Generating folded protein structures with a lattice chain growth algorithm

Hin Hark Gan[a)]
Department of Chemistry and Courant Institute of Mathematical Sciences, New York University and the Howard Hughes Medical Institute, 31 Washington Place, Room 1021 Main, New York, New York 10003

Alexander Tropsha[b)]
Laboratory for Molecular Modeling, School of Pharmacy, University of North Carolina at Chapel Hill, CB# 7360, Beard Hall, Chapel Hill, North Carolina 27599-7360

Tamar Schlick[c)]
Department of Chemistry and Courant Institute of Mathematical Sciences, New York University and the Howard Hughes Medical Institute, 251 Mercer Street, New York, New York 10012

(Received 17 March 2000; accepted 3 July 2000)

We present a new application of the chain growth algorithm to lattice generation of protein structure and thermodynamics. Given the difficulty of *ab initio* protein structure prediction, this approach provides an alternative to current folding algorithms. The chain growth algorithm, unlike Metropolis folding algorithms, generates independent protein structures to achieve rapid and efficient exploration of configurational space. It is a modified version of the Rosenbluth algorithm where the chain growth transition probability is a normalized Boltzmann factor; it was previously applied only to simple polymers and protein models with two residue types. The independent protein configurations, generated segment-by-segment on a refined cubic lattice, are based on a single interaction site for each amino acid and a statistical interaction energy derived by Miyazawa and Jernigan. We examine for several proteins the algorithm's ability to produce nativelike folds and its effectiveness for calculating protein thermodynamics. Thermal transition profiles associated with the internal energy, entropy, and radius of gyration show characteristic folding/unfolding transitions and provide evidence for unfolding via partially unfolded (molten-globule) states. From the configurational ensembles, the protein structures with the *lowest* distance root-mean-square deviations (dRMSD) vary between 2.2 to 3.8 Å, a range comparable to results of an exhaustive enumeration search. Though the *ensemble-averaged* dRMSD values are about 1.5 to 2 Å larger, the lowest dRMSD structures have similar overall folds to the native proteins. These results demonstrate that the chain growth algorithm is a viable alternative to protein simulations using the whole chain.
[S0021-9606(00)50337-7]

I. INTRODUCTION

Current equilibrium simulations of proteins are based on Metropolis algorithms[1] and, in more recent years, multicanonical[2] or entropy[3–6] sampling schemes. These algorithms generate valuable structural and thermodynamic properties of peptides and proteins. Despite advances in algorithmic development, *ab initio* structure prediction remains challenging, and the computation of thermodynamic functions is costly. Thus, it is worthwhile exploring a third alternative to these problems based on the chain growth algorithm.

This algorithm was originally developed by Rosenbluth and Rosenbluth for self-avoiding walk chains[7] and extended by Meirovitch to chains with attractive potentials.[8,9] Meirovitch and others later applied algorithmic extensions to single-chain polymers,[8,9] simple protein models with two residue types (hydrophobic and hydrophilic) or HP models,[10] and peptides.[11] More recently, Grassberger and co-workers introduced the pruned-enriched Rosenbluth method for polymers and HP protein models.[12] Their method incorporates a mechanism for favoring the selection of chain growth configurations with high statistical weights over those with low weights.

An important element of chain growth schemes is that the transition probability for growing the links is guided by Boltzmann statistics so as to generate configurations that contribute significantly to the thermodynamic average at a given temperature. For example, compact configurations are sampled more frequently than open configurations at low temperatures; the reverse is true at high temperatures. Meirovitch[8] developed a scanning method where future continuations of the chain (involving several links) are searched before the current growth direction is selected. Meirovitch and co-workers found that this scanning approach samples more efficiently peptide conformations.[11] In addition, the thermodynamic free energy of chain molecules[11] could be evaluated efficiently, as well as computation of their thermodynamic transition profiles.[10] These features of the chain growth algorithm offer valuable information when analyzing

[a)] Electronic mail: hgan@biomath.nyu.edu

[b)] Electronic mail: alex_tropsha@unc.edu

[c)] Author to whom all correspondence should be addressed. Electronic mail: schlick@nyu.edu conformations and thermodynamic properties of polymers and proteins.

Another proposed variant of the chain growth idea is the configurational-biased Monte Carlo method.[13] In this algorithm, a new configuration of the chain is sampled by regrowing only part of the chain, and the acceptance or rejection of the new configuration is judged by an appropriately weighted transition probability. Hao and Scheraga have applied this configuration sampling procedure successfully to their entropy sampling Monte Carlo simulations of β-protein models with polar (P), hydrophobic (H), and neutral (N)[3–6] residue types.

Apart from the chain growth approach, simulations of protein folding with Metropolis Monte Carlo methods have been found to be quite successful in structure prediction, especially with distance constraints obtained either from NMR or theoretical predictions.[1,14] Furthermore, Metropolis methods can yield folding pathway information[15] while chain growth approaches generate statistically independent configurations. Since the chain growth configurations are not dynamically connected, they can lead to enhanced exploration of configuration space.

Here we apply the chain growth method to prediction of protein structure and thermodynamics from the amino acid sequence. We determine the quality of the algorithm for this application by analyzing the computed configurational ensembles and thermodynamic functions. We use a low-resolution protein model with the two-body Miyazawa–Jernigan (MJ) residue interaction potential.[16,17] Although such a model is not expected to yield accurate folded native protein structures, it can be used to rapidly compute protein thermodynamics and assess the ability of the algorithm to generate nativelike configurations for known test cases. When additional experimental or theoretical information is incorporated, such approaches are also viable prediction tools.

Specifically, we describe results of the chain growth method guided by the MJ contact potential on Cd-7 Metallothionein-2 protein (30 residues, PDB code 2mhu) and helical proteins 434 Repressor (63 residues, 1r69) and 434 Cro (65 residues, 2cro). Applications are also performed for Protein G (56 residues, 1pgb) and ColE1 Repressor of Primer protein (63 residues, 1rop), both associated with the solution to the Paracelsus challenge.[18] We analyze the resulting configurational ensembles using a novel statistical-weight-based scheme to select nativelike conformations. We also calculate several thermodynamic properties such as internal energy, entropy, and free energy at different temperatures using both simulations and theory, the latter of which is based on an analytical extrapolation formula. We find a good agreement between simulated and predicted thermal transition curves. Moreover, the thermal transition profiles show evidence of unfolding via molten globule or intermediate states. From the configurational ensembles generated, the lowest distance RMSD (dRMSD) structures have the correct overall folds with dRMSD of 2–4 Å or coordinate RMSD (cRMSD) of 3–6 Å as compared to native structures. Furthermore, our method can explore these reasonable structures rapidly, faster than unbiased searches or dynamic schemes which may be confined to small regions of configurational space. The results thus demonstrate that the chain growth algorithm is a viable alternative for protein structure and thermodynamics studies.

In Sec. II, we present the protein model and associated lattice and energy functions. The methodology of the chain growth approach is described in Sec. III, which includes the process of chain generation, factors affecting the algorithm's convergence, thermodynamic functions, and an efficient method for evaluating thermal averages. Section IV describes the analysis of configurational ensembles and protein thermodynamics. We summarize our findings and conclusions in Sec. V. In the Appendix, we elaborate on some aspects of the chain growth algorithm.

II. PROTEIN MODEL, LATTICE MOVES, AND POTENTIAL FUNCTION

The choice of an appropriate lattice/protein model represents a balance between the accuracy of the attainable results and the overall computational complexity. Lattice geometries that have been used to simulate protein structures range from low to high resolution models. Examples of low resolution lattices include simple cubic[19,20] and diamond[21] models. Higher resolution lattices for protein simulations include octahedral[22] and a family of refined cubic[15,23,24] models. Low resolution lattices are more efficient in sampling the conformational space—even exact enumeration of all compact configurations for short chains is possible, such as for a diamond lattice[21]—but high resolution lattices are necessary to reproduce more accurately the secondary and tertiary structural elements.

Here, we consider a protein model defined on a moderate resolution cubic lattice that is similar to the family investigated by Kolinski and Skolnick.[24] Each residue is represented by an interaction site and the residues interact via the two-body MJ potential.[16,17] In the following, we describe the lattice used, the protein model parameters, and the form of the pairwise interaction potential.

A. Geometry of protein model and lattice moves

In our simplified protein model, the interaction sites are located at the C_α positions; side chains are not modeled. We reproduce the geometric characteristics of polypeptide chains by restricting the C_α pseudobond angles and by preventing the overlap of excluded-volume sites. The pseudobond angles are limited to the range of 63°–143°; excluded-volume sites will be introduced below after the lattice is defined. Our objective is to grow configurations of lattice proteins that satisfy these geometric constraints. Protein configurations on lattices are generally not invariant with respect to rotations, although the effects of anisotropy are not severe for refined cubic lattices with high coordination numbers.

The cubic lattice we employ is a (311) model with 24 allowed moves from a given site. On this lattice, the possible growth directions are given by the direction vectors $\mathbf{v}=x\hat{\mathbf{i}}+y\hat{\mathbf{j}}+z\hat{\mathbf{k}}$ where $(x,y,z)\in\{(\pm1,\pm3,\pm1),(\pm3,\pm1,\pm1),(\pm1,\pm1,\pm3)\}$. We illustrate the related (31) lattice move directions in 2D in Fig. 1(a). In the actual chain growth

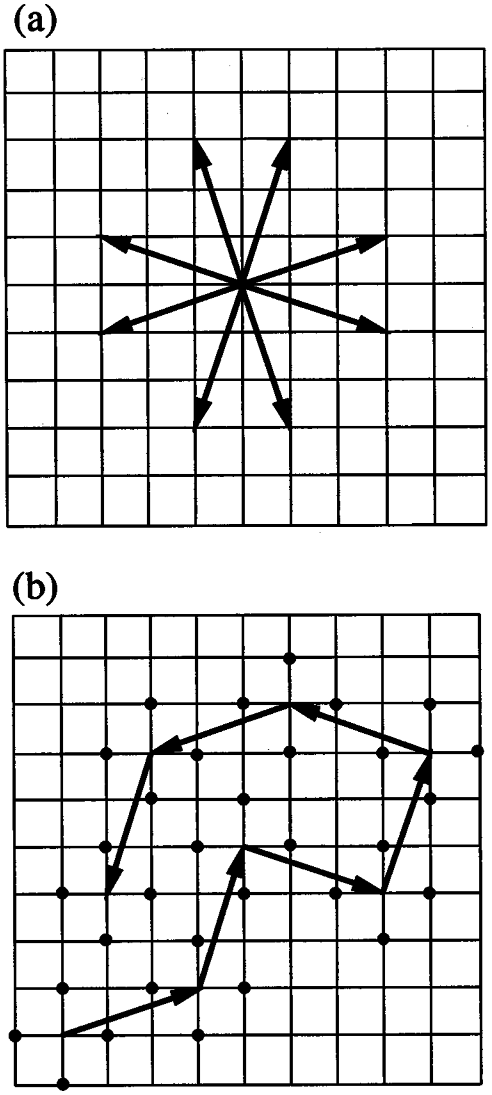

FIG. 1. Illustration of the chain growth procedure on a 2D lattice: (a) lattice vectors of the (31) moves of a growing chain and (b) associated allowed configuration in 2D. The constant vector length (bond length) determines the spacing of the underlying square lattice. The growth directions of the allowed configuration (b) are prescribed by (31) lattice moves in (a). The chain vertices represent the C_α positions with excluded-volume sites placed at its nearest-neighbor sites (filled circles). Possible new growth directions must not overlap with any of the previous C_α, the excluded-volume sites, and the pseudo-bond angle constraints.

implementation, many of these move directions are prohibited by the pseudobond angle restrictions and the excluded-volume requirements (see below).

The lattice parameters and mapping accuracy for protein molecules depend on the type of lattice. Since the distance between two adjacent C_α positions in proteins is 3.8 Å , the lattice spacing in our (311) lattice is $L=3.8/\sqrt{x^2+y^2+z^2}$ Å $=1.146$ Å . The discrete nature of the formulation leads to an accuracy for lattice-mapped structures of native proteins of about 1.5 Å in cRMSD, as found empirically. Better mapping resolutions (e.g., cRMSD of 0.7 Å or less) can be achieved at the expense of computational cost with lattices that have greater number of move vectors and more sophisticated move protocols.[24] Still, our current resolution of 1.5 Å is sufficient to reproduce elements of the secondary structures and overall protein folds.

The possible move directions on the (311) lattice are further restricted by the condition that the near neighbors of C_α sites do not overlap. On the (311) lattice in 3D, there are 26 such near neighbors which are separated from the central C_α site by the vectors $\{(\pm1,\pm1,\pm1),(\pm1,\pm1,0),(\pm1,0,\pm1),(0,\pm1,\pm1),(\pm1,0,0),(0,\pm1,0),(0,0,\pm1)\}$. Note that the magnitudes of lattice vectors are in units of $L=1.146$ Å. These excluded-volume sites, together with the finite repulsive energy at short distances described below, define the geometric characteristics of protein backbones. We illustrate in Fig. 1b an allowed configuration of a growing chain on a (31) lattice in 2D.

B. Potential function

We choose a simple square-well function to parametrize the residue/residue potential. The attractive interactions are represented by the residue-based contact energies of statistical potentials which are derived from solved protein structures. For each i and j pair representing two residues, with distance separation R_{ij}, our potential has the form

$$u_{ij}(R_{ij})=\begin{cases}\epsilon_r & \text{if } R_{ij}<4\text{ Å}\\ \epsilon_{ij} & \text{if } 4\text{ Å}\leqslant R_{ij}\leqslant 6.5\text{ Å},\\ 0 & \text{if } R_{ij}>6.5\text{ Å}\end{cases} \tag{1}$$

where ϵ_r is a residue-independent finite repulsive energy, and ϵ_{ij} is a contact-energy value derived from experimental protein databases. The short-range repulsive energy ensures minimal overlap between protein cores. The value of ϵ_r is set simply as: $\epsilon_r=5\max_{\{ij\}}|\epsilon_{ij}|$; we found results to be insensitive to the precise value of ϵ_r within a wide range.

The contact energies $\{\epsilon_{ij}\}$ of statistical potentials are derived from the observed occurrences of residue pairs i,j in protein structure databases. Statistical potentials assume that the contact energy ϵ_{ij} is related to the observed occurrences of residue pairs i,j via Boltzmann's relation: $\epsilon_{ij}=-k_BT\ln(F_{ij}/R_{ij})$, where F_{ij} is the observed contact frequency for the pair i,j, and R_{ij} is its corresponding reference state.[16,25,26]

The reference state R_{ij} has been defined based on both the random or uncorrelated residue pairs and the solvent-mediated contact pairs in protein structures. For the random reference state, $R_{ij}\propto n_i n_j$ where n_i,n_j are the frequencies of residues in the protein structure database. The solvent-mediated reference state is calculated as follows:[17] $R_{ij}=n_{si}n_{sj}/n_{ss}$ where n_{si},n_{sj} refer to solvent (s) with residue (i or j) contact numbers in protein structures; n_{ss} is solvent/solvent contact number which is evaluated by using the estimated number of effective solvent molecules in native protein structures. The choice of the reference state affects the scale of the residue contact energies. Contact energies $\{\epsilon_{ij}\}$ calculated with the random reference state reflect the residual nonrandom preferences for residue/residue contacts in proteins; energies calculated with solvent-mediated reference state are related to the preferences for residue/residue over residue/solvent contacts. The MJ energies are derived using a solvent-mediated reference state and are correlated with experimental hydrophobic energies of residues.[16] Thus they implicitly incorporate effects due to dispersive, electrostatic, and other interactions that are present in protein structures. The attractive hydrophobic energies drive the protein chain toward a folded conformation. Since the MJ approach models short-range contacts, there are neither long-range correlations nor specialized terms for helices, sheets, etc.

We represent the pairwise energies ϵ_{ij} by a modified MJ matrix. The modification involves a simple shift: $\epsilon_{ij} \leftarrow M_{ij} + 2$, where M_{ij} is the MJ interaction matrix as re-evaluated in 1996;[17] the energies are expressed in $k_B T_0$ units, where k_B is Boltzmann's constant and T_0 is the room temperature. This modification is effectively similar to the statistical potential derived by Skolnick and co-workers[27] in modeling the residue/residue contact energies with a random reference state that accounts for chain connectivity rather than the solvent-mediated as in the original MJ formulation.[16] Betancourt and Thirumalai[28] have shown that the interaction energies of Skolnick *et al.*[27] can be related to the MJ energies by a linear fit $\tilde{\epsilon}_{ij} \leftarrow a M_{ij} + b$ where a and b are constants ($a = 0.37$ and $b = 1.26$ in Ref. 28) in $k_B T_0$ units. Our simple shift similarly weakens the attractive energies between the residues and was found to yield reasonable global properties of proteins in our simulation protocol. To arrive at our constant shift value, we experimented with a range of values; large shifts favor open structures, and small shifts tend to result in overcompact structures.

III. METHODS

This section presents the temperature-dependent chain growth transition probability, thermal averages derived from the importance sampling procedure, and an efficient technique for computing thermal curves. This is followed by a brief discussion on the factors that affect the convergence of the chain growth algorithm. We also derive thermodynamic averages for chain growth configurations. Detailed formulations of the chain growth algorithm are found in the Appendix.

A. The chain growth method: Overview

Here we integrate central concepts in the chain growth algorithm, reported previously for different systems; quantitative formulations are provided in the Appendix. Essentially, the chain growth algorithm generates chain configurations by sequential addition of links until the full length of the chain is reached. Since each configuration is generated *de novo*, configurations are statistically independent, unlike the standard Metropolis algorithms. For HP protein models, comparisons between the Metropolis and chain growth approaches show that the latter is more efficient for sampling as well as for computation of thermal curves.[10] These main findings motivate our present application of the algorithm to proteins.

For chains where the properties are independent of temperature, all possible self-avoiding growth directions have equal transition probabilities. To treat self-avoiding and temperature-dependent polymers, Meirovitch[8] used the Boltzmann-weighted transition probability, which is a modification of the form for athermal chains. This transition probability favors growth directions with low energy contacts; it allows generation of both open and compact chain configurations depending on the temperature. This form of the transition probability is not unique, and other forms have been used by Grassberger and co-workers to study polymer and simple protein systems.[12] Here we use the transition probability proposed by Meirovitch.

Greater efficiency in chain growth algorithms can be introduced by adding more than one link at each step. This procedure confers greater foresight to growth process and reduces the frequency of ''dead-end'' configurations (i.e., those that cannot be fully grown given the near neighbor constraints; see below). Examples of multilink steps are the scanning[8] and double-scanning[9] procedures, and multilink insertion technique.[10] Since these procedures require assessment of lattice occupancy and energy evaluations of possible future configurations, greater computational costs are involved especially for high coordination lattices. For this reason, we use a simple approach of one link per step; see also Sec. III D on convergence.

After fully grown chains are generated by the chain growth method, ensemble averages are calculated using an importance sampling procedure where each contributing configuration is appropriately weighted.[11] The weight is derived from the known probability of generating a configuration. A unique feature of the chain growth algorithm is that the temperature dependence of ensemble averages can be efficiently computed using an analytic extrapolation technique.[10] This technique greatly facilitates calculation of thermal transition curves, often costly to reproduce by other methods.

B. Chain generation by temperature-dependent transition probability and computing thermal averages

The direction chosen at each step of the chain growth process is determined by the transition probability. If the first $i-1$ links of a chain with N links have been placed, the temperature-dependent transition probability P_i at step i, as proposed by Meirovitch,[8,11] is given by

$$P_i(\mathbf{R}_i + \mathbf{v}_{k_i} | \mathbf{R}_1, \ldots, \mathbf{R}_i; \beta) = \exp[-\beta u_i(\mathbf{R}_i + \mathbf{v}_{k_i})] \Big/ \sum_{k_i=1}^{C_i} \exp[-\beta u_i(\mathbf{R}_i + \mathbf{v}_{k_i})], \tag{2}$$

where the incremental, nonbonded potential energy is

$$u_i(\mathbf{R}_i + \mathbf{v}_{k_i}) = \sum_{j=1}^{i-1} u_{ij}(R_{ij}). \tag{3}$$

Other symbols are defined as follows: temperature parameter $\beta = 1/k_B T$; $\mathbf{R}_1, \ldots, \mathbf{R}_i$ are position vectors of (interaction) sites $1,2,\ldots,i$; $\mathbf{v}_{k_i}$ is the lattice vector for the chosen direction k_i; and C_i is the number of vacant sites at step i. Since the growing chain configuration must be self-avoiding, the transition probability depends on the coordinate vectors of sites $1,2,\ldots,i$; P_i is also normalized to unity. The above temperature-dependent transition probability favors configuration states with large Boltzmann weights and/or growth directions with favorable energetic contacts. Configurations thereby generated are compact at low temperatures and more open at high temperatures.

We generate chains on (311) cubic lattice guided by the transition probability (2). The first link can be placed in any direction, and the move directions for subsequent links are selected according to P_i by a Monte Carlo procedure (see the Appendix). This growth process is continued until the entire chain length (N links) is reached. Each configuration Λ is generated with a probability

$$P_\Lambda(\beta)=\prod_{i=1}^{N} P_i(\mathbf{R}_i+\mathbf{v}_{k_i}|\mathbf{R}_1,\ldots,\mathbf{R}_i;\beta)$$
$$=\exp(-\beta E_\Lambda)/W_\Lambda(\beta), \tag{4}$$

where E_Λ is the energy for configuration Λ and the statistical weight

$$W_\Lambda(\beta)=\prod_{i=1}^{N}\left\{\sum_{k_i=1}^{C_i}\exp[-\beta u_i(\mathbf{R}_i+\mathbf{v}_{k_i})]\right\}. \tag{5}$$

If a dead-end configuration, i.e., $C_i=0$, is encountered before the chain is fully grown, the growth process is terminated and the chain discarded, the next chain is then regrown from scratch. In this way, successive configurations are not correlated. To obtain accurate estimates of thermal averages, we generate millions of configurations.

Once configurations are generated as above, the average of a property A in canonical ensemble is given by

$$\langle A\rangle_\beta=\sum_{\{\Lambda\}} A_\Lambda W_\Lambda(\beta)\Big/\sum_{\{\Lambda\}} W_\Lambda(\beta), \tag{6}$$

where A_Λ is the value of property A for configuration Λ. We have used an importance sampling procedure to obtain the above statistical average, which ensures that the (biased) chain growth configurations are assigned appropriate weights.[11] All successfully grown configurations are counted, each with a statistical weight W_Λ. The statistical weight,

$$W_\Lambda(\beta)=P_\Lambda(\beta)^{-1}\exp(-\beta E_\Lambda), \tag{7}$$

represents a correction of the Boltzmann weight ($\exp[-\beta E_\Lambda]$) of the canonical ensemble. Since each chain growth configuration is generated with a probability $P_\Lambda(\beta)$, the correction factor $P_\Lambda(\beta)^{-1}$ in Eq. (7) removes this bias (see also the Appendix). To ensure accuracy and convergence of the average $\langle A\rangle_\beta$, the size of the configurational sample must be sufficiently large so that configurations with large weights belong to the sample.

C. An extrapolation technique for calculating thermal profiles

Temperature-dependent quantities are expensive to compute because many configurational ensembles are often needed to reproduce thermal profiles. An efficient way of obtaining thermal curves is by an analytic extrapolative technique employed by O'Toole and Panagiotopoulos[10] where only a few configurational samples are needed. In this approach, the thermal average $\langle A\rangle_{\beta'}$ can be computed based on configurational ensemble generated at the temperature parameter β as follows:[10]

$$\langle A\rangle_{\beta'}\equiv\sum_{\{\Lambda\}} A_\Lambda W_\Lambda(\beta')\Big/\sum_{\{\Lambda\}} W_\Lambda(\beta')$$
$$\approx\sum_{\{\Lambda\}} A_\Lambda S_\Lambda(\beta',\beta)\Big/\sum_{\{\Lambda\}} S_\Lambda(\beta',\beta), \tag{8}$$

where

$$S_\Lambda(\beta',\beta)=W_\Lambda(\beta)\exp[-(\beta'-\beta)E_\Lambda]. \tag{9}$$

In this approximation of $\langle A\rangle_{\beta'}$, the statistical weight $W_\Lambda(\beta')$ is effectively replaced by the Boltzmann-corrected weight $S_\Lambda(\beta',\beta)$. Equation (8) is exact when $\beta=\beta'$ and a poor approximation of $\langle A\rangle_{\beta'}$ for non-negligible values of $|\beta'-\beta|$. The approximation is reasonable because energy distributions of two configurational ensembles at β' and β overlap significantly when $|\beta'-\beta|$ is small. Hence, the above extrapolation of thermal averages is justified within a certain temperature range.

D. Factors affecting convergence

The convergence of the average property $\langle A\rangle_\beta$ depends on the size of the configurational ensemble $\mathcal{N}$, the number of links b placed at each step of the growth process, and the lattice coordination number n_c. Since the chain growth algorithm explores all available growth directions at each step, the number of energy evaluations per step is proportional to $(n_c)^b$. The cost per step must be balanced with the overall added efficiency possible by the ''scanning'' approach ($b>1$) of Meirovitch[8] and its variants such as double-scanning procedure[9] and multilink insertion technique,[10] all of which are generally efficient for polymers when implemented on low coordination square[9] and cubic[10] lattices. For our (311) lattice with $n_c=24$, $(n_c)^b$ is large even for small values of b.

We thus choose $b=1$ but compensate by generating a large sample size on the order of 10 million. Larger b would mean reducing the sample size $\mathcal{N}$ and likely to decrease the effectiveness of the importance sampling approach to computing thermal averages. As our thermodynamic results show, our statistical fluctuations are not large, but larger b values might be considered in the future.

Our simulations are performed on a 300 MHz R12000 SGI Origin2000 computer at New York University. Sampling 5 million configurations for a 30 residue protein requires about 10 CPU hours.

E. Thermodynamic functions

We now present the expressions for the thermodynamic energy, free energy, entropy, and heat capacity. These thermal averages are computed based on the importance sampling procedure and the analytic extrapolation technique outlined in Secs. III B and III C (see also subsection 2 of the Appendix). The changes in thermodynamic quantities are more meaningfully measured with respect to a reference state. The reference state for a protein chain molecule is taken to be the ideal random walk chain on the (311) lattice.

We define the excess free energy as

$$F(\beta)=-k_BT\ln[Z(\beta)/Z_{\mathrm{ref}}], \tag{10}$$

where $Z(\beta)$ and Z_{ref} are the configuration partition functions of the protein and reference chains, respectively. For an ideal random walk chain on the lattice, the number of states is $(n_c)^N$ where $n_c=24$ for the (311) lattice. Therefore, the exact reference partition function is

$$Z_{\mathrm{ref}}=\sum_{\Lambda=1}^{\mathcal{N}} n_c^N=\mathcal{N}n_c^N, \tag{11}$$

where $\mathcal{N}$ is the number of configurations. As expected, Z_{ref} only depends on the chain and lattice characteristics.

We rewrite the protein partition function, $Z(\beta)=\Sigma_{\{\Lambda\}}\exp(-\beta E_\Lambda)$, using an importance sampling procedure for chain growth configurations as $Z(\beta)=\Sigma_{\{\Lambda\}}W_\Lambda(\beta)$. When the extrapolation technique is employed to evaluate temperature dependence of $Z(\beta)$, we have $Z(\beta)=\Sigma_\Lambda S_\Lambda(\beta,\beta')$; therefore the excess free energy function is given by

$$F(\beta)=-k_BT\ln\left[\frac{1}{Z_{\mathrm{ref}}}\sum_{\Lambda=1}^{\mathcal{N}}S_\Lambda(\beta,\beta')\right]. \tag{12}$$

Similarly, the excess internal energy is given by

$$E=\langle E\rangle_\beta=\sum_{\{\Lambda\}}E_\Lambda S_\Lambda(\beta,\beta')\Bigg/\sum_{\{\Lambda\}}S_\Lambda(\beta,\beta'), \tag{13}$$

where $E_\Lambda=\Sigma_{i<j}^N u_{ij}$ and, from thermodynamic relations, the excess entropy is

$$S/k_B=\beta(E-F). \tag{14}$$

While we might also evaluate the entropy and free energy by using the probability of configurations P_Λ, such estimates lead to results that are poorer than the weight-corrected formulas above.[11]

The evaluation of specific heat is of significant interest because this quantity can be measured in calorimetry experiments.[29,30] The specific heat capacity C corresponds to second energy moment or energy fluctuations, given as

$$\frac{C}{k_B}=\frac{d\langle E\rangle_\beta}{dT}=\beta^2[\langle E^2\rangle_\beta-\langle E\rangle_\beta^2]. \tag{15}$$

These thermodynamic functions are expressed in a form that allows rapid evaluation of thermal transition curves. Below we compare this method of determining thermodynamic functions to direct simulation results at different temperatures.

IV. RESULTS AND DISCUSSION

A. Analysis of configurational ensembles

1. Energy distributions

The energy distribution of configurations reflects the nature of the algorithm used to produce it. Below, we characterize the energy distributions, examine average energies in comparison to native energies, and discuss the relationship between statistical weights and energies.

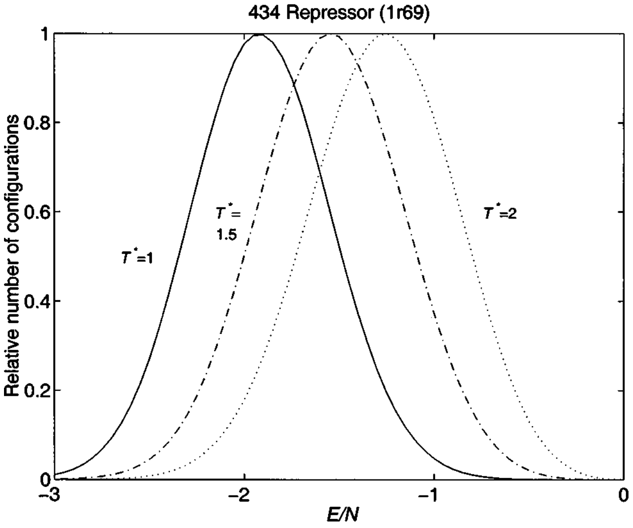

FIG. 2. Energy distribution of configurations for a helical protein 434 Repressor (1r69) at reduced temperatures $T^*=T/T_0=1$, 1.5, and 2 where $T_0=298$ K. Ensembles of 10 million configurations were generated to produce the curves. E/N is the excess internal energy per residue, expressed in units of k_BT_0.

Figure 2 shows the energy distributions of protein 434 Repressor (1r69) for ensembles generated at three different temperatures: $T^*=T/T_0=1$, 1.5, and 2, where T_0 is the room temperature (298 K). No weighting by $W_\Lambda(\beta)$ was done to obtain these distributions. We see that, as expected, higher temperature distributions are peaked at higher energies. The distributions are nearly Gaussian and they have similar widths. The widths can be parametrized by a function of the form $\exp\{-[E-E_0(T)]^2/2\sigma^2\}$ where $E_0(T)$ and σ specify the location of the peak and width of the distribution, respectively. This characterization arises since configurations $\{\Lambda\}$ and corresponding energies $\{E_\Lambda\}$ of the chain growth algorithm are statistically uncorrelated.

The energy distributions in Fig. 2 also show that the energy states at different temperatures overlap significantly. This suggests that an ensemble generated at a temperature can be used to approximate thermodynamic averages at neighboring temperatures according to the analytic extrapolation of Eq. (8). Indeed, we show later that only a few ensembles are needed to reproduce the temperature curves for various average properties.

Next we assess the energies and configurations of the chain growth ensemble compared to the native states. Table I shows that the ensemble-averaged energy, $\langle E\rangle_\beta$, of 1r69 is considerably lower than the value for the native protein, which is close to the peak of the energy distribution (Fig. 2) at room temperature. This follows the fact that the larger weights are associated with the lower energy configurations. Indeed, observations (data not shown) give an almost linear correlation between weight $\ln W_\Lambda(\beta)$ and energy $-E_\Lambda/k_BT$ at $T^*=1$. Hence, $W_\Lambda(\beta)$ approximates the Boltzmann weight at low temperature and, from Eq. (4), P_Λ is a constant of temperature and energy. In other words, configu-

TABLE I. Comparison of the global properties of chain growth (CG) configurations and native structures for three proteins. Results include excess internal energy per residue (E/N), mean radius of gyration (R_G), ensemble-averaged dRMSD ($\sqrt{\langle D^2_{\mathrm{drms}}\rangle}$) and the lowest dRMSD values in the generated ensembles along with their corresponding cRMSD values. The native values of energy refer to computed energies of C_α proteins in their native configurations. The energy is expressed in units of k_BT_0 where T_0 is room temperature and the RMSD values are in Å .

Proteins		Size (N)	E/N	R_G (Å)	$\sqrt{\langle D^2_{\mathrm{drms}}\rangle}$	Lowest dRMSD	Lowest cRMSD
2mhu	native	30	−1.46	8.47			
2mhu	CG	30	−2.93	7.15	3.71	2.18	3.27
1r69	native	63	−1.92	10.22			
1r69	CG	63	−3.90	10.46	5.74	3.82	5.48
2cro	native	65	−2.11	10.22			
2cro	CG	65	−4.85	10.24	5.37	3.82	5.67

rations generated at $T^*=1$ are unbiased. By contrast, for athermal chains or high temperature situations, the open configurations have larger weights than compact ones since $W_\Lambda \sim \Pi_i^N C_i$. This predicted lower equilibrium energy of generated configurations could be attributed to our use of relatively simple MJ interaction energies and possibly the protein model as well. The calculated radius of gyration, however, agrees well with the native value, as also shown in Table I. Thus, it may be generally more difficult to reproduce the native energy than the size and overall structure. These observations also hold for the proteins 2mhu and 2cro (see Table I).

2. dRMSD distributions and 3D structures

The RMSD distribution yields a more direct measure of the quality of configurational ensembles than the energy distribution. The distance RMSD, D_{drms}, is defined by

$$D_{\mathrm{drms}} = \sqrt{\frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}(D_{ij}-D_{ij}^{\mathrm{nat}})^2}, \tag{16}$$

where D_{ij} and D_{ij}^{nat} are distances between residues i,j of a generated configuration and the native protein, respectively. We use the dRMSD measure to allow rapid evaluation of deviations of about 10 million configurations from the native structure. We also compute C_α cRMSD, obtained by optimal superposition of structures using Kabsch's method.[31,32] The cRMSD value is typically about 40%–50% larger than that of the dRMSD.

As shown in Fig. 3, the dRMSD distributions for the proteins 2mhu and 2cro are rather broad, but the peaks are located at the lower ends of the distributions which means that most configurations have lower dRMSD values. Conversely, a distribution peaked at the high end of the dRMSD range would indicate a poor algorithm or/and energy function. Certainly, the goal is to devise a selection criterion to pick the structures in the ensemble with the lowest dRMSD values. This is discussed in the next subsection, but now we examine the best structure or the one with the lowest dRMSD value in a given ensemble of configurations. We see in Table I that the best structures have a dRMSD value of 2.2 Å for the small protein 2mhu (30 residues) and 3.8 Å for the larger proteins 1r69 (63 residues) and 2cro (65 residues). For comparison, the best dRMSD values obtained by Hinds and Levitt[21] using an exact enumeration method on diamond lattice are in the range of 3.17– 4.01 Å for proteins having 52 to 68 residues. Thus our results are comparable in quality with those by Hinds and Levitt.

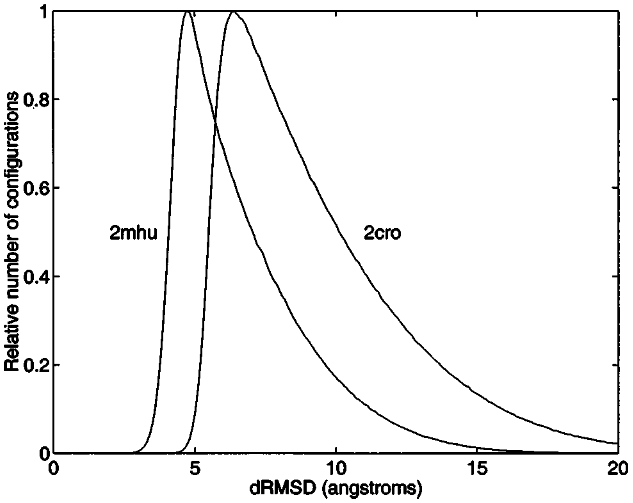

FIG. 3. Distribution of distance root-mean-square deviations (dRMSD) for Cd-7 Metallothionein-2 (2mhu) and 434 Cro (2cro) proteins at reduced temperature $T^*=1$ as computed from an ensemble of 5 million configurations for 2mhu and 10 million configurations for 2cro.

Some of the factors determining the lowest dRMSD value obtainable include quality of the algorithm and energy function, the sequence length, and nature of protein fold. For example, as the sequence length increases the number of possible configurations grows exponentially, making it much more difficult to obtain low dRMSD values. By comparing the best dRMSD values of 2mhu (30 residues) and 2cro (65 residues), we see that doubling the protein length nearly doubles the dRMSD value. Moreover, simple protein folds are expected to be easier to generate than complicated protein folds, as the comparisons of three-dimensional structures below illustrate.

Best dRMSD structures (with corresponding cRMSD value) for the three test proteins 2mhu, 1r69, and 2cro are compared to their native structures in Fig. 4. These configurations with cRMSD of about 3–6 Å or dRMSD of 2–4 Å reproduce the rough overall folds of their native structures. However, secondary structures of 1r69 and 2cro are not evident in these calculated configurations; specific short range potentials are needed to reproduce these features.

To further explore the performance of our approach on a relevant protein with unknown native structure, we consider the Janus sequence, a solution to the Paracelsus challenge of Rose and Creamer:[33] transform the conformation of a parent globular protein into a target protein by altering less than 50% of the parent protein sequence. The solution to the challenge, based on the parent (1pgb) and target (1rop) proteins,[18] is the Janus sequence that has an experimentally inferred fold that is similar to the target protein 1rop.[18] However, its experimental three dimensional structure has not been solved. We use our algorithm to generate the folds of

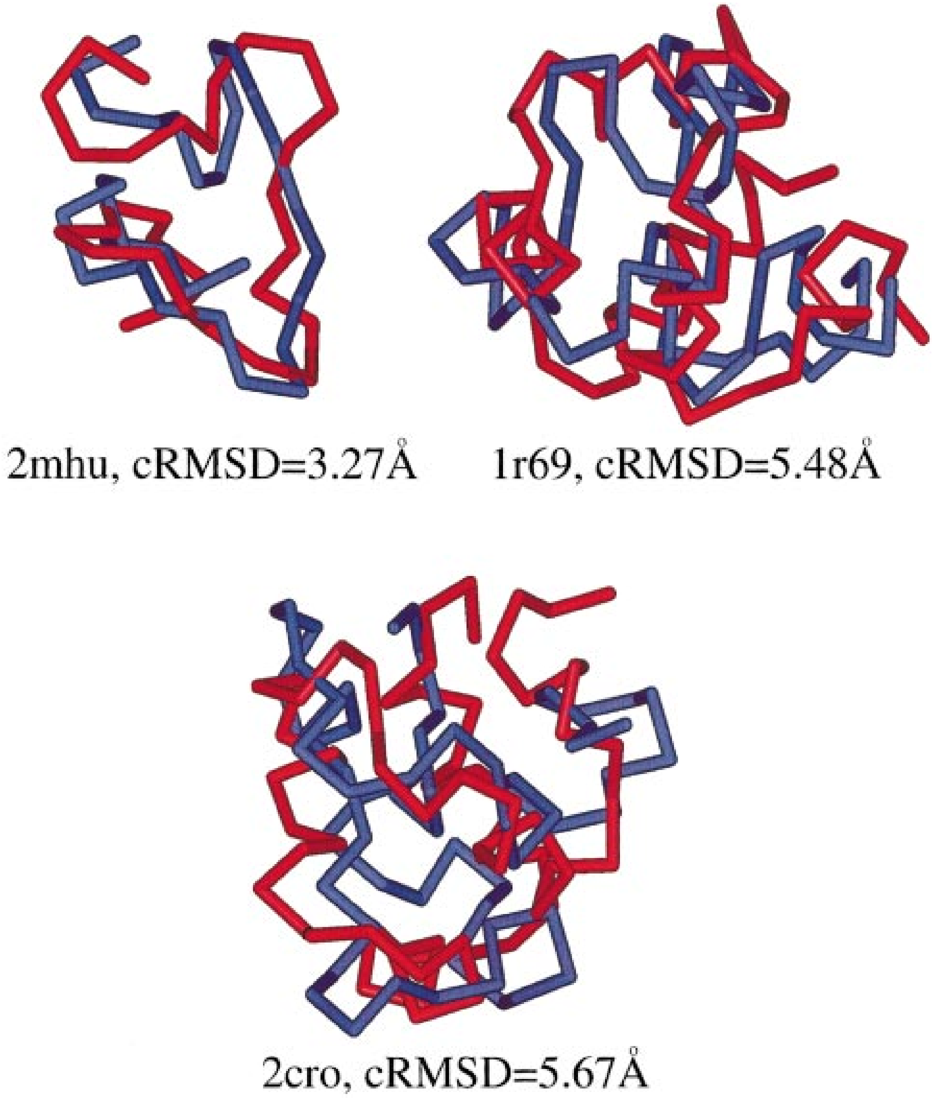

FIG. 4. (Color) Comparison of native C_α trace (red) and the corresponding structure generated by the chain growth algorithm (blue) with the lowest dRMSD or cRMSD. Superpositioning was done using the molecular graphics program Insight II. The cRMSD value was calculated using the optimal superposition method of Kabsch (Ref. 31). Results were obtained using 5 million configurations for protein 2mhu and 10 million each for proteins 1r69 and 2cro.

the parent, target, and Janus proteins; 5 million configurations were generated for each of the proteins. Since 1pgb (56 residues) and 1rop (63 residues) proteins do not have the same number of residues, the last seven residues at the unstructured C-terminal tail of 1rop are not considered in our simulations.

Figure 5(a) shows how well the best generated structures compare with the native 1pgb and 1rop proteins: the cRMSD for 1rop is 4.9 Å and for 1pgb it is 6.2 Å . Better accuracy is achieved for the all-α protein because it is topologically simpler than the mainly β protein 1pgb.

Since the Janus structure is not available, we next compare the configurations generated with this sequence to the parent and target structures. From the Janus configurational ensemble, we select two structures where one of them has the lowest dRMSD with respect to the parent protein and the other to the parent protein. The two pairs of superimposed structures are shown in Fig. 5b. The best Janus structure for the target protein has a low cRMSD of 5 Å , but the equivalent structure for the parent protein has a larger cRMSD of 7 Å . Thus, our approach can approximate the target structure using its native (1rop) and Janus sequences.

The algorithm's ability to generate accurate folds is a prerequisite for predicting protein structures. *De novo* predictions must be based on some selection criteria such as free energy. For completeness, we also compare the lowest energy structure of Janus in our ensemble to the parent and target proteins in Fig. 5(c). Disappointingly, the Janus structure has cRMSD values of 8.1 and 11.8 Å with respect to the parent β protein and target α protein, respectively. This failed prediction may be explained by the fact that the lowest energy conformation from our model is a compact structure whereas the 1rop structure is an elongated fold. A better assessment of a prediction algorithm is to test its predictive capability on a set of proteins. On this measure, our algorithm's accuracy of predicted folds generally varies between 6 to 8 Å for a set of α proteins.[34] Recent assessment of blind predictions (CASP3, critical assessment of structure predictions) from *ab initio* methods shows that for most target structures the cRMSD value is about 10 Å , although a few significantly better predictions are reported.[35] While progress in *ab initio* approaches has been made, all approaches are far from being sufficiently accurate to be useful for detailed functional studies (e.g., prediction of active sites).

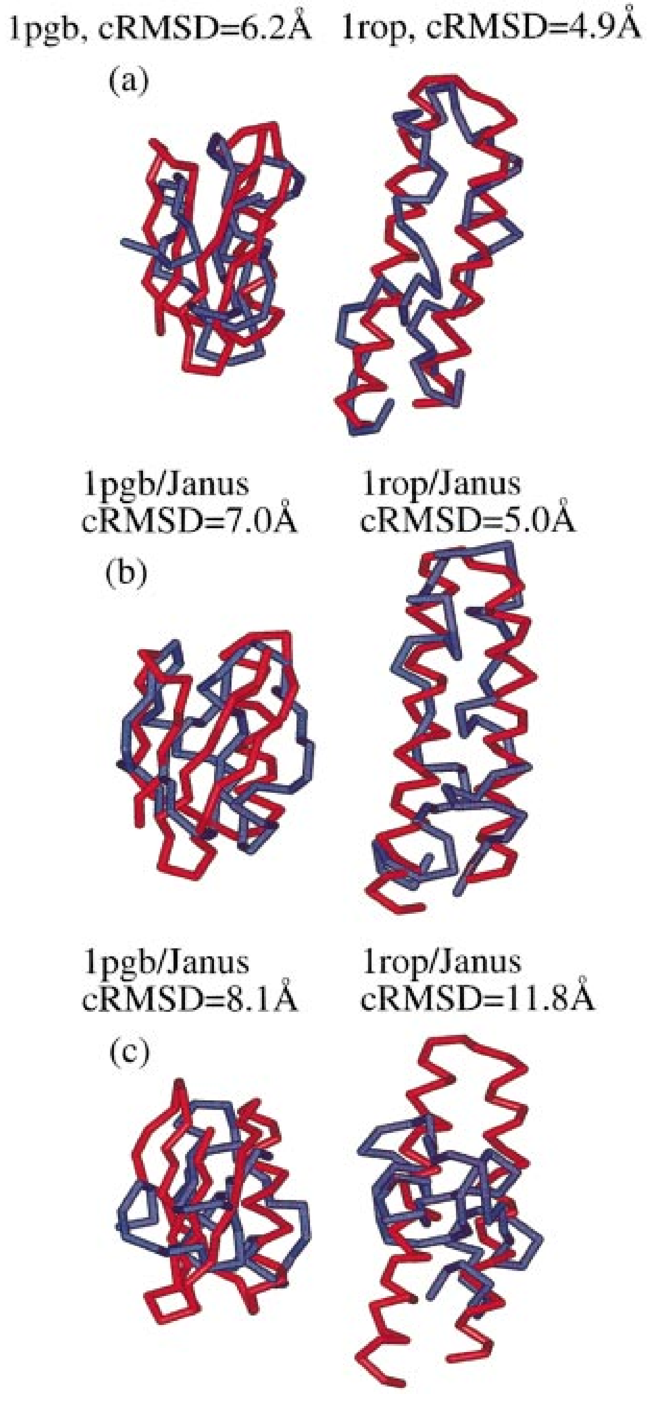

FIG. 5. (Color) Comparison of native C_α trace (red) and the corresponding structure generated by the chain growth algorithm (blue): (a) lowest dRMSD or cRMSD structures in configurational ensembles for 1pgb (left) and 1rop (right) proteins; (b) two structures from the Janus ensemble with the lowest dRMSD with respect to 1pgb and 1rop proteins; (c) the generated Janus structure with the lowest energy is superimposed with 1pgb and 1rop proteins. Note that the last seven unstructured residues of the C-terminal of 1rop are not shown, and the structures superimposed have 56 residues each. Five million configurations were produced for each of the protein sequences.

3. Correlations between RMSD and energy/weight

Next we assess the quality of configurational ensembles by displaying the correlation between RMSD and energy for the configurations. We also introduce the ensemble-averaged dRMSD and an energylike quantity based on the statistical weight W_Λ. Energy and statistical weight are different criteria for selecting favorable configurations whereas ensemble-averaged dRMSD is a global measure of ensemble's quality.

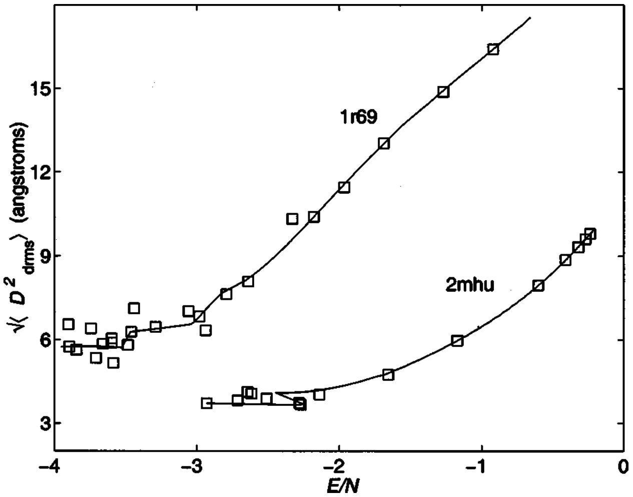

FIG. 6. Ensemble-averaged distance RMSD vs E/N, the excess internal energy per residue, for the two proteins 2mhu and 1r69. The square symbols denote simulation points, and the composite curves were computed using the analytic extrapolation formula Eq. (8) given four configurational ensembles at $T^*=1$, 1.25, 1.5, and 2.

We define the thermal or ensemble-averaged dRMSD value for the configurational ensemble at inverse temperature parameter β as follows:

$$\langle D^2_{\rm drms}\rangle_\beta = \sum_\Lambda D^2_{\rm drms} S_\Lambda(\beta,\beta') \Bigg/ \sum_\Lambda S_\Lambda(\beta,\beta'), \qquad (17)$$

where $S_\Lambda(\beta,\beta')$ is defined in Eq. (9). This average is a more meaningful measure of expected deviations produced by the model and algorithm than other measures based on specific configurations, especially for equilibrium configurations with considerable flexibility. It is also equivalent to the long-time average of dRMSD in equilibrium molecular dynamics simulations. The dependence of $\sqrt{\langle D^2_{\rm drms}\rangle}$ on internal energy $E(T)=\langle E_\Lambda\rangle_\beta$ for the proteins 2mhu and 1r69 is shown in Fig. 6. In this plot, we have calculated the mean energy per residue using both direct simulations (points) and an extrapolation technique whose results are shown as composite curves. The composite curves from the extrapolation procedure were obtained by using formula [Eq. (8)] with four configurational ensembles generated at $T^*=1$, 1.25, 1.5, and 2. Figure 6 shows that the ensemble-averaged dRMSD value is a monotonic decreasing function of the internal energy. It emphasizes that a low average dRMSD corresponds to a low average energy, although below a certain threshold energy the average dRMSD does not seem to improve. For the configurations with the lowest energies (i.e., near room temperature) their dRMSD values vary between 4 Å and 6 Å . These values are about 2 Å larger than the best values in the ensembles and slightly lower than the dRMSD values at which the peaks of the dRMSD distributions are found (Fig. 3).

The average RMSD is a statistical measure of the RMSD for an ensemble of configurations. Often, it is desirable to compare specific, favorable configurations to the native structure; this requires a way of ranking the configurations.

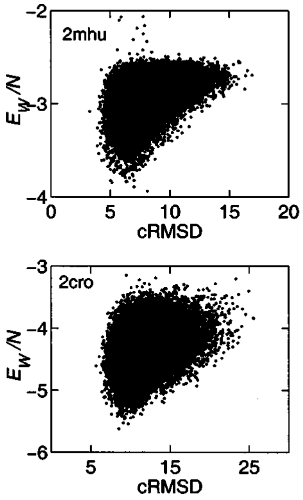

FIG. 7. Scatter plots of the weight-based energy per residue, E_W/N [Eq. (18)], vs cRMSD for three proteins. Energy is expressed in units of k_BT_0 where T_0 is the room temperature and cRMSD is in Å . Each plot displays 50 000 randomly chosen configurations.

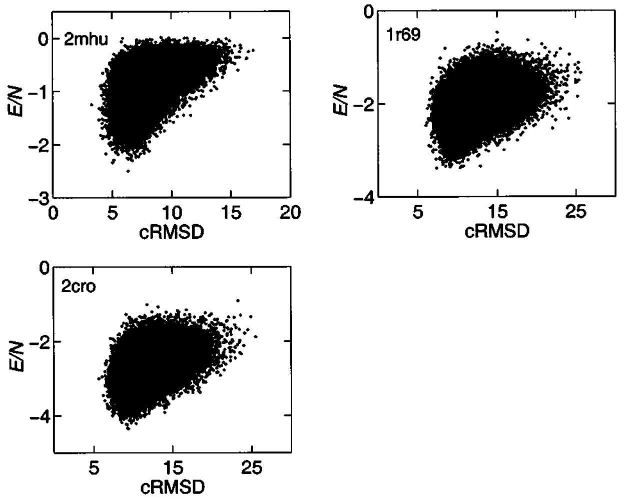

FIG. 8. Scatter plot as in Fig. 7, but for internal energy per residue, E/N, vs cRMSD.

Most ranking of proximity to the native state is based on the internal energy by assuming that the entropic contribution is small. Here we employ both energy and weight-based selection criteria to display their correlations with RMSD.

The new weight-based selection criterion is rationalized as follows. Recall that the excess free energy $F(\beta)\propto -\ln\Sigma_{\{\Lambda\}}W_{\Lambda}(\beta)$. This function cannot be used to rank configurations because a free-energy like function that is configuration specific is required. We achieve this by using the following weight-based energy for configuration Λ:

$$E_W(\Lambda)=-k_BT\ln W_{\Lambda}\,. \tag{18}$$

Here, configurations with large weights correspond to low energies. Recall that statistical weight W_{Λ} is related to chosen growth directions which are in turn determined by the Boltzmann factors in the transition probability. Moreover, the weight-based energy is correlated with both the internal and the free energies. In fact, our simulation results show that $-k_BT\ln W_{\Lambda}$ is correlated with internal energy E_{Λ} at room temperature. It is also apparent that $E_W(\Lambda)$ is obtained from the free energy when a configuration in the ensemble has a dominant weight. Thus, $E_W(\Lambda)$ offers a different selection criterion but is still related to the internal and free energies.

Folded nativelike conformations are expected to have much lower energies than unfolded conformations. Therefore, there should be a good correlation between the configuration weight-based energy and cRMSD value. Scatter plots of E_W/k_BT versus cRMSD for the three proteins are shown in Fig. 7 using the cRMSD corresponding to optimal superposition of the two structures. The shapes of the scatter plots in Fig. 7 indicate a general correlation between low cRMSD value and energy. However, this correlation is not strong since configurations with high E_W values can also have lower relative cRMSD values. In Fig. 8, similar scatter plots are shown using the internal energy E as selection criterion instead of the free-energy like E_W. The general trend is similar to those in Fig. 7. Others have also found comparable or weaker degree of energy/cRMSD correlations when statistical potentials are used to compute the protein energy or used as selection criteria.[36] Stronger correlations between cRMSD and energy are possible with all-atom potentials.[37]

B. Thermodynamic analysis of temperature-induced transitions

The native states of proteins are special points in the thermodynamic phase diagram. Experimental studies of protein folding/unfolding provide essential information about the behavior of proteins under a wide range of temperature and solvent conditions.[29,30] The validity of protein interaction potentials and the algorithms designed to find the global free energy minimum are more fully tested when computed properties of proteins in various regions of the thermodynamic phase diagram are assessed with respect to experimental data. Below we report on the response of thermodynamic functions to variation of temperature. Since the solvent degrees of freedom are not included explicitly, variation of thermodynamic properties with respect to solvent conditions cannot be adequately studied here.

1. Statistical fluctuations in thermodynamic quantities

We begin with a consideration of statistical errors in the computation of thermodynamic quantities. Statistical errors in thermodynamic quantities are estimated by performing different runs for the same protein at a given temperature, and repeating this process for various temperatures of interest. The number of configurations sampled is 5×10^6 for 2mhu and 10^7 for the larger proteins 1r69 and 2cro. Computational time for a single run for protein 2mhu is about 10 hours on SGI Origin2000 and about four times longer for the other two proteins (double the size of 2mhu). Typically, over 90% of the trial configurations are successfully grown (less than 10% are discarded because of dead-end configurations).

Table II summarizes the standard deviations for various calculated thermodynamic quantities at the reduced temperature $T^*=1.5$, which are expressed as percentages of the av-

TABLE II. Standard deviations of excess internal energy (E), free energy (F), specific heat (C), and entropy (S) per residue, and mean radius of gyration (R_G) and ensemble-averaged dRMSD ($\sqrt{\langle D^2_{\mathrm{drms}}\rangle}$) at reduced temperature of $T^*=1.5$ as obtained from 20 runs each for two proteins (ensembles of 5 million configurations for 2mhu and 10 million configurations for 1r69). The standard deviations are expressed as a percentage of the mean value and the energy is in units of k_BT_0 where T_0 is the room temperature.

T^*	E/N	F/N	C/k_BN	S/k_BN	R_G (Å)	$\sqrt{\langle D^2_{\mathrm{drms}}\rangle}$ (Å)
2mhu						
1.5	3.0	21.7	25.4	2.5	0.9	2.9
1r69						
1.5	3.5	3.2	73.6	4.3	5.6	12.8

erage values; each value was obtained from 20 different runs. We have chosen the temperature to lie in the transition region of thermal curves where fluctuations are most significant. As shown in Table II, all quantities except the specific heat capacity have small to moderate standard deviations. The heat capacity, reflecting energy fluctuations or the second energy moment, is subject to larger numerical uncertainties than, for example, the internal energy.

2. Thermodynamic profiles

In the chain growth approach thermodynamic functions can be calculated using both the direct simulation approach and the theoretical extrapolation formula (8). The direct simulation approach requires that many points be calculated at different temperatures whereas the theoretical formula of Eq. (8) can rapidly estimate the shape of the curve based on a few configurational ensembles generated. The extrapolation formula is expected to be most effective in moderate and high temperature regions. Because the low temperature region is dominated by a few specific configurations, more configurational ensembles may be needed to obtain accurate thermodynamic curves. To reproduce a thermodynamic curve, we generate ensembles at $T_1, T_2, \ldots, T_n$ where n is typically small. A composite curve is obtained by extrapolating known thermal average values $A(T_1), A(T_2), \ldots, A(T_n)$ using Eq. (8). We compare direct simulation results with thermodynamic curves produced in this way.

The temperature dependence of the excess free energy, internal energy, and entropy [Eqs. (12)–(14), respectively] is shown in Fig. 9 for the three proteins 2mhu, 1r69, and 2cro; the corresponding results for the radius of gyration and ensemble-averaged dRMSD are shown in Fig. 10. For each protein, the actual simulation values are shown as points and the curves are derived from Eq. (8). The extrapolated curves are based on four ensembles at reduced $T^*=1$, 1.25, 1.5, and 2. We generally note a good fit between computations and theory, for moderate and high temperatures. The best agreement is in the free energy curves in Fig. 9. The composite curves are discontinuous in the lower temperature region, and there is evidence that simulated points behave similarly. Better agreement can be achieved when the composite curves are constructed based on more configurational ensembles, at the expense of computational cost. For comparison, thermodynamic functions of simple protein models on cubic lattices calculated by a chain growth algorithm can be obtained with greater efficiency than with our more realistic protein model.[10] Here, the statistical fluctuations are small and the extrapolation technique works well over a wide temperature range.

The excess energy and entropy functions in Fig. 9 show marked increases when the temperature exceeds about 300 K. We see an initial increase in the temperature range 300–500 K and further increases at higher temperatures as the proteins unfold. The small protein 2mhu tends to show an abrupt transition from native to unfolded state. The transition curves of the larger proteins 1r69 and 2cro are not smooth and the transition is spread over a broader temperature range. The larger proteins tend to reach their high temperature satu-

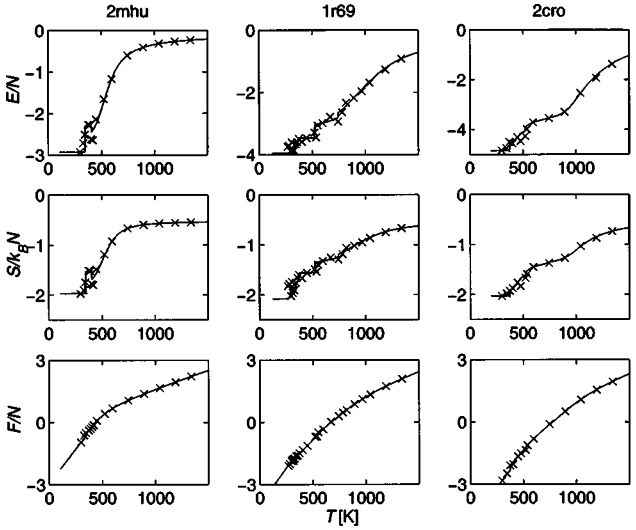

FIG. 9. Temperature dependence of excess internal energy (E/N), entropy (S/k_BN), and free energy (F/N) per residue for three proteins. The symbols are simulation values and the composite curves were obtained via analytic extrapolation formula Eq. (8) given four configurational ensembles at reduced temperatures $T^*=1$, 1.25, 1.5, and 2. The energies are expressed in units of k_BT_0 where T_0 is room temperature.

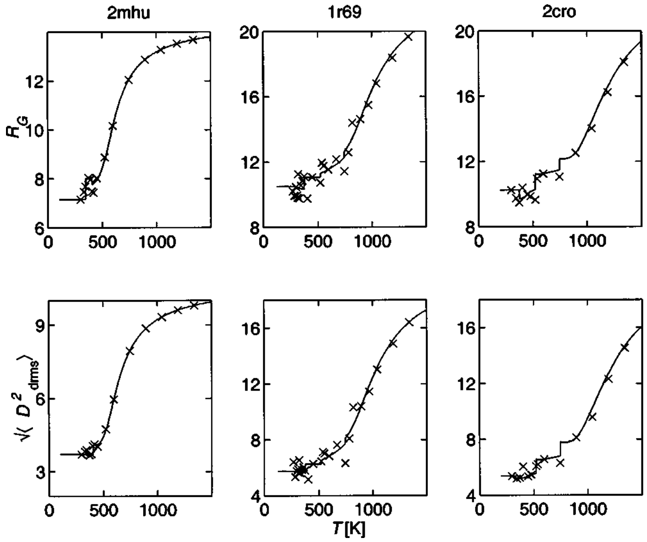

FIG. 10. Temperature-dependent studies as in Fig. 9, but for the radius of gyration (R_G) and ensemble- averaged dRMSD, in units of Å .

ration value at higher temperatures. This difference is partly due to the higher internal energies (per residue) for the larger proteins; see Table I. For temperatures less than 500 K, the radius of gyration and ensemble-averaged dRMSD in Fig. 10 only change moderately. Large increases for these global properties occur when the temperature approaches two to three times the native or room temperature. These temperatures correspond to about 2/3 of the internal energy per residue at $T^*=1$. At these temperatures, all degrees of freedom in the proteins are essentially thermalized, resulting in dramatic increases in the radius of gyration and thermodynamic curves (Figs. 9 and 10). Proteins are unfolded under these conditions.

3. Relation to other theoretical studies

Protein denaturation has been rationalized using both the two-state transition[29] and transition via intermediate states.[30,38] Recent entropy-sampling Monte Carlo studies by Hao and Scheraga[3–6] on thermal transitions of model proteins support the view that the transitions are essentially of the two-state character. The entropy sampling method is similar to the multicanonical sampling approach which has been applied to proteins.[2] In this view of protein denaturation transition, the protein is either in the native state or the unfolded states with few stable intermediate states. The free energy of *states*, as opposed to thermodynamic free energy, is found to have a double-well shape. The free energy of states cannot be calculated using the chain growth method because the density of states is not determined. Calculated free energy curves in Fig. 9 for all three proteins 2mhu, 1r69, and 2cro are smoothly varying functions of temperature, which show lack of abrupt changes as seen in energy, entropy, and radius of gyration transition curves.

Protein denaturations associated with heat and acid denaturants are also interpreted in terms of partial unfolding or formation of intermediate molten globule state.[30] Complete unfolding occurs only in strong concentrations of urea. This type of transition is akin to melting of crystals to the disordered liquid state where the density of molecular packing is not greatly changed. Such a theoretical model has been proposed for protein denaturation.[38] As remarked before, below 500 K the values of the radius of gyration of proteins in Fig. 10 increase only moderately or by about 10% even though the energy and entropy in Fig. 9 show significant changes. Our results indicate that below 500 K the computed proteins are far from being fully unfolded. This discussion emphasizes that thermodynamic functions and mean radius of gyration do not necessarily increase at the same rate.

Our calculated results are more consistent with the picture that proteins denature via partial unfolding, although we caution that our calculations are limited to only three proteins using a low-resolution model. To support either view of protein denaturation, more protein cases must be examined and the level of accuracy of the model must be enhanced.

V. SUMMARY AND CONCLUSIONS

We have examined the feasibility of using the chain growth algorithm for computing the configurational properties of proteins. The properties analyzed include structural features, correlations between energy and RMSD for configurations, and thermal transition profiles. Our protein model is more realistic than those in previous studies using chain growth algorithms[10,12] because it includes all 20 residue types, a physical range of pseudobond angles, and the simulation was performed on a refined cubic lattice.

We show that our model/algorithm combination is capable of generating configurations that are reasonably close to the native structures. The ensemble-averaged deviations from the native structures are only about 2 Å poorer than the best structures obtained, although the dRMSD distributions of the configurational ensembles are fairly broad. In addition, we presented an interesting application of the chain growth algorithm to the parent and target sequences, and the transformed sequence Janus associated with the solution to the Paracelsus challenge. Our algorithm reproduced the target protein structure with reasonable accuracy based on the best structures in the configurational ensembles generated for target and Janus sequences. We also introduced weight-based energy to evaluate energy-RMSD correlations for configurations; we found that present configurational ensembles yield only a moderate level of correlation. Collectively, these results appear to be satisfactory in view of our relatively simple C_α protein model which does not incorporate any specific secondary biases or tertiary restraints. Further improvements should be possible with more accurate representations of the protein chain and newer chain growth algorithms that are currently being developed.[12]

The thermodynamic curves of proteins complement our results on structural features and analysis of configurational ensembles. We showed that the chain growth method can reproduce the thermal transition curves with only a few simulation runs. This technical advantage should prove useful in studies of protein transitions. Computed protein thermodynamic functions show that proteins undergo a transition to the unfolded state above the room temperature. Temperature dependence of the radius of gyration and ensemble-averaged dRMSD indicates that the rate of change may be different from that of thermodynamic functions. We discussed our results in the context of current views on protein transitions. However, computed specific heat capacity was subject to significant statistical fluctuations which prevented meaningful comparison to experimental results.

The chain growth method of generating configurations is conceptually distinct from the Metropolis algorithm. Ensembles of statistically independent configurations provide efficient computations of thermodynamic functions for the test cases presented. Furthermore, current developments in the chain growth algorithm may enhance its computational efficiency for realistic protein models which is one of the goals of protein structure prediction research. Our structural and thermodynamic results demonstrate that the algorithm is a viable alternative to other Monte Carlo algorithms that are currently being employed in theoretical studies of protein structure prediction and thermodynamics. In a forthcoming paper, we use the chain growth algorithm to evaluate the ability of different energy functions to discern nativelike structures for a set of proteins.[34]

ACKNOWLEDGMENTS

This work was supported by NIH Grant No. GM55164 and NSF Grants Nos. BIR-94-23827EQ and ASC-9704681 to T. Schlick. T. Schlick is an investigator of the Howard Hughes Medical Institute. A. Tropsha gratefully acknowledges receiving support from NIH Research Resource Grant No. RR08102.

APPENDIX: FORMULATIONS OF THE CHAIN GROWTH ALGORITHM

Here we present formulations of the chain growth algorithm for athermal and temperature-dependent chains based on Rosenbluth and Rosenbluth,[7] Meirovitch,[9] and Grassberger.[12] The concepts of transition probability, statistical weight, and importance sampling are described. An overview and summary of the algorithm are given in Secs. III A and III B.

1. Transition probabilities and statistical weights

For a chain molecule with $N+1$ interaction sites, the chain growth process is governed by the transition probability P_i at step i that specifies the probability of selecting a given direction for the next link after $i-1$ links have been placed. For self-avoiding chains, the new direction must not overlap with any of the previous sites. Thus the transition probability P_i depends on the position vectors $\mathbf{R}_1,\ldots,\mathbf{R}_i$ of sites $1,\ldots,i$. We denote by $\mathbf{v}_{k_i}$ the lattice vector for the chosen direction k_i at step i. The transition probability P_i for athermal chains is given by

$$P_i(\mathbf{R}_i+\mathbf{v}_{k_i}|\mathbf{R}_1,\ldots,\mathbf{R}_i)=1/C_i, \tag{A1}$$

where C_i is the number of vacant sites at step i and $P_i(\mathbf{R}_i+\mathbf{v}_{k_i})$ is normalized: $\Sigma_{k_i=1}^{C_i}P_i(\mathbf{R}_i+\mathbf{v}_{k_i})=1$. If n_c is the coordination number of a lattice, the maximum value of C_i is n_c-1 for self-avoiding chains and n_c for ideal chains with allowed overlapping sites. For example, $n_c=6$ for a simple cubic lattice and 24 for the refined cubic lattice (311). For a chain with N links, the probability of generating the full configuration of the chain $\Lambda=\Lambda(\mathbf{R}_1,\ldots,\mathbf{R}_{N+1})$ is then

$$P_\Lambda=\prod_{i=1}^{N} P_i(\mathbf{R}_i+\mathbf{v}_{k_i}|\mathbf{R}_1,\ldots,\mathbf{R}_i)=1/W_\Lambda, \tag{A2}$$

where the weight $W_\Lambda=\Pi_{i=1}^{N}C_i$. Configurations of athermal chains are mostly expanded states, and their average dimension (e.g., radius of gyration) obeys well-known scaling laws.

To derive the analogous expressions for temperature-dependent chains, the above transition probability P_i is modified to allow sampling of both open and compact configurations when $\beta\neq 0$ and the potential has a finite range. Meirovitch[8,11] proposed a transition probability where a uniform weighting of Eq. (A1) is modified by a Boltzmann-weighted energy factor as follows:

$$P_i(\mathbf{R}_i+\mathbf{v}_{k_i}|\mathbf{R}_1,\ldots,\mathbf{R}_i;\beta)$$
$$=\exp[-\beta u_i(\mathbf{R}_i+\mathbf{v}_{k_i})]\Big/\sum_{k_i=1}^{C_i}\exp[-\beta u_i(\mathbf{R}_i+\mathbf{v}_{k_i})], \tag{A3}$$

where the relevant incremental, nonbonded potential energy at the ith step is

$$u_i(\mathbf{R}_i+\mathbf{v}_{k_i})=\sum_{j=1}^{i-1}u_{ij}(R_{ij}). \tag{A4}$$

In practice, we ensure that the probability P_i given in Eq. (A3) is associated with the selected direction k_i by considering several growth directions k_i' and associated energies $u_i(\mathbf{R}_i+\mathbf{v}_{k_i'})$ through comparison to a uniform random variate x in [0,1] as follows. We define

$$J_{k_i}=\sum_{k_i'=1}^{k_i}P_i(\mathbf{R}_i+\mathbf{v}_{k_i'}) \tag{A5}$$

and increase k_i from 1 up to the smallest integer k_i that yields the sum $J_{k_i}>x$. The chain growth process is terminated when no vacant sites are available, i.e., $C_i=0$.

The probability of generating a configuration at temperature β is now given by

$$P_\Lambda(\beta)=\prod_{i=1}^{N}P_i(\mathbf{R}_i+\mathbf{v}_{k_i}|\mathbf{R}_1,\ldots,\mathbf{R}_i;\beta)$$
$$=\exp(-\beta E_\Lambda)/W_\Lambda(\beta), \tag{A6}$$

where the statistical weight

$$W_\Lambda(\beta)=\prod_{i=1}^{N}\left\{\sum_{k_i=1}^{C_i}\exp[-\beta u_i(\mathbf{R}_i+\mathbf{v}_{k_i})]\right\}. \tag{A7}$$

In the athermal limit, $\beta=0$ or $E_\Lambda=0$, we recover the Rosenbluth transition probability [Eq. (A1)]. In fact, the configurational probabilities of Rosenbluth [Eq. A2)] and Meirovitch [Eq. A6)] satisfy the relation

$$P_\Lambda(\beta)W_\Lambda(\beta)=\exp(-\beta E_\Lambda). \tag{A8}$$

Thus other forms of transition probabilities and associated weights are possible provided this relation is satisfied. Indeed, Grassberger and co-workers have used this relation to design other forms for weights and transition probabilities in the context of polymers and simple protein models.[12]

2. Ensemble averaging by importance sampling

Once configurations are generated as above, an ensemble is used to evaluate a thermodynamic property A in canonical ensemble by a modification of the standard average

$$\langle A\rangle_\beta=\sum_{\{\Lambda\}}A_\Lambda\exp(-\beta E_\Lambda)\Big/\sum_{\{\Lambda\}}\exp(-\beta E_\Lambda), \tag{A9}$$

where E_Λ is the total potential energy for configuration Λ, and A_Λ is the value of property A for Λ.

A modification of Eq. (A9) is needed to remove the bias inherent in the growth process since the configuration probability P_Λ depends on the sequence of selected growth directions. This can be done by using an importance sampling procedure which assigns a counter-weight P_Λ to each A_Λ term. The thermal average in Eq. (A9) becomes

$$\begin{aligned}\langle A\rangle_\beta &= \sum_{\{\Lambda\}} A_\Lambda P_\Lambda^{-1}(\beta)\exp(-\beta E_\Lambda) \Bigg/ \sum_{\{\Lambda\}} P_\Lambda^{-1}(\beta) \\ &\quad \times \exp(-\beta E_\Lambda) \\ &= \sum_{\{\Lambda\}} A_\Lambda W_\Lambda(\beta) \Bigg/ \sum_{\{\Lambda\}} W_\Lambda(\beta).\end{aligned} \tag{A10}$$

The configurations $\{\Lambda\}$ above now refer to those generated by the chain growth method. In this average, all successfully grown configurations are counted, each with a statistical weight W_Λ [as defined in Eq. (A7)]. This procedure generalizes the Rosenbluth weighting for statistical averages.[7] Naturally, to enhance the accuracy of the estimate of thermal averages, the configurational sample must be sufficiently large so that configurations with large weights belong to the ensemble. Thus, the importance sampling procedure will generally be ineffective for small samples. An alternative procedure for ensuring correct weighting of chain growth configurations for thermal averages is described in Refs. 11 and 39.

[1] J. Skolnick and A. Kolinski, Adv. Chem. Phys. **105**, 203 (1999).

[2] Y. Okamoto, Recent Res. Devel. Pure Appl. Chem. **2**, 1 (1998);U. Hansmann and Y. Okamoto, Curr. Opin. Struct. Biol. **9**, 177 (1999); J. Chem. Phys. **110**, 1267 (1999).

[3] M. H. Hao and H. A. Scheraga, J. Phys. Chem. **98**, 4940 (1994).

[4] M. H. Hao and H. A. Scheraga, J. Phys. Chem. **98**, 9882 (1994).

[5] M. H. Hao and H. A. Scheraga, J. Chem. Phys. **102**, 1334 (1995).

[6] M. H. Hao and H. A. Scheraga, J. Mol. Biol. **277**, 973 (1999).

[7] M. N. Rosenbluth and A. V. Rosenbluth, J. Chem. Phys. **23**, 356 (1955).

[8] H. Meirovitch, Macromolecules **16**, 1628 (1983).

[9] H. Meirovitch, J. Chem. Phys. **89**, 2514 (1988).

[10] I. Szleifer, E. M. O'Toole, and A. Z. Panagiotopoulos, J. Chem. Phys. **97**, 6802 (1992); E. M. O'Toole and A. Z. Panagiotopoulos, *ibid.* **97**, 8644 (1992).

[11] H. Meirovitch, M. Vasquez, and H. A. Scheraga, Biopolymers **27**, 1189 (1988).

[12] P. Grassberger, Phys. Rev. E **56**, 3682 (1997); U. Bastolla, H. Frauenkron, E. Gerstner, P. Grassberger, and W. Nadler, Proteins: Struct., Funct., Genet. **32**, 52 (1998).

[13] J. I. Siepmann and D. Frenkel, Mol. Phys. **75**, 59 (1992).

[14] A. R. Ortiz, A. Kolinski, and J. Skolnick, J. Mol. Biol. **277**, 419 (1998).

[15] J. Skolnick and A. Kolinski, J. Mol. Biol. **221**, 449 (1991).

[16] S. Miyazawa and R. Jernigan, Macromolecules **18**, 534 (1985).

[17] S. Miyazawa and R. Jernigan, J. Mol. Biol. **256**, 623 (1996).

[18] S. Dalal, S. Balasubramaniam, and L. Regan, Nature Struc. Biol. **4**, 548 (1997).

[19] D. G. Covel, Proteins: Struct., Funct., Genet. **14**, 409 (1992).

[20] D. G. Covel, J. Mol. Biol. **235**, 1032 (1994).

[21] D. Hinds and M. Levitt, Proc. Natl. Acad. Sci. U.S.A. **89**, 2536 (1992).

[22] L. Toma and S. Toma, Protein Sci. **8**, 196 (1999).

[23] A. Kolinski, W. Galazka, and J. Skolnick, J. Chem. Phys. **108**, 2608 (1998).

[24] A. Kolinski and J. Skolnick, *Lattice Models of Protein Folding, Dynamics, and Thermodynamics* (Landes, Austin, 1996).

[25] R. L. Jernigan and I. Bahar, Curr. Opin. Struct. Biol. **6**, 195 (1996).

[26] M. Sippl, J. Mol. Biol. **213**, 859 (1990).

[27] J. Skolnick, L. Jaroszewski, A. Kolinski, and A. Godzik, Protein Sci. **6**, 676 (1997).

[28] M. R. Betancourt and D. Thirumalai, Protein Sci. **9**, 361 (1999).

[29] P. L. Privalov, Adv. Protein Chem. **33**, 167 (1979).

[30] O. B. Ptitsyn, Adv. Protein Chem. **47**, 83 (1995).

[31] W. Kabsch, Acta Crystallogr., Sect. A: Cryst. Phys., Diffr., Theor. Gen. Crystallogr. **32**, 922 (1976); **34**, 827 (1978).

[32] We thank M. Mezei for the subroutine for calculating cRMSD described in M. Mezei, Protein Engineering **7**, 331 (1994).

[33] G. D. Rose and T. P. Creamer, Proteins: Struct., Funct., Genet. **19**, 1 (1994).

[34] H. H. Gan, A. Tropsha, and T. Schlick (unpublished).

[35] C. A. Orengo, J. E. Bray, T. Hubbard, L. LoConte, and I. Sillitoe, Proteins: Struct., Funct., Genet. **3**, 149 (1999).

[36] B. H. Park, E. S. Huang, and M. Levitt, J. Mol. Biol. **266**, 831 (1997).

[37] T. Larazidis and M. Karplus, J. Mol. Biol. **288**, 477 (1999).

[38] E. I. Shakhnovich and A. V. Finkelstein, Biopolymers **28**, 1667 (1989).

[39] K. E. Schmidt, Phys. Rev. Lett. **51**, 2175 (1983).