Multisequence algorithm for coarse-grained biomolecular simulations: exploring the sequence-structure relationship of proteins

A. Aina[1] and S. Wallin[1, a)]

Memorial University of Newfoundland, Department of Physics and Physical Oceanography, A1B 3X7 St John's, NL, Canada

(Dated: 1 August 2017)

We consider a generalized-ensemble algorithm for coarse-grained simulations of biomolecules which allows the thermodynamic behavior of two or more sequences to be determined in a single multisequence run. By carrying out a random walk in sequence space, the method also enhances conformational sampling. Escape from local energy minima is accelerated by visiting sequences for which the minima are more shallow or absent. We test the method on an intermediate-resolution coarse-grained model for protein folding with 3 amino acid types and explore the potential for large-scale coverage of sequence space by applying the method to sets of more than 1,000 sequences. The resulting thermodynamic data is used to analyze the structures and stability properties of sequences covering the space between folds with different secondary structures.

PACS numbers: 87.14.E; 87.15.A; 05.10.Ln

Keywords: Monte Carlo, generalized ensembles, protein folding, protein fold switching

[a)] Electronic mail: swallin@mun.ca

I. INTRODUCTION

Recent years have seen important advances in biomolecular simulation methods, including improvements to standard molecular dynamics force fields,[1] the advent of several alternative atomistic simulation approaches,[2–5] and new techniques for conformational sampling.[6] Together with the ever-increasing availability of computational resources, these advances have triggered a few major efforts[7–11] to characterize the dynamics of biomolecular systems of various sizes, e.g., a small native protein on the millisecond scale[10] and a comprehensive model cytoplasm on the nanosecond scale.[11] While encouraging and insightful, these large-scale simulations have also highlighted that severe tradeoffs in size and time scales will likely persist for the foreseeable future.

One way to expand the range of biomolecular simulations is to turn to coarse-grained (CG) models, where the basic aim is to simplify the description of physical interactions while retaining the essential physics of the system under study.[12] Ingolfsson *et al.* list 4 main factors that make CG models computationally fast: reduction in the number of degrees of freedom, faster simulation dynamics, emphasis on short-range interactions and the ability of using larger integration time steps.[13] To this list can be added that a CG representation of either the interaction potential or the molecular geometry often opens up for alternative sampling schemes beyond traditional molecular dynamics approaches, which can further speed up conformational sampling. Examples of such sampling schemes include activation-relaxation kinetics,[14] discrete molecular dynamics[15] and various Monte Carlo (MC)-based techniques such as cluster moves.[16]

The challenges of achieving representative conformational sampling of individual biomolecular systems notwithstanding, many biologically motivated problems naturally call for the investigation and comparison of molecular variants, e.g., determining the molecular mechanisms of specificity in protein-protein[17,18] or protein-nucleotide[19] interactions, or the role of mutations in molecular disease processes.[20] Another example is protein folding, where unique insight has been achieved by comparing sequences within and between protein families.[21,22] In a situation with extremely rapid growth of sequence information,[23] it is of interest to explore ways to efficiently sample multiple sequences in biomolecular simulations.

To this end, we consider in this work an MC-based algorithm that can calculate the thermodynamics of multiple sequences in a single run and apply it to a coarse-grained model for protein folding.[24] This multisequence (MS) method was originally developed in the context of homo- and heteropolymer simulations[25] and was later adapted for the characterization of peptide-protein binding specificity.[26,27] To our knowledge, it has not been previously tested in realistic protein folding simulations. The MS algorithm carries out a simulation in a generalized ensemble that performs a random walk in sequence space. Hence, there are two main types of updates: conformational updates $r \to r'$ and sequence updates $s \to s'$. This strategy is straightforward when r and s are "perpendicular" coordinates, as illustrated in Fig. 1, such that the potential energy of the model can be written in terms of two independent variables, $E(s, r)$.

As a test application of the MS algorithm, we selected the phenomenon of protein fold switching which recently was demonstrated in a handful of natural and engineered proteins. These special proteins exhibit a unique ability to reversibly switch between entirely different folds, with accompanying changes in secondary structure, hydrophobic core packing and overall shape.[28] The fold switching transitions found in natural proteins typically play a functional role. For example, rare transitions to an alternative fold in the protein KaiB provide a crucial time delay mechanism in the circadian clock cycle of cyanobacteria.[29] Fold switching can occur either spontaneously[30] or be triggered by various signals including changes to solution conditions,[31] subdomain detachment,[32] ligand binding[33] and point mutations.[34] Computational studies, using coarse-grained[35–38] or atomistic[39–41] models, have attempted to explain how proteins can exhibit multiple folding funnels and how they are altered in response to binding events or changes in sequence.

The discovery that proteins can be driven to switch folds through an accumulation of point mutations, in particular, holds implications for protein evolution as it suggests a simple mechanism of fold evolution.[42] Alexander *et al.* demonstrated that the similarly sized but structurally distinct A (3α) and B ($4\beta + \alpha$) domains of protein G could, after extensive mutations leaving their respective folds undisturbed, be triggered to switch folds by applying one additional mutation, Y45L, located at the edge of the hydrophobic core in the B domain.[34] This remarkable discovery suggests the possibility that the two domains might be evolutionary related despite a lack of detectable similarity in either sequence or structure in wild-type protein G,[43] although this has yet to be proven. Moreover, it is unclear how common such fold-to-fold transitions are and how they might occur in evolutionary processes.[44] In previous work[35,36] we showed that mutational paths with abrupt fold switching exist

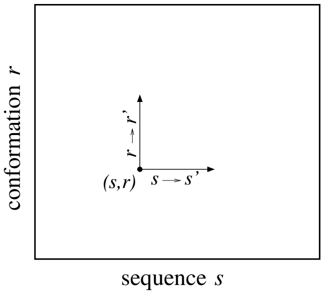

FIG. 1. The two types of Monte Carlo updates in the multisequence Monte Carlo algorithm.

between two other pairs of smaller protein folds within the framework of our CG model.[24]

In demonstrating mutation-induced fold switching in our model we characterized the folding of a set of 144 different model sequences with 16 amino acids. This set (denoted here $\mathrm{S16}_{144}$) was constructed to sparsely span the sequence space between two ideally designed sequences, A1 and N1, folding into an α-helix and a β-hairpin, respectively, as shown in Fig. 2. Here we use the set $\mathrm{S16}_{144}$ to validate the MS method and compare its computational efficiency to a standard generalized-ensemble method.[45,46] We thereafter greatly enlarge $\mathrm{S16}_{144}$ to a set with 1,024 sequences as well as another set of the same size spanning two 35-amino acid sequences, A2 and TN, that fold into two-helical bundle and mixed α-β structures, respectively (see Fig. 2). Besides demonstrating that the MS method can be applied to large numbers of sequences, the results allow us to carry out a more systematic analysis of the biophysical properties of sequences along mutational pathways connecting these two pairs of basic folds than has been previously possible.

II. THEORY

A. Generalized-ensemble algorithms and simulated tempering

Conventional Metropolis Monte Carlo simulations of the canonical distribution is problematic at low temperatures for many physical systems because simulations tend to become trapped in local energy minima and hamper representative sampling of configurational space. The basic idea of generalized-ensemble algorithms is to alleviate this trapping problem by sampling states using a non-Boltzmann weight factor and/or expand the state space with additional dynamical parameters[47] (for a recent historical account see Ref. 48). Generalized-ensemble methods that have been frequently used for biomolecular simulations include simulated tempering (ST),[45,46] replica exchange[49] or parallel tempering,[50] and metadynamics.[51]

ST is a direct extension to the Metropolis algorithm in which the temperature T becomes a dynamic parameter. In this way, frequent visits to high T allow simulations to readily escape from local energy minima. The algorithm thus simulates the joint probability distribution

$$P(m, r) = \frac{1}{\hat{Z}} e^{-\beta_m E(r) + g_m} \,, \tag{1}$$

where $\beta_m = 1/k_{\mathrm{B}} T_m$, $\{T_m\}_{m=1}^{\mathrm{M}}$ a set of temperatures and k_{B} is Boltzmann's constant. The normalization constant in Eq. 1 is

$$\hat{Z} = \sum_r \sum_{m=1}^{\mathrm{M}} e^{-\beta_m E(r) + g_m} \,, \tag{2}$$

where the first sum is over all conformations r. The simulation parameters g_m control the marginal probability distribution

$$P(m) = \frac{1}{\hat{Z}} \sum_r e^{-\beta_m E(r) + g_m} \,, \tag{3}$$

and must therefore be carefully chosen. A common and convenient choice is $g_m \approx \beta_m F_m$, where F_m is the free energy at temperature T_m. With this choice, $P(m)$ becomes approximately flat ensuring all temperatures are frequently visited.

B. Multisequence algorithm

The basic idea of the MS algorithm for biomolecular simulation is to let the sequence s become a dynamic parameter rather than the temperature as in ST. A dynamic s is technically feasible when the potential energy can be written as $E(s, r)$, where s and r are independent variables. This is the case in our coarse-grained protein model which has only backbone degrees of freedom. It can also be achieved in some more detailed models.[26,27]

Similarly to ST, the MS algorithm simulates the joint probability distribution

$$P(s, r) = \frac{1}{Z} e^{-\beta E(s,r) + h(s)} \,, \tag{4}$$

where

$$Z = \sum_{s \in \mathrm{S}} \sum_r e^{-\beta E(s,r) + h(s)} \tag{5}$$

and S is a set of pre-selected sequences, i.e., the sequences to which visits are allowed during a simulation. The simulation parameters $h(s)$, similar to the parameters g_m in ST, control the marginal distribution $P(s) = Z^{-1}\sum_r e^{-\beta E(s,r)+h(s)} = \tilde{Z}^{-1}e^{-\beta F(s)+h(s)}$ and a roughly flat $P(s)$ can be achieved by choosing $h(s) \approx \beta F(s)$, where $F(s)$ is the free energy of sequence s at temperature T.

Two types of MC updates are required to sample from the distrubution in Eq. 4, ordinary conformational updates $r \to r'$ and sequence updates $s \to s'$. The acceptance probability for the latter becomes

$$P_{\text{acc}}(s \to s') = \min[1, \exp\{-\beta\Delta E + \Delta h\}]\,, \tag{6}$$

where $\Delta E = E(s', r) - E(s, r)$ and $\Delta h = h(s') - h(s)$.

Picking a new sequence s' in a sequence update $s \to s'$ can be done in several ways. One possibility is to draw s' randomly from the set S, such than $s' \neq s$. Alternatively, a type of "mutational" move can be used where an amino acid position is first picked and then assigned a new amino acid type. The selection of position and type would have to be chosen such that s' does not end up outside S. In this work, we use the former update which is general and guarantees that ergodicity is fulfilled for any S. Importantly, both updates fulfill detailed balance and therefore lead to the same estimates of equilibrium quantities, such as native state stabilities, for the different sequences in S.

III. MODEL AND METHODS

A. Coarse-grained 3-letter model for protein folding

All calculations were carried out using the coarse-grained model for protein folding developed in Ref. 24. In this model, there are 3 different amino acid types: hydrophobic (h), polar (p) and turn-type (t). The backbone chain is represented atomistically by the N, H, C_α, $H_{\alpha 1}$, C′ and O atoms. By contrast, the sidechain represention is simplified to a single enlarged C_β atom, which is geometrically identical for h and p types. The sidechain is absent for the t type which instead has an $H_{\alpha 2}$ atom. The t type is therefore closely related to glycine. All bond lengths, bond angles, and peptide plane angles (180°) are held fixed. Hence, an N-amino acid chain conformation r can, for any sequence s, therefore be described by the set of $2N$ backbone torsional angles $\{\phi_i,\, \psi_i\}_{i=1}^{N}$.

TABLE I. List of 6 model sequences of different lengths N studied in this work.

Name	N	Sequence
A1	16	pphpphhpphpphhpp
N1	16	phphphpttphphphp
R1	16	pphhphpptthpphhpp
R2	16	ppphphhtthhphppp
A2	35	(A1)ttt(A1)
TN	35	(A1)ttt(N1)

This geometrical description is paired with a simplified but finely tuned energy function with 4 terms: $E = E_{\text{ev}} + E_{\text{loc}} + E_{\text{hb}} + E_{\text{hp}}$. The first two, E_{ev} and E_{loc}, represent excluded-volume effects and local electrostatic effects, respectively. The hydrogen-bond energy, E_{hb}, represents directionally dependent interactions between NH and CO groups and is necessary for secondary structure formation. Finally, the "hydrophobicity" term, E_{hp}, implements pairwise Lennard-Jones-like interactions between the C_β atoms of h amino acids which are necessary for driving chain collapse during folding. Various model parameters, e.g., the strengths of hydrophobic attractions and hydrogen bonding, were determined based on the ability of the model to spontaneously fold a set of model sequences with 18-54 amino acids into structurally diverse and thermodynamically stable native states with both β and α-structure. As it turned out, this strategy made the model robust enough to fold sequences designed to have mixed α and β structures.

B. Model sequences

Six of the model sequences studied in this work, A1, N1, R1, R2, A2, and TN, are given in Table I. In addition, we study two sequence sets S16_{1024} and S35_{1024} with 1,024 sequences each derived from the A1-N1 and A2-TN pairs, respectively, through mutational combinations, as well as the set S16_{144} taken from Ref. 35.

C. Monte Carlo simulation parameters and updates

Both ST and MS simulations are carried out with two types of conformational updates $r \to r'$: (1) a global pivot move (20%) which randomly picks a ϕ_{i} angle or ψ_{i} angle and assign a new value between $-\pi$ and π; and (2) a semi-local move (80%) which turns the ϕ_{i} and ψ_{i}-angles of 4 consecutive amino acids in a coordinated manner.[52]

In our MS simulations, sequence updates $s \to s'$ are carried out in the following way. First, a new sequence s' is picked randomly from the set of pre-selected (allowed) sequences S, such that $s' \neq s$. This new sequence s' therefore differs from s in one or more amino acid positions. Thereafter, the sidechains of the protein, which remains in an unchanged (backbone) conformation r, is re-built according to the new sequence s'. Practically this means that, at the position(s) where the amino acid type has changed, the sidechain is altered according to the type change. For example, if p → t, the C_β atom is removed and replaced with an $\mathrm{H}_{\alpha 2}$ atom or, if p → h, the C_β remains in place but its character is changed to hydrophobic. Finally, the change in total energy ΔE is calculated and the accept-reject criterion in Eq. 6 is applied. If rejected, the old state (s, r) is restored.

A sequence update is attempted every $1,000$ MC steps while temperature updates $m \to m'$ are attempted every 100 steps. The computational cost for sequence updates is somewhat higher than for temperature updates. The latter update does not require any energy calculation and is thus extremely rapid. For the purpose of comparing computational efficiencies of ST and MS, we therefore chose sequence updates to be slightly less frequent than temperature updates while both are fairly frequent. All simulations carried out in this work are summarized in Table II.

D. Observables

Fold stabilities are calculated as in Ref. 36 and described briefly below. First we define two structural similarity measures Q_{IA} and Q_{IB} for folds IA and IB, respectively, indicating the fraction of the fold-specific backbone-backbone hydrogen bonds that have been formed. The fold IA-hydrogen bonds are (2,6), (3,7), (4,8), (5,9), (6,10), (7,11), (8,12), (9,13), (10,14), (11,15) and the fold IB-bonds are (3,14), (5,12), (7,10), (10,7), (12,5), (14,3), where (i,j) indicates a hydrogen bond between the CO group of amino acid i and the NH group of

TABLE II. List of simulations carried out in this work.

Runs	Algorithm	$k_B T$	MC steps/run[a]	Sequences
32	ST	0.43–0.65	1×10^7	A1
32	ST	0.43–0.65	1×10^7	N1
32	ST	0.43–0.65	1×10^7	R1
32	ST	0.43–0.65	1×10^7	R2
32×8[b]	MS	0.43–0.65	18×10^7	$S16_{144}$
16	MS	0.43	5×10^9	$S16_{1024}$
16	MS	0.46	4×10^9	$S35_{1024}$

[a] Excludes a thermalization step with 10^7 MC steps/run.

[b] 32 runs per temperature at 8 different temperatures.

amino acid j. The stabilities of folds IA and IB are defined as the probabilities $P_{\rm IA} = P(Q_{\rm IA} \geq 0.8)$ and $P_{\rm IB} = P(Q_{\rm IB} \geq 0.8)$, respectively, i.e., the probability that at least 80% of the fold's hydrogen bonds are formed. $P_{\rm IA}$ and $P_{\rm IB}$ thus depend on both sequence s and temperature T. For example, $P_{\rm IA} = 0.875 \pm 0.003$ for A1 and $P_{\rm IB} = 0.785 \pm 0.008$ for N1 at $k_{\rm B}T = 0.43$. Structural similarity measures for 35-amino acid folds IIA and IIB are defined as $Q_{\rm IIA} = (Q_{\rm IA}^{1-16} + Q_{\rm IA}^{20-35} + Q_{\rm tert})/3$ and $Q_{\rm IIB} = (Q_{\rm IA}^{1-16} + Q_{\rm IB}^{20-35} + Q_{\rm tert})/3$, respectively, where superscripts on $Q_{\rm IA}$ and $Q_{\rm IB}$ indicate over which amino acid positions those measures are applied to within the longer 35 amino acid sequences and $Q_{\rm tert}$ is a measure that counts the number of C_β-C_β contacts between the two secondary structure elements of these folds.[36] In analogy with $P_{\rm IA}$ and $P_{\rm IB}$, we define the stabilities of folds IIA and IIB as $P_{\rm IIA} = P(Q_{\rm IIA} \geq 0.8)$ and $P_{\rm IIB} = P(Q_{\rm IIB} \geq 0.8)$, respectively. The root-mean-square-deviation, RMSD, is calculated over all C_α atoms.

IV. RESULTS

A. Computational efficiency

We start by applying the MS algorithm to the set $S16_{144}$ across a range of temperatures T (see Table II). Two of the sequences in $S16_{144}$ are A1 and N1 (see Table I) which fold into stable α-helix and β-hairpin structures, respectively, as shown in Fig. 2B. Because A1

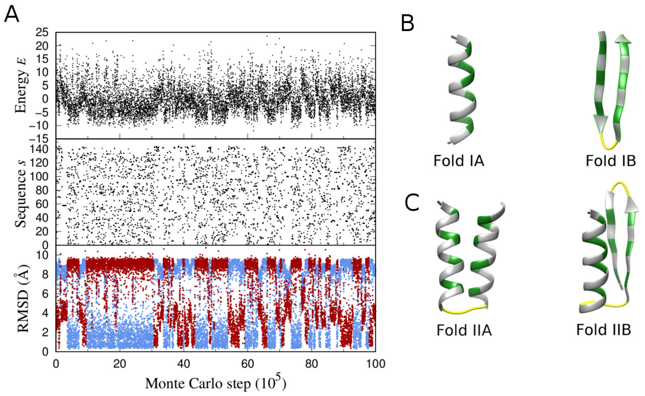

FIG. 2. (A) Example of an MS simulation of the sequence set $S16_{144}$ carried out at $k_{\mathrm{B}}T = 0.43$. The plot shows the MC evolution of the sequence s (numbered 1–144), the total potential energy E and the root-mean-square deviation (RMSD) calculated against the representative fold IA (light blue) and fold IB (dark red) structures in (B). Representative structures of folds (B) IA, IB, (C) IIA and IIB, chosen to be the minimum-energy conformations found for the sequences A1, N1, A2 and TN, respectively.

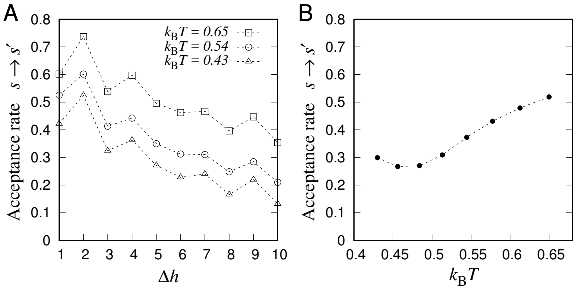

FIG. 3. Acceptance rates for $s \to s'$ updates in MS simulations of the $S16_{144}$ sequence set as a function of (A) the number of changed amino acid positions Δh and (B) temperature T. Acceptance rates for 3 different T's are shown in (A).

and N1 differ at 10 positions, 10 consecutive point mutations can transform A1 into N1, and vice versa. The binary sequence space between A1 and N1 in which any combination of these mutations have been carried out, therefore contains $2^{10} = 1,024$ sequences. The 144 sequences in $S16_{144}$ were selected from this binary space with the constraints that the total number of hydrophobic amino acids are not too high and that they are not too unevenly distributed along the sequence.[35]

Figure 2 illustrates a typical MS simulation trajectory carried out at the lowest studied temperature which is below the folding temperature of both A1 and N1.[35,36] From the MC evolution of the total energy E, sequence index s, and RMSD values from the representative structures in Fig. 2B, it is evident that visits to various sequences drive transitions into a range of structural states. In particular, there are frequent visits to α-helix and β-hairpin structures and transitions between them are accompanied by a shift in which sequences are preferably visited. For example, visits to high s-indices, including N1 with index 144, tend to coincide with formation of β-hairpin structures as required to generate correct equilibrium conformational ensembles.

One might have suspected that the MS algorithm would be hampered by poor acceptance rates for sequence updates. However, this is not the case in our model. We carry out updates $s \to s'$ by picking a new random sequence $s' \neq s$ from the set of allowed sequences. The (average) acceptance rate depends on both T and the step in sequences space Δh, i.e., the number of amino acid positions changed, as shown in Fig. 3. At the lowest T and highest Δh, acceptance rates are only around 0.1-0.2. However, for most other T and Δh the overall acceptance rates are substantially higher and often above the oft-quoted rule-of-thumb value 0.25[53] (see Fig 3B). An increased acceptance rate can easily be achieved by restricting proposed updates such that $\Delta h \leq \Delta h_{\max}$, where $\Delta h_{\max}$ is a maximum step size, which might be necessary for longer chains. For example, $\Delta h_{\max} = 1$ would be equivalent to applying only a "mutational" update, i.e., picking a random (allowed) position and changing the amino acid type at that position.

We now compare the results from our MS calculations with simulated tempering (ST) simulations carried out on 4 of the 144 sequences, namely A1 and N1 and two random sequences, R1 and R2, chosen at distances $h = 4$ and $h = 6$ from A1, respectively (see Table I). While ST provides the thermodynamics of a given sequence across a range of T in a single run, an MS simulation provides the thermodynamics of all 144 sequences at one T. We adjust the simulation lengths for ST and MS runs such that roughly the same number of sampled conformations are obtained for each s and T combination, thus ensuring that similar computational resources are used for the two algorithms (see Table II). We first validate the MS algorithm by comparing the average total energy, $\langle E \rangle$, calculated for these 4 sequences with the two different methods (see Supplementary Information). The two sets of results are entirely consistent showing that, for a given s and T, the MS and ST algorithms indeed sample the same (Boltzmann) distribution.

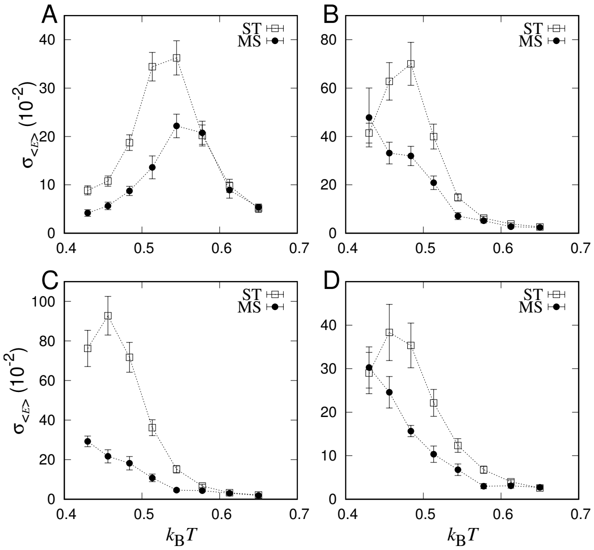

FIG. 4. Comparing sampling efficiencies of the MS and ST algorithms. Statistical errors $\sigma_{\langle E\rangle}$ of the average total energy $\langle E\rangle$ obtained for the sequences (A) A1, (B) N1, (C) R1 and (D) R2 (see Table I) at different temperatures T. Simulation lengths in the two methods are adjusted such that the number of conformations sampled per sequence and temperature is roughly the same (see text).

As a way to assess conformational sampling efficiency, we compare in Fig. 4 the statistical error, $\sigma_{\langle E\rangle}$, of the average energy $\langle E\rangle$ for the 4 sequences obtained using ST and MS, respectively. Because approximately the same number of sampled conformations were obtained for each combination of s and T, we compare the statistical errors directly. At the highest studied T, which is well above the folding temperature of both A1 and N1, the two algorithms give almost identical statistical errors. This can be understood by noting that at high-T the free-energy landscape is relatively smooth and conformational space requires little difficulty to sample. The benefit of adding a dynamic parameter, whether s or T, is apparently minimal under these conditions. However, at lower T, the $\sigma_{\langle E\rangle}$ values from MS is often smaller than those from ST and never significantly higher. For example, at the lowest T, the precision in the estimate of $\langle E\rangle$ is roughly twice as high in MS than ST for A1 and R1, and roughly the same for N1 and R2.

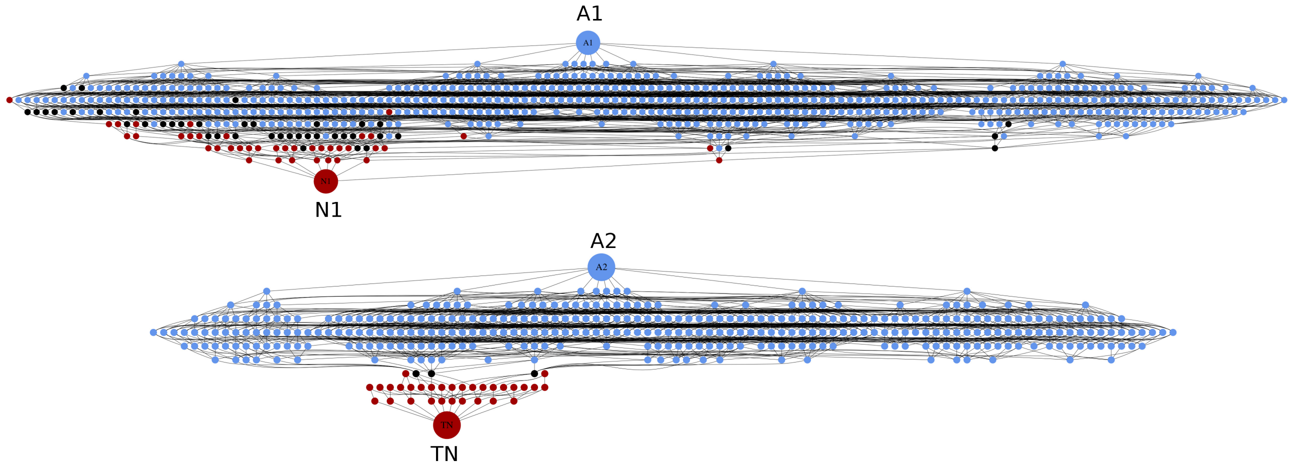

FIG. 5. Networks of sequences connecting folds IA and IB (top) and folds IIA and IIB (bottom). Each node represents a stable sequence ($P_{\text{tot}} \geq P_{\text{cut}}$ where $P_{\text{cut}} = 0.50$) that folds into either IA or IIA (light blue), IB or IIB (dark red), or is classified as bistable ($B > 0.5$, black). A line between two nodes indicates that the sequences differ at only one position. Graph created using the tool Graphviz[54] obtained from www.graphviz.org.

B. Exploring sequence space: IA/IB and IIA/IIB fold connectivities

We now turn to the full binary sequence sets S16_{1024} and S35_{1024} with 1,024 sequences each. By applying the MS method to these two sets (see Table II), we determine the low-T thermodynamic behavior of each included sequence. In particular, we calculate the stabilities of folds IA and IB, P_{IA} and P_{IB}, for all sequences in S16_{1024} and the stabilities of folds IIA and IIB, P_{IIA} and P_{IIB}, for all sequences in S35_{1024} (see Methods). The relative statistical errors on these quantities vary but are only a few percent at the most, despite the large number of sequences included.

Having calculated these fold stabilities, we are in a position to determine if there are pathways in sequence space that lead to abrupt IA-IB or IIA-IIB fold changes, i.e., paths that do not pass through any unstable intermediate sequence. To this end, we construct graphs in which each stable sequence is represented by a node and any two nodes are connected if their sequences differ at a single amino acid position. To determine if a sequence is stable we use the criterion $P_{\text{tot}} > P_{\text{cut}}$, where $P_{\text{tot}} = P_{\text{IA}} + P_{\text{IB}}$ and $P_{\text{IIA}} + P_{\text{IIB}}$ for the IA-IB and IIA-IIB fold pairs, respectively; P_{tot} thus indicates the total stability of a sequence across the two competing folds. The precise network depends, of course, on the cut-off value P_{cut} and a higher P_{cut} generally means a selection of more stable pathways.

Fig. 5 illustrates the networks obtained with $P_{\mathrm{cut}} = 0.50$ showing that both the IA-IB and IIA-IIB fold pairs are connected in sequence space at this stability threshold. A precise analysis shows that there are 516,972 viable IA-IB paths and 57,912 viable IIA-IIB paths. These paths represent 14.2 % and 1.6 % of all possible paths, respectively, because there are a total of $10! = 3,628,800$ possible paths between start and end points in both cases. Hence, folds IA and IB are rather highly connected in our model for $P_{\mathrm{cut}} = 0.50$. For $P_{\mathrm{cut}} = 0.60$, the numbers are 104,640 paths (2.9%) for IA-IB and 22,512 (0.6%) paths for IIA-IIB. We find that there are no possible IA-to-IB or IIA-to-IIB paths when $P_{\mathrm{cut}} \geq 0.74$ and ≥ 0.66, respectively.

C. Biophysical properties of fold-to-fold mutational pathways

An apparently general characteristic of designed and natural proteins that exhibit mutation-induced fold switching is a reduced stability near the switch point.[34,37,38,40,43] Our model proteins exhibit a similar trend. Fig. 6A and B show the average total stability P_{tot} for sequences found at different Hamming distances h from the starting point. Intermediate sequences are less stable than sequences at distances $h = 0$ (A1 or A2) and $h = 10$ (N1 or TN), although there are large variations between sequences as indicated by the upper and lower bounds. There is nonetheless a clear statistical trend that sequences become gradually less stable as successive mutations are applied to any of the 4 start and end points until a minimum is reached.

However, the smooth stability trends in Fig. 6A and B belie the real character of the individual mutational pathways which tend to exhibit an abrupt switch between the two folds. To see this and to further examine the character of the fold transitions in our model, we make a distinction between two types of stable sequences: those that fold into a single unique fold, thus behaving as classical proteins, and those that display substantial stabilities of both folds. Such “bistable” sequences are interesting from a biophysical perspective in that they are able to fold into two alternative folds. Indeed, bistable sequences have been proposed to play a role in the evolution of new protein folds.[55] We consider a sequence to be

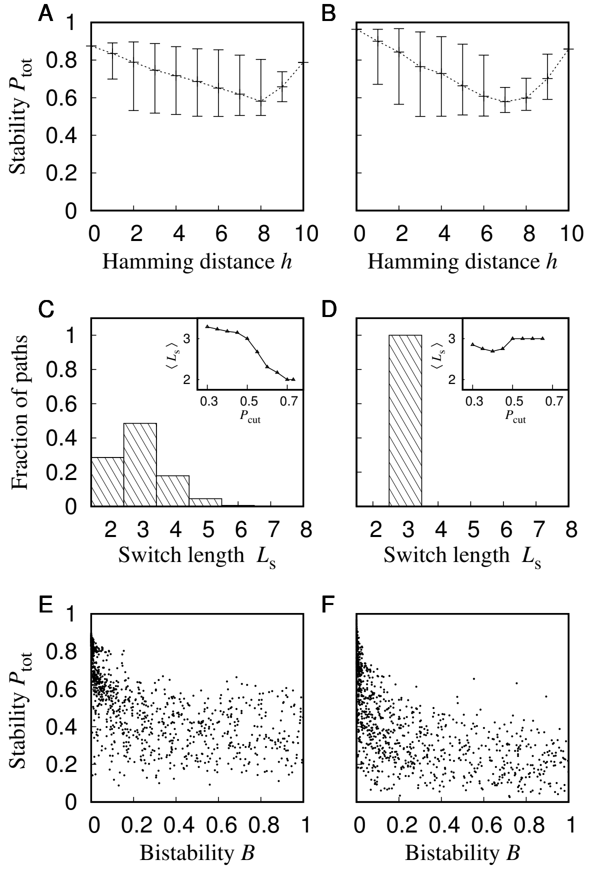

FIG. 6. Stability properties of mutational pathways. The total stability P_{tot} as a function of the distance h from A1 averaged over all (A) IA-IB and (B) IIA-IIB mutational paths obtained with $P_{\mathrm{cut}} = 0.50$. Error bars indicate maximum and minimum P_{tot} values. The distribution of switch lengths L_{s} for the (C) IA-IB and (D) IIA-IIB mutational paths ($P_{\mathrm{cut}} = 0.50$). C and D insets: Average switch length $\langle L_{\mathrm{s}} \rangle$ across all paths as a function of P_{cut}. Scatter plots of P_{tot} versus bistability B for all sequences in (E) $\mathrm{S16}_{1024}$ and (F) $\mathrm{S35}_{1024}$, where $B = 1 - \Delta P / P_{\mathrm{tot}}$ and $\Delta P = |P_{\mathrm{IA}} - P_{\mathrm{IB}}|$ or $|P_{\mathrm{IIA}} - P_{\mathrm{IIB}}|$.

bistable if $B > 0.5$, where B is a bistability measure (see Fig. 6 legend). In principle, a fold transition can then occur directly between two classical proteins with unique native folds, or it can proceed via one or more intermediate bistable sequences which populate both folds. We define the switch length of a mutational pathway $L_{\mathrm{s}} = 2 + n_{\mathrm{B}}$, where n_{B} is the number of bistable sequences in between the two classical sequences that define the switch point. Hence, a path with $L_{\mathrm{s}} = 2$ accomplishes a fold switch in a single mutational step without going through a bistable point. From the distributions of L_{s} in Fig. 5C and D, taken over all pathways with $P_{\mathrm{cut}} = 0.50$, it can be seen that fold switching along individual pathways are typically completed in only 1-2 mutations and a single step is often sufficient to switch between the IA and IB folds. Hence, fold switching is typically abrupt and, for $P_{\mathrm{cut}} = 0.5$, it is fairly common that viable pathways pass through one or two bistable sequences.

Interestingly, switching between folds IA and IB through one or more bistable sequences become less and less frequent as selections for more stable pathways are made. This can be seen from the decrease in $\langle L_{\mathrm{s}} \rangle$ as a function of P_{cut} (see Fig. 5C(inset)). For $P_{\mathrm{cut}} \geq 0.70$, there are no longer any remaining path between the α-helix and β-hairpin that passes through a bistable sequence because $\langle L_{\mathrm{s}} \rangle = 2$. An underlying reason for the occurrence of sharper fold switches for more stable mutational pathways is apparent from a comparison between P_{tot} and B across all sequences in $\mathrm{S16}_{1024}$. As shown in Fig. 5E, sequences with the highest P_{tot} tend to exhibit very little bistability. Hence, highly stable paths are therefore forced to go through abrupt switch points where they transition directly between folds in a single step. The situation for the IIA-IIB fold pair is more complicated. We find that, just as for $\mathrm{S16}_{1024}$, sequences in $\mathrm{S35}_{1024}$ follow the trend that the highest P_{tot} values occur for only classical, low-B proteins. One might therefore expect that selection of more stable IIA-IIB paths would decrease $\langle L_{\mathrm{s}} \rangle$, however, this is not the case as such abrupt switch points between the IIA and IIB are not available for $P_{\mathrm{cut}} \geq 0.50$ (cf. Fig. 5 bottom). As a result, bistable sequences do play a crucial role in bridging the IIA and IIB folds, although passing though these sequences lead to additional reduction in stability at the switch point.

V. DISCUSSION

We have evaluated a biomolecular simulation algorithm that works by making the biological sequence a dynamic parameter. As a test, we applied it on a CG model for protein folding. The results indicate that there are two main benefits of this approach. Firstly, it provides a convenient way to sample the canonical distributions of large numbers of sequences in a single run and, secondly, it enhances the sampling of conformational space meaning it can be applied directly to low temperatures. The conformational sampling efficiency can be assessed from the comparison with ST. Although there is no single “fair” way to compare the two methods, we chose as a measure of efficiency the statistical error of the total energy, $\sigma_{\langle E \rangle}$, obtained with roughly the same computational cost per temperature and sequence. At the highest studied temperatures, we find that the statistical errors $\sigma_{\langle E\rangle}$ are basically the same. This finding is not unexpected because conformational sampling of short polymers at high T does not involve crossing any major energy barriers. As a result, successively sampled conformations for a given combination of sequence s and temperature T, are likely uncorrelated in both methods which leads to similar $\sigma_{\langle E\rangle}$ values.

At lower temperatures, we find that the MS simulations often yields significantly smaller $\sigma_{\langle E\rangle}$ than ST simulations. It is notable that this acceleration in conformational sampling vis-á-vis ST is achieved despite that simulations are carried out at constant T. Therefore, rather than promoting escape from local minima by visits to higher T, as in ST[45,46] or temperature replica-exchange,[49] MS simulations must escape the minima occurring for one sequence through visits to other sequences. How this process works can be envisioned by considering an MS simulation that is visiting a sequence s and is trapped in a local minimum, requiring that a high free-energy barrier is overcome for escape. The trapped state could be, e.g., a compact β-sheet rich state with a particular register. Eventually, sequence updates will carry the simulation to other sequences $\hat{s}$ while it conformationally still remains in the trapped state. However, the free energy barrier of escape might be substantially lower for $\hat{s}$ than for s, or the barrier may even be altogether absent if the trapped state is unstable for $\hat{s}$, leading to rapid escape from the minimum. The above reasoning also implies that the performance of the MS algorithm likely depends on the size of the sequence set S as well as the conformational properties of the sequences. Specifically, the performance of MS simulations of proteins at low T may benefit from the inclusion of at least a few sequences with poor stability properties, such that partial unfolding of the chain is regularly triggered and thus promoting transitions to new conformational states.

Individual protein sequences that exhibit spontaneous transitions between widely different competing conformational states, such as fold switching proteins, are especially challenging to molecular simulation methods. Representative sampling in such cases requires multiple transitions between highly different states, which can be a slow process. This problem has been addressed by using Hamiltonian replica-exchange techniques to couple Gō-like terms, i.e., energy terms with an artificial bias towards a given structure, with a physical force field.[56–58] This way exchange moves "feed" diverse conformations into the replica corresponding to the physical force field and enhance sampling.[58] We did not specifically study sampling efficiency for sequences exhibiting competing states, such as bistable sequences. It appears likely, however, that a very similar type of "feeding" of conformations would take place in MS simulations although the coupling occurs instead with other sequences rather than with Gō-type terms.

We emphasize that the MS method should not be seen as a general technique to speed up conformational sampling. However, our results indicate that for CG models that permit sequence updates to be carried out as a simple Metropolis step and when visits to higher T is unwanted (or unnecessary), our method can be a highly efficient way to sample the equilibrium behavior of many sequences. This opens up for various applications, such as exploring sequence effects on the conformational properties of disordered[59] or denatured[60] proteins, or as a tool in efforts to combine population genetics and molecular simulations.[61] While applied to proteins in this work, we note that the theoretical framework of the MS algorithm is equally valid for other bio-macromolecules, including DNA and RNA. The ability to promote conformational sampling without resorting to an increase in T may make the method useful in simulations of biomolecules in ordered phases, such as lipid bilayers or double-stranded DNA, where escape from local minima can be especially challenging[62,63] and elevated T is typically avoided in simulations because it may lead to unwanted perturbations or unraveling of the basic underlying structure.

VI. CONCLUSION

We have evaluated an algorithm for biomolecular simulations that allows the thermodynamics of multiple sequences to be calculated in a single run. We applied the algorithm to protein folding and showed that the thermodynamic behavior of >1,000 amino acid sequences with up to 35 amino acids could be determined in an intermediate-resolution CG model. The method performs a random walk in sequence space which is especially useful at low temperature as it promotes escape from local minima present in the free energy landscapes of individual sequences. The method might be suitable for CG simulations of various other biomolecular systems, such as peptides in phospholipid bilayers, where sampling at elevated temperatures is not desirable.

VII. SUPPLEMENTARY MATERIAL

See supplementary material for the validation of the MS algorithm.

VIII. ACKNOWLEDGMENTS

This work was supported by grants from Memorial University and National Sciences and Engineering Research Council of Canada (RGPIN-2016-05014).

REFERENCES

[1] S. Piana, J. L. Klepeis, and D. E. Shaw, Curr Opin Struct Biol **24**, 98 (2014).

[2] F. Ding, D. Tsao, H. Nie, and N. V. Dokholyan, Structure **16**, 1010 (2008).

[3] A. Irbäck and S. Mohanty, J Comput Chem **27**, 1548 (2006).

[4] A. Verma and W. Wenzel, Biophys J **96**, 3483 (2009).

[5] J. S. Yang, W. W. Chen, J. Skolnick, and E. I. Shakhnovich, Structure **15**, 53 (2007).

[6] R. C. Bernardi, M. C. Melo, and K. Schulten, Biochim Biophys Acta **1850**, 872 (2015).

[7] S. R. McGuffee and A. H. Elcock, PLOS Comput Biol **6**, e1000694 (2010).

[8] Y. Miao, J. E. Johnson, and P. J. Ortoleva, J Phys Chem B **114**, 11181 (2010).

[9] J. R. Perilla, J. A. Hadden, B. C. Goh, C. G. Mayne, and K. Schulten, J Phys Chem Lett **7**, 1836 (2016).

[10] K. Lindorff-Larsen, S. Piana, R. O. Dror, and D. E. Shaw, Science **334**, 517 (2011).

[11] I. Yu, T. Mori, T. Ando, R. Harada, J. Jung, Y. Sugita, and M. Feig, ELife **5** (2016).

[12] S. Riniker, J. R. Allison, and W. F. van Gunsteren, Phys Chem Chem Phys **14**, 12423 (2012).

[13] H. I. Ingólfsson, C. A. Lopez, J. J. Uusitalo, D. H. de Jong, S. M. Gopal, X. Periole, and S. J. Marrink, WIREs Comput Mol Sci **4**, 225 (2014).

[14] L. K. Beland, P. Brommer, F. El-Mellouhi, J. F. Joly, and N. Mousseau, Phys Rev E **84**, 046704 (2011).

[15] E. A. Proctor, F. Ding, and N. V. Dokholyan, WIREs Comput Mol Sci **1**, 80 (2011).

[16] A. Vitalis and R. V. Pappu, Annu Rep Comput Chem **5**, 49 (2009).

[17] A. Zarrinpar, S. H. Park, and W. A. Lim, Nature **426**, 676 (2003).

[18]L. Hakes, S. C. Lovell, S. G. Oliver, and D. L. Robertson, Proc Natl Acad Sci USA **104**, 7999 (2007).

[19]R. Rohs, X. Jin, S. M. West, R. Joshi, B. Honig, and R. S. Mann, Annu Rev Biochem **79**, 233 (2010).

[20]C. A. Ross and M. A. Poirier, Nat Med **10 Suppl**, S10 (2004).

[21]F. O. Tzul, D. Vasilchuk, and G. I. Makhatadze, Proc Natl Acad Sci USA **114**, E1627 (2017).

[22]B. G. Wensley, S. Batey, F. A. Bone, Z. M. Chan, N. R. Tumelty, A. Steward, L. G. Kwa, A. Borgia, and J. Clarke, Nature **463**, 685 (2010).

[23]O. G. Vukmirovic and S. M. Tilghman, Nature **405**, 820 (2000).

[24]A. Bhattacherjee and S. Wallin, Biophys J **102**, 569 (2012).

[25]A. Irbäck and F. Potthast, J. Comp. Phys. **103**, 10298 (1995).

[26]A. Bhattacherjee and S. Wallin, PLOS Comput Biol **9**, e1003277 (2013).

[27]S. Wallin, Methods Mol Biol **1561**, 201 (2017).

[28]P. N. Bryan and J. Orban, Curr Opin Struct Biol **20**, 482 (2010).

[29]Y. G. Chang, S. E. Cohen, C. Phong, W. K. Myers, Y. I. Kim, R. Tseng, J. Lin, L. Zhang, J. S. Boyd, Y. Lee, S. Kang, D. Lee, S. Li, R. D. Britt, M. J. Rust, S. S. Golden, and A. LiWang, Science **349**, 324 (2015).

[30]S. Meier, P. R. Jensen, C. N. David, J. Chapman, T. W. Holstein, S. Grzesiek, and S. Ozbek, Curr Biol **17**, 173 (2007).

[31]E. S. Kuloglu, D. R. McCaslin, J. L. Markley, and B. F. Volkman, J Biol Chem **277**, 17863 (2002).

[32]B. M. Burmann, S. H. Knauer, A. Sevostyanova, K. Schweimer, R. A. Mooney, R. Landick, I. Artsimovitch, and P. Rösch, Cell **150**, 291 (2012).

[33]X. Luo, Z. Tang, G. Xia, K. Wassmann, T. Matsumoto, J. Rizo, and H. Yu, Nat Struct Mol Biol **11**, 338 (2004).

[34]P. A. Alexander, Y. He, Y. Chen, J. Orban, and P. N. Bryan, Proc Natl Acad Sci USA **106**, 21149 (2009).

[35]C. Holzgräfe and S. Wallin, Biophys J **107**, 1217 (2014).

[36]C. Holzgräfe and S. Wallin, Phys Biol **12**, 026002 (2015).

[37]M. Kouza and U. H. Hansmann, J Phys Chem B **116**, 6645 (2012).

[38]L. Sutto and C. Camilloni, J. Comp. Phys. **136**, 185101 (2012).

[39] C. A. Ramirez-Sarmiento, J. K. Noel, S. L. Valenzuela, and I. Artsimovitch, PLOS Comput Biol **11**, e1004379 (2015).

[40] T. Sikosek, H. Krobath, and H. S. Chan, PLOS Comput Biol **12**, e1004960 (2016).

[41] N. Hansen, J. R. Allison, F. H. Hodel, and W. F. van Gunsteren, Biochemistry **52**, 4962 (2013).

[42] T. Sikosek and H. S. Chan, J R Soc Interface **11**, 20140419 (2014).

[43] Y. He, Y. Chen, P. A. Alexander, P. N. Bryan, and J. Orban, Structure **20**, 283 (2012).

[44] L. L. Porter, Y. He, Y. Chen, J. Orban, and P. N. Bryan, Biophys J **108**, 154 (2015).

[45] E. Marinari and G. Parisi, Europhys Lett **19** (1992).

[46] A. P. Lyubartsev, A. A. Martsinovski, S. V. Shevkunov, and P. N. Vorontsov-Velyaminov, J. Comp. Phys. **96**, 1776 (1992).

[47] A. Mitsutake, Y. Sugita, and Y. Okamoto, Biopolymers **60**, 96 (2001).

[48] B. A. Berg, Eur Phys J Spec Top **226**, , 551 (2017).

[49] R. H. Swendsen and J.-S. Wang, Phys. Rev. Lett. **57**, 2607 (1986).

[50] U. H. E. Hansmann, Chem Phys Lett **281**, 140 (1997).

[51] A. Laio and M. Parrinello, Proc Natl Acad Sci USA **99**, 12562 (2002).

[52] G. Favrin, A. Irbäck, and F. Sjunnesson, J. Comp. Phys. **114**, 8154 (2001).

[53] A. Gelman, G. O. Roberts, and W. R. Gilks, Bayesian Statist **5**, 599 (1996).

[54] E. R. Gansner and S. C. North, Software – practice and experience **30**, 1203 (2000).

[55] T. Sikosek, E. Bornberg-Bauer, and H. S. Chan, PLOS Comput Biol **8**, e1002659 (2012).

[56] J. H. Meinke and U. H. E. Hansmann, J Phys Condens Matter **19**, 285215 (2007).

[57] W. Zhang and J. Chen, J Chem Theory Comput **10**, 918 (2014).

[58] N. A. Bernhardt, W. Xi, W. Wang, and U. H. Hansmann, J Chem Theory Comput **12**, 5656 (2016).

[59] R. K. Das and R. V. Pappu, Proc Natl Acad Sci USA **110**, 13392 (2013).

[60] W. Meng, B. Luan, N. Lyle, R. V. Pappu, and D. P. Raleigh, Biochemistry **52**, 2662 (2013).

[61] A. W. Serohijos and E. I. Shakhnovich, Curr Opin Struct Biol **26**, 84 (2014).

[62] T. Bereau and M. Deserno, J Membr Biol **248**, 395 (2015).

[63] J. Curuksu and M. Zacharias, J. Comp. Phys. **130**, 104110 (2009).