.696)
1f7d	0.147 (0.669)	0.293 (0.726)	0.260 (0.648)	0.343 (0.716)	0.200 (0.688)
1f86	0.144 (0.527)	0.217 (0.469)	0.330 (0.430)	0.339 (0.415)	0.198 (0.443)
2trx	0.044 (0.564)	0.102 (0.466)	0.196 (0.471)	0.120 (0.427)	0.129 (0.418)
1bkf	0.159 (0.526)	0.117 (0.564)	0.130 (0.691)	0.131 (0.667)	0.164 (0.584)
1lkk	0.264 (0.214)	0.146 (0.216)	0.128 (0.239)	0.128 (0.246)	0.121 (0.228)
1puc	0.009 (0.181)	0.083 (0.157)	0.103 (0.134)	0.069 (0.137)	0.068 (0.132)

TABLE III. The numbers of violations of distance constraints of the recovered structures and the fittest native structures (in parentheses) of ten proteins of different classes. % of SRD: the percentage of SRDs included in the constraint set $\mathcal{D}$.

	Number of violated distance constraints				
PDB ID	Structure recovered from % of SRD				
	20%	40%	60%	80%	100%
2mhr	67 (61)	158 (137)	224 (211)	288 (303)	383 (377)
256b	63 (74)	123 (143)	202 (231)	271 (313)	318 (372)
1cmc	36 (60)	92 (107)	150 (176)	211 (236)	281 (302)
1btn	62 (93)	188 (175)	267 (263)	386 (352)	427 (439)
1f7d	81 (83)	183 (184)	260 (294)	375 (383)	428 (463)
1f86	72 (96)	191 (210)	279 (302)	349 (384)	455 (504)
2trx	63 (79)	133 (156)	214 (233)	276 (313)	351 (398)
1bkf	84 (98)	160 (182)	237 (292)	311 (383)	414 (489)
1lkk	82 (79)	137 (158)	249 (242)	299 (327)	376 (402)
1puc	30 (47)	87 (88)	133 (134)	150 (180)	171 (216)

ture. The number of missed contacts, extraneous contacts, and rmsd to native structures measured in angstroms of the recovered structures are reported in Table IV. Figure 7 shows rmsd of the recovered structures to native structures. Again, we found that the recovered structures of alpha helical proteins have smaller rmsd to the native structures.

C. Result of generating contact maps and conformations from indirect distance constraints by ϕ-values

This section depicts the result of generating TSE from ϕ-value constraints.

We generate TSE of bovine acyl-coenzyme A-binding protein, a length 86 protein with experimental ϕ-values. The PDB entry of the protein is 1nvl. The experimental ϕ-values are plotted in Fig. 8. More details of the experimental ϕ-values can be found in Ref. 48. We follow Ref. 16 and define TSE as the conformations satisfying $|\phi_i^{\text{calc}} - \phi_i^{\text{exp}}| < 0.15$ for all residue i with experimental measured ϕ-value. Hence, the target distribution is the uniform distribution of all conformations satisfying these constraints.

We generate $m^* = 10\,000$ contact map samples using Algorithm 2, among which 1000 contact maps are chosen with probability proportional to their corresponding weights. For each chosen contact map, Algorithm 1 is used to generate conformations. At most $m_{\max} = 1000$ conformations are generated for each contact map. Figure 8 reports ϕ_i^{calc} of the generated TSE. It is seen that the generated TSE can faithfully reproduce the ϕ-values. The average rmsd between TSE and the native structure of 1nvl is 11.3 Å. The result shows that the conformations of TSE can be far away from the native structure.

IV. DISCUSSION

Obtaining molecular structures from incomplete and inaccurate distance information provided by experiments is an important problem. Several global optimization methods has been applied to solve this problem,[9,12–14] in which the goal is

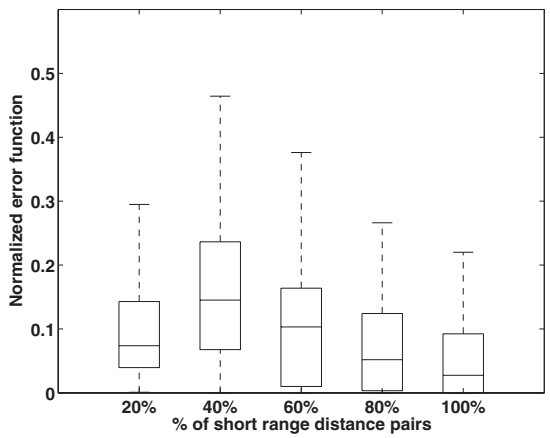

FIG. 5. Box plot of normalized error function $\mathcal{E}(\boldsymbol{x}_n)/|\mathcal{D}|$ of recovered structures of 189 proteins with length of 80–120 when distance constraints are constructed based on the fittest native structures. The boxes have lines at the lower quartile, median, and upper quartile values. The lines extending from each end of the boxes are to show the rest of the data. X axis is the percentage of native SRD pairs included in the constraint set $\mathcal{D}$.

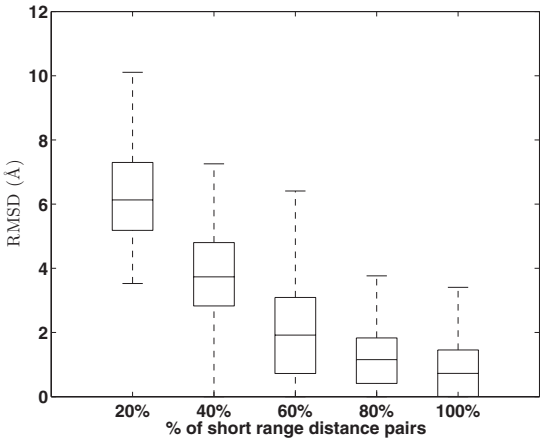

FIG. 6. Box plot of rmsd's measured in Å of the recovered structures to the fittest native structures of 189 proteins with length of 80–120 when distance constraints are constructed based on the fittest native structures. The boxes have lines at the lower quartile, median, and upper quartile values. The lines extending from each end of the boxes are to show the rest of the data. X axis is the percentage of native SRD pairs included in the constraint set $\mathcal{D}$.

TABLE IV. List of proteins of different classes used to recover structures form complete native contact maps. The number of all native contacts, the number of missed contacts, the number of false positive contacts, rmsd to native structure in Å are also listed.

PDB ID	Protein class	Protein length	Number of native contacts	Number of missed contacts	Number of extraneous contacts	rmsd (Å)
1ptq	Small protein	50	164	8	9	1.7
1cse	Small protein	63	172	10	11	2.6
1utg	All alpha	70	206	6	3	2.5
1hyp	All alpha	75	232	12	7	1.7
1lmb	All alpha	87	280	8	5	2.0
1plc	All beta	99	391	28	51	2.6
256b	All alpha	106	363	17	7	1.9
2mcm	All beta	112	414	36	36	2.3
2mhr	All alpha	118	352	17	17	2.3
1dz3	Alpha/beta	123	413	22	15	3.4
1mdc	All beta	131	474	40	22	2.3
1stm	All beta	141	521	68	77	4.1
1mba	All alpha	146	530	38	53	2.7
1byr	Alpha/beta	152	641	52	42	1.6
4dfr	Alpha/beta	159	598	49	49	2.3
3dfr	Alpha/beta	162	578	45	65	2.6
1v37	Alpha/beta	171	672	60	58	2.3
1dgw	All beta	178	617	65	52	3.9
1fvk	Alpha/beta	188	664	58	36	2.0
1o7n	All beta	193	622	77	85	3.7

to minimize some error function derived from the provided distance information. In this study, we use SMC method to recover protein structures.

Compared to global optimization methods, an important advantage of our approach is that it can generate a set of conformations that are properly weighted with respect to a specified target distribution. Hence, in addition to recovering structures, we can also provide estimate of important physical parameters of the molecular ensemble, including thermodynamics properties such as energy and entropies under a given energy function.[18,19] In this paper, the average rmsd to native structure for TSE conformations is a consistent estimate of how close the native structure and TSE satisfying the distance constraints indirectly provided by ϕ-values are.

A difficulty in growth-based method, such as SMC method, is that the distance information of future residues cannot be directly used for placing current residue. To circumvent this problem, we develop a new growth potential function that can incorporate the distance information of future residues. In this potential function, we convert upper bound constraints of distance for a subset of residue pairs to global distance upper bound constraints of all possible residue pairs. In addition, we introduce reference points of future residues to be placed.

We have used this algorithm to generate protein conformations from constraints in the form of small intervals of distances between a subset of residue pairs, from contact map, and from indirect distance constraints by ϕ-values. This algorithm can effectively recover native structures and can generate conformations satisfying any given set of dis-

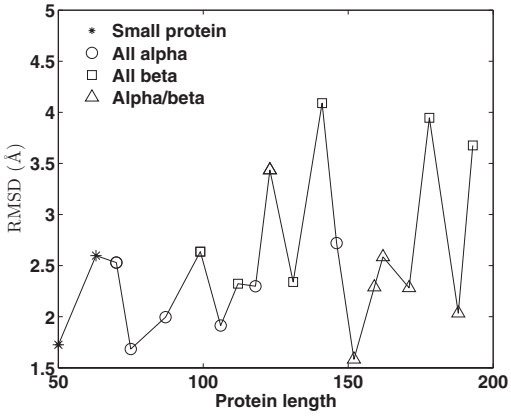

FIG. 7. rmsd of structures recovered from complete native contact maps to native structures for 20 proteins of different classes with length of 50–200. *X* axis is the protein length, *Y*-axis is the rmsd value of generated conformation that best fit the contact map to the native structure measured in Å.

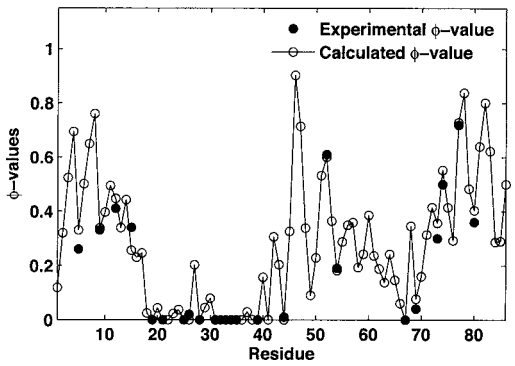

FIG. 8. Comparison of the experimental ϕ-values and calculated ϕ-values of the generated TSE of 1nvl. The filled circles represent the experimental ϕ-values, empty circles represent the calculated ϕ-values of the generated TSE.

tance constraints. The conformations generated by this method can also be used as the initial conformations for further refinement.[9–12]

In this study, a discrete model for protein structures was used for simplicity, at the price of model accuracy.[22] We expect further improvement by extending our model to continuous space, with additional steps of local move refinement, as demonstrated in Refs. 14 and 49.

ACKNOWLEDGMENTS

This work was supported by NIH Grant Nos. GM079804-01A1 and GM081682 and by NFS Grant Nos. DBI-0646035 and DMS-0800257.

[1] G. Rhodes, *Crystallography Made Crystal Clear: A Guide for Users of Macromolecular Models* (Academic, New York, 1999).
[2] G. M. Crippen and T. F. Havel, *Distance Geometry and Molecular Conformation* (Wiley, New York, 1988).
[3] W. Rieping, M. Habeck, and M. Nilges, Science **309**, 303 (2005).
[4] C. M. Falcon and K. S. Matthews, Biochemistry **40**, 15650 (2001).
[5] K. Cai, R. Langen, W. L. Hubbell, and H. G. Khorana, Proc. Natl. Acad. Sci. U.S.A. **94**, 14267 (1997).
[6] C. Altenbach, T. Marti, H. Khorana, and W. L. Hubbell, Science **248**, 1088 (1990).
[7] C. Altenbach, K. J. Oh, R. J. Trabanino, K. Hideg, and W. L. Hubbell, Biochemistry **40**, 15471 (2001).
[8] B. Berger, J. Kleinberg, and T. Leighton, J. ACM **46**, 212 (1999).
[9] J. J. Moré and Z. Wu, in *Global Minimization of Nonconvex Energy Functions: Molecular Conformation and Protein Folding*, edited by P. M. Pardalos, D. Shalloway, and G. Xue (American Mathematical Society, Providence, 1996), pp. 151–168.
[10] W. Glunt, T. L. Hayden, and M. Raydan, J. Comput. Chem. **14**, 114 (1993).
[11] J. J. Moré and Z. Wu, in *Encyclopedia of Nuclear Magnetic Resonance*, edited by D. M. Grant and R. K. Harris (Wiley, New York, 1995), pp. 1701–1710.
[12] A. Grosso, M. Locatelli, and F. Schoen, "Solving molecular distance geometry problems by global optimization algorithms," Optim. (to be published).
[13] G. A. Williams, J. M. Dugan, and R. B. Altman, J. Comput. Biol. **8**, 523 (2001).
[14] M. Vendruscolo, E. Kussell, and E. Domany, Folding Des. **2**, 295 (1997).
[15] M. Vendruscolo and E. Domany, Folding Des. **3**, 329 (1998).
[16] M. Vendruscolo, E. Paci, C. Dobson, and M. Karplus, Nature (London) **409**, 641 (2001).
[17] M. N. Rosenbluth and A. W. Rosenbluth, J. Chem. Phys. **23**, 356 (1955).
[18] J. Liang, J. Zhang, and R. Chen, J. Chem. Phys. **117**, 3511 (2002).
[19] J. Zhang, M. Lin, R. Chen, J. Liang, and J. S. Liu, Proteins **66**, 61 (2007).
[20] D. Frenkel and B. Smit, *Understanding Molecular Simulation: From Algorithms to Applications* (Academic, San Diego, 1996).
[21] D. P. Landau and K. Binder, *Monte Carlo Simulations in Statistical Physics* (Cambridge University Press, Cambridge, 2000).
[22] J. Zhang, R. Chen, and J. Liang, Proteins **63**, 949 (2006).
[23] J. J. Moré and Z. Wu, J. Global Optim. **15**, 219 (1999).
[24] G. Hom, S. Mayo, and N. Pierce, J. Comput. Chem. **24**, 232 (2002).
[25] A. E. Keating, V. N. Malashkevich, B. Tidor, and P. S. Kim, Proc. Natl. Acad. Sci. U.S.A. **98**, 14825 (2001).
[26] J. W. Moore and R. G. Pearson, *Kinetics and Mechanism* (Wiley, New York, 1981).
[27] A. R. Fersht, L. S. Itzhaki, N. Elmasry, J. M. Matthews, and D. E. Otzen, Proc. Natl. Acad. Sci. U.S.A. **91**, 10426 (1994).
[28] L. Li and E. I. Shakhnovich, Proc. Natl. Acad. Sci. U.S.A. **98**, 13014 (2001).
[29] S. Ozkan, I. Bahar, and K. A. Dill, Nat. Struct. Biol. **8**, 765 (2001).
[30] T. Lazaridis and M. Karplus, Science **278**, 1928 (1997).
[31] A. Li and V. Daggett, Proc. Natl. Acad. Sci. U.S.A. **91**, 10430 (1994).
[32] A. R. Fersht, R. J. Leatherbarrow, and T. N. Wells, Biochemistry **26**, 6030 (1987).
[33] G. Winter, A. R. Fersht, A. J. Wilkinson, M. Zoller, and M. Smith, Nature (London) **299**, 756 (1982).
[34] E. Paci, K. Lindorff-Larsen, C. Dobson, M. Karplus, and M. Vendruscolo, J. Mol. Biol. **352**, 495 (2005).
[35] S. E. Jackson, M. Moracci, N. ElMasry, C. M. Johnson, and A. R. Fersht, Biochemistry **32**, 11259 (1993).
[36] S. Miyazawa and R. Jernigan, Macromolecules **18**, 534 (1985).
[37] X. Li, C. Hu, and J. Liang, Proteins **53**, 792 (2003).
[38] J. Zhang, R. Chen, J. Liu, and J. Liang, Proteins **63**, 949 (2006).
[39] X. Li and J. Liang, *Computational Algorithms for Protein Structure Prediction* (Springer, New York, 2006).
[40] A. Marshall, in *Symposium on Monte Carlo Methods*, edited by M. Meyer (Wiley, New York, 1956), pp. 123–140.
[41] J. Liu and R. Chen, J. Am. Stat. Assoc. **93**, 1032 (1998).
[42] J. S. Liu, *Monte Carlo Strategies in Scientific Computing* (Springer, New York, 2001).
[43] P. Fearnhead and P. Clifford, J. R. Stat. Soc. Ser. B (Stat. Methodol.) **65**, 887 (2003).
[44] M. J. Atallah, *Algorithms and Theory of Computation Handbook* (CRC, Boca Raton, FL, 1998).
[45] A. S. Householder, *Principles of Numerical Analysis* (McGraw-Hill, New York, 1953).
[46] S. X. Chen and J. S. Liu, Stat. Sin. **7**, 875 (1997).
[47] Y. Chen, P. Diaconis, S. P. Holmes, and J. S. Liu, J. Am. Stat. Assoc. **100**, 109 (2005).
[48] B. B. Kragelund, P. Osmark, T. B. Neergaard, J. Schiødt, K. Kristiansen, J. Knudsen, and F. M. Poulsen, Nat. Struct. Biol. **6**, 594 (1999).
[49] J. Zhang, S. C. Kou, and J. S. Liu, J. Chem. Phys. **126**, 225101 (2007).