MBAR-enhanced Lattice Monte Carlo Simulation of the Effect of Helices on Membrane Protein Aggregation

Yuanwei Xu[a,*], P. Mark Rodger[a,b]

[a] *Centre for Scientific Computing, University of Warwick, UK*
[b] *Department of Chemistry, University of Warwick, UK*

Abstract

We study the effect of helical structure on the aggregation of proteins using a simplified lattice protein model with an implicit membrane environment. A recently proposed Monte Carlo approach, which exploits the proven statistical optimality of the MBAR estimator in order to improve simulation efficiency, was used. The results show that with both two and four proteins present, the tendency to aggregate is strongly expedited by the presence of amphipathic helix (APH), whereas a transmembrane helix (TMH) slightly disfavours aggregation. When four protein molecules are present, partially aggregated states (dimers and trimers) were more common when the APH was present, compared with the cases where no helices or only the TMH is present.

Keywords: lattice Monte Carlo, parallel tempering, multicanonical sampling, protein aggregation, twin-arginine translocation

1. Background and Introduction

Proteins are macromolecules that are essential to the functioning of living organisms. The primary structure of a protein consists of a chain of amino acids that is coded by genes. Typically, upon synthesis on ribosomes, the extended chain must fold into a specific three-dimensional structure called the

[*] Corresponding author
Email addresses: yuanwxu@warwickgrad.net (Yuanwei Xu), p.m.rodger@warwick.ac.uk (P. Mark Rodger)

native structure. Only in its native structure can a protein be fully functional. The correct folding to native state depends both on interactions among different residual pairs of the molecule and on multiple contributing factors from the crowded cellular milieu [1]. Changes in the cellular condition—e.g. pH or temperature—or mutations in the protein, can all lead to misfolding, which is then prone to self-assembly and formation of aggregates [2]. One type of aggregate, known as amyloid fibrils, has an abundance of β sheet structures and is the most studied form of aggregate due to its toxic, pathological nature and association with numerous human diseases. In fact, it has been suggested that amyloid fibrils were linked with approximately 50 disorders including such neurodegenerative diseases as Alzheimers, mad cow and Parkinsons [3]. The study of amyloid fibrils has been so dominant in protein aggregation research that a recent review on the computational aspects of protein aggregation focused entirely on this type of aggregate [4].

Whereas amyloid fibrils may be considered the end product of aggregation, in the sense that the process is irreversible, there are cases where proteins are recruited only to form a functional aggregate, and then dissociated back into a monomeric state once that functional role is fulfilled. For instance, our work in this paper is motivated by certain membrane protein, TatA, as a key component in a protein translocation mechanism—the twin-arginine translocation (Tat) pathway [5]. The aggregation of TatA is essential to the process in that it is believed to form the translocation channel [6].

Protein aggregation is a very challenging computational problem because the length scale of the process can be one to hundreds of nanometers and the time scale ranges from milliseconds to a year [4]. It is therefore a prime application for coarse-graining and multi-scaling methods. A comprehensive review of the computational models and methodologies used for protein aggregation, applied in particular to amyloid fibrils, can be found in [4]. From a modeling perspective, there are models of different resolutions, from simple models that sacrifice sequence-level resolution, lattice models that represent an amino acid by one or more beads, coarse-grained models that balance accuracy and computational efficiency, to atomistic models that are more accurate and detailed but too computationally prohibitive to study the full assembly process of interest. On the other hand, computational approaches can be applied to models of various resolutions, although certain methods may be more suitable to a particular class of models.

Molecular dynamics (MD) is a widely used and perhaps mainstream computational approach to the study of protein aggregation. Replica exchange molecular dynamics (REMD) [7], metadynamics [8], and their variants are among the commonly used methods. While MD is limited to a small time step to ensure accuracy of the time integration method, Monte Carlo (MC) methods are more flexible in terms of choosing an appropriate move set which may not be physical but is chosen to suit the purpose of the study. Note that the aforementioned MD methods have their MC counterparts: the REMD method is adapted from the parallel tempering method of [9] and [10], while metadynamics may be viewed as the MD incarnation of the Wang-Landau sampling [11]. Both parallel tempering and Wang-Landau sampling predate, respectively, REMD and metadynamics, which might suggests that advances in MD methods that enhance sampling are often inspired by those advances in MC methods. The versatility of MC makes it a handy tool to capture events occurring at a much longer timescale than can be probed by MD, or to study simplified systems that embody the right physical properties. The insights gained from MC modeling of the simplified system can be exploited to guide the design of experiments in relation to the underlying complex physical or biological processes, or be used to help in making a better choice of the simulation protocol of a high-resolution atomistic simulation.

This article is concerned with the Monte Carlo modeling of TatA aggregation within a biological membrane, using a minimal lattice model that encodes the relevant structural characteristics of the TatA molecule. We have chosen a lattice MC model because we wish to probe the underlying equilibrium between aggregated and dispersed states, and the necessary ergodic sampling is simply not viable with higher resolution models. We use a novel MC method that optimally combines parallel tempering (PT) and multicanonical sampling (MUCA) [12], through the use of the multistate Bennett acceptance ratio (MBAR) estimator [13], an approach first introduced in [14] and which we refer to as MBAR-enhanced MC. It provides a framework to calculate density of states from multiple equilibrium simulations using the MBAR estimator, and so there is a wide choice of methods that can be used in place of PT and MUCA as are used here.

We emphasize that the TatA aggregation in the Tat mechanism serves as a motivation to the current lattice model, and that our simulation results are not intended for detailed quantitative comparison with experiment at this stage. To our knowledge, there has not been a computational study of the TatA assembly process within a membrane, although MD simulations have been utilized to study the stability of a preformed TatA oligomer [15]. Given this context, the purpose of this paper is twofold. In the first place, this paper presents a feasibility study for conducting equilibrium membrane protein aggregation studies with MBAR-enhanced lattice MC. At the same time, it serves as a motivation to more research, especially computational studies, into this aspect, and we hope that our work can draw interest from computational scientists in protein aggregation to the Tat translocation process, for which the underlying mechanistic understanding is at least as embryonic as that for amyloid fibrils. In the rest of this section, we shall briefly review the Tat process.

1.1. The Twin-arginine Translocation pathway

The Twin-arginine translocation (Tat) pathway is one of two major pathways cells have for transporting proteins across membranes. It is involved in the export of proteins across bacterial cytoplasmic membranes and across the thylakoid membranes in plant chloroplasts, and is essential for bacterial pathogenesis and for plant photosynthesis [16]. The translocated proteins are referred to as substrates; these are proteins that need to be transported to perform their functions either within the cell or in extracellular space. One distinctive feature of the Tat mechanism is that substrate proteins are transported in a folded manner, contrary to the general secretory (Sec) pathway which transports proteins in an unfolded state [16]. The name *Tat* is an acronym for “twin-arginine translocation” and comes from the unique, consensus twin-arginine (RR) motif that is a key feature of the amino acid sequence of the signal peptide that triggers Tat translocation. Major components of the Tat translocon are membrane proteins from the TatABC family; these are small integral membrane proteins that, when forming complexes that have the right structural organization, allow the folded substrate to be translocated without compromising the permeability of the lipid bilayer. TatA consists of a single transmembrane helix (TMH), and an amphipathic helix (APH) that lies along the membrane surface. The two helices form approximately a right angle and are connected by a small loop. The TatA structure of bacterium *Bacillus subtilis* has been determined in atomic resolution by nuclear magnetic resonance spectroscopy [17]. TatC consists of six TMHs and has limited conformational flexibility [5]. TatB has similar structure to TatA and the two are best discriminated by their biochemical behaviour: whereas TatA proteins oligomerize to form the translocation channel, TatB proteins form a 1 : 1 complex with TatC and plays a role in substrate recognition prior to the transient translocation process [5].

The manner in which the components just described combine to translocate folded proteins is generally described in terms of three major steps. While many of the mechanistic details about these steps remains uncertain, the overall pattern for the mechanism is well accepted.

1. The twin arginine signal peptide is added to the protein; additional subunits or cofactors may also bind to the protein at this stage.
2. Through the twin arginine signal peptide, the protein recognises, and binds to, a TatBC complex within the membrane.
3. This binding event then nucleates the aggregation of a variable number of TatA proteins to form an oligomer with the TatBC-protein complex, thereby generating a pore that is permeable to the folded protein.
4. The complex (protein/signal peptide/TatBC) then dissociates, releasing the protein, and triggering the dispersion of the TatA oligomer.

The dispersion of the TatA oligomer once translocation is completed is an essential step, since a persistent oligomer would have undesirable consequences for the cell, such as ion leakage [18]. Also, experimental results show that the TatA channel can vary its diameter to accommodate substrates of different sizes [6], implying that translocation is mediated by oligomerization of variable amounts of TatA monomer.

The fact that Tat is able to transport folded proteins makes it particularly challenging compared to the Sec pathway, because membranes must maintain a permeability barrier to ions and small molecules during transport. As an example, the *Escherichia. coli* Tat pathway is able to transport substrates of up to 70 Å in diameter, whereas an unfolded polypeptide chain is only about 12 Å in diameter [19]. Due to this distinctive feature that the Tat mechanism possesses, models of the dynamics of the translocation process have been proposed (see [20] for a review), including one that predicts a local weakening of the membrane that is sufficient for the substrate to move through [21].

Clearly, the structure of the TatA oligomer is crucial to understanding the Tat mechanism. However, due to the transient nature of an active translocation complex, it is difficult to conduct experimental analyses and thereby establish its structural organization during translocation. Computer simulation can be an indispensable complementary tool in elucidating the Tat mechanism. Indeed, molecular dynamics simulations have been applied to investigate the stability of an experimentally-proposed solubilized TatA oligomer structure in membrane [15]; however it is unclear whether TatA assembly in the native membrane environment will result in the same oligomeric structure as determined in a detergent solution.

Our lattice models (Section 2), on the other hand, allow us to simulate the assembly process in an (implicit) membrane environment and explore its equilibrium properties rigorously. Once the equilibrium behaviour of TatA aggregation is adequately explained, we can then begin to model how the TatBC complex and the substrate interact with, and influence, the aggregation of TatA within the Tat process. We stress that, at this early stage, our model of the TatA is

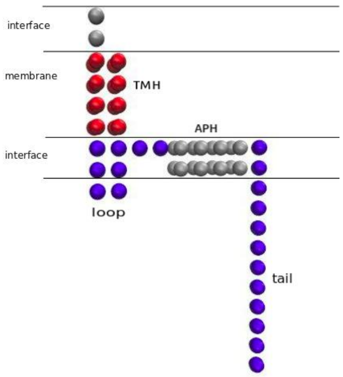


Figure 1: Lattice model of TatA. Four regions are identified: transmembrane helix (TMH), loop, amphipathic helix (APH) and the hydrophilic tail. Boundaries of membrane and interface are also shown. Color scheme: hydrophobic—red, hydrophilic—blue, amphipathic—gray. Chain sequence: $(P)_{12}(H2)_{16}(P)_8(H)_{16}(H2)_2$, where the notation $(X)_n$ means that bead type X is repeated n times.

intended to give qualitative insight into the key features, particularly relating to secondary structure, that would need to be incorporated into a subsequent, more quantitative, model.

2. Model and method

2.1. Lattice Model

As the resolution of the model should depend on the smallest important length scale of the system [4], we employ a lattice model and Monte Carlo method in order to be able to observe frequent transitions between monomeric and aggregated states. The HP model [22] has been used to represent the protein, extended by the incorporation of an amphipathic bead type (H2) to supplement the hydrophobic (H) and polar (P) beads (Figure 1). The H2 beads are used mainly to model the amphipathic helix in TatA, while the TMH is hydrophobic (H) and the loop and tail regions hydrophilic (P).

Membrane: while the transmembrane helix (TMH) spans the membrane normal, the amphipathic helix (APH) lies in parallel with the membrane surface. As a result, we have modelled the membrane through fixed spatial zones defining membrane, water and interfacial regions. As explained below (*The Hamiltonian*), the different zones give rise to different one- and two-particle energy terms , thus allowing us to capture both hydrophilic/hydrophobic solvation and membrane-mediated interactions within a continuum membrane model.

The Hamiltonian: full details of the Hamiltonian are provided in Appendix, and a summary presented here for convenience. Three types of interactions are considered: intra-polymer, inter-polymer, and the implicit interactions associated with being in the membrane, water or interface environments. The total potential energy is the sum of the energies defined by these three types of interactions, i.e. $E = E_{\mathrm{intra}} + E_{\mathrm{inter}} + E_{\mathrm{im}}$. Both intra-polymer and implicit interactions are defined as the sum of terms corresponding to each individual polymer, the inter-polymer interaction is the sum of interactions between all polymer pairs. In other words,

$$E_{\mathrm{intra}} = \sum_k E^k_{\mathrm{intra}}, \qquad E_{\mathrm{im}} = \sum_k E^k_{\mathrm{im}},$$

and

$$E_{\mathrm{inter}} = \sum_{s<t} E^{s,t}_{\mathrm{inter}},$$

where k, s and t index polymers.

The implicit term for each polymer is the sum of one-particle energies for the beads in the molecule, representing the interaction of the bead with its immediate environment. These are defined such that hydrophobic beads are favoured in membrane, polar beads are favoured in water and amphipathic beads are favoured in interfaces.

Both E_{intra} and E_{inter} are sums over pairs of beads, differing only in whether the beads belong to the same (intra) or different (inter) proteins. Both the type (H or P) and the environment (membrane or water) determine the strength of the interaction in a way that mimics a good solvent model. Thus H-H interactions are the same as H-membrane interactions, while P-P interactions match those of a P bead with water. The net effect is that an H-H attraction is present in water, and P-P attraction in the membrane. The strength of the interaction is then scaled linearly with distance, becoming zero beyond a specified cutoff (2 lattice sites).

In order to generate secondary structure, we have also incorporated hydrogen bond interactions between beads displaced along the same chain and located within either the TMH or APH segments. These serve to stabilize helical configurations in these segments. Since our focus is on association/dissociation of the chains, and not on helix formation itself, the hydrogen bonding energies were sufficiently strong to ensure the helices—if present—were stable at all temperatures studied. A switch was incorporated into the Hamiltonian to allow us to include or exclude either or both of these helices in any given simulation.

2.2. Monte Carlo Method

Since our model exhibits a near-first-order phase transition, we used multicanonical sampling (MUCA) [12], following a parallel tempering (PT) simulation [9, 10]. The PT part produced estimates of the weights that were subsequently fed into the MUCA simulation. To estimate these weights efficiently, the logarithm of the density of states was expressed in terms of free energy differences and the MBAR estimator [13] was used to best estimate these quantities [14]. We applied MBAR to the full dataset of the PT simulation and averaged over the resulting estimates of the density of states obtained from each temperature trajectory. For most simulations, using the weights estimated from the PT stage led to a suitably enhanced sampling of conformation space, hence MUCA recursions [23] were not required.

In general, methods that bias sampling in a generalized ensemble involve generating configurations with *a priori* unknown weights, which are refined iteratively as more data are accumulated. Examples include the multicanonical sampling, the $1/k$-ensemble method [24], Wang-Landau sampling, and transition matrix Monte Carlo [25]. The initial weights are typically set to be equal to one, indicating an infinite-temperature setup of the system, and proposal moves leading to unvisited states are definitely accepted to enable exploration of the whole energy spectrum. In some cases, there could be physical and geometric constraints on the systems of interest and it may be more convenient to just focus on a subset of the energy spectrum. One reason that we use PT as the first stage of our simulation instead of directly applying MUCA is because we wish both to preserve specified helical structure in TatA and to make sure that the molecule resides in the membrane during the course of the simulation. Since both the medium and structure of the molecule are encoded through our force field, an infinite-temperature simulation would inevitably sample the entire energy spectrum and explore part of phase space which may not be interesting to us. On the other hand, a PT simulation with suitably chosen temperature ladder allows us to monitor, e.g. helicity, in the simulated temperature range.

Pull moves [26] and translation moves are used as trial moves, although biased moves such as used in the configuration bias Monte Carlo [27] are also desirable. A translation move just displaces the entire chain by a random number of lattice sites. Specifically, one of the four directions corresponding to $\pm x, \pm y$ is chosen, and the chain is shifted with an amount uniformly chosen between 1 and 10 lattice sites. Pull moves start by creating a square in the chain and successively pulling the beads along until an existing square is undone or until the terminal bead is reached when there is no such square along the path. The chain can be pulled in either direction. One feature of the pull move is that it stops early whenever possible, resulting in a higher acceptance probability compared with moves that displace many beads in a configuration. The correction noted in [28] was taken into account to ensure reversibility and ergodicity of pull moves.

3. Results A: Membrane with two protein molecules

We first examined the dispersed/condensed equilibrium in a membrane containing just two protein molecules. To consider how the secondary structure affects this equilibrium, simulations were conducted with all possible combinations of the TatA secondary structure present: no helices, only the TMH, only the APH, or both. The same amino acid sequence was used in each case.

3.1. Case 1: no secondary structure—both TMH and APH are absent

Here, and in other cases, we monitor six observables of interest: the heat capacity (C_v), the inter-polymer contribution to total potential energy (U_{inter}), the number of inter-polymer tail-tail contacts (N_{tt}), the number of inter-polymer tail-loop contacts (N_{tl}), the number of inter-polymer contacts ($N_{\text{inter}} = N_{\text{tt}} + N_{\text{tl}}$), and the number of intra-polymer P-P contacts per chain (N_{intra}). Because of the way the force field is defined, the P-P contacts in N_{intra} must happen within the membrane, so there is a competition within the membrane between intra- and inter-polymer P-P contacts. Results for these observables in the simulated temperature range are shown in Figure 2.

The heat capacity shows two distinct states with a transition temperature centred at $T = 1.2$, but spanning 1.0 to 1.4. Also, the inter-polymer energy tends towards 0 as T increases, with almost no interactions between the two polymers when $T > 1.4$. Hence we see that a dispersed state dominates high temperatures ($T > 1.2$) and the dimer state dominates low temperatures ($T < 1.2$). All other quantities, i.e. the various contact numbers, show the same trend as the temperature varies: the number of contacts decreases as T increases.

In addition, comparing N_{inter} with N_{intra}, we see that when $T < 1.2$, more inter-polymer contacts were observed on average than the number of intra-polymer contacts. In particular, at $T = 0.8$, N_{inter} is about 5 contacts larger than N_{intra}. The distribution of the pair ($N_{\text{intra}},\ N_{\text{inter}}$) was examined more closely at the lowest temperature. The 10 most frequent combinations of N_{inter} and N_{intra} were identified in each of the 10 MUCA runs, and their probabilities are presented in (Figure 3). It is observed that some pairs occur repeatedly in many runs, such as $(11, 16)$, $(12, 16)$, $(12, 15)$ and $(13, 15)$; however, none of them has probability above 0.03, suggesting that there is not a favourable combination of intra- and inter-polymer contacts. In addition, N_{inter} is no less than N_{intra} in most cases. Similar results were observed with the other secondary structure combinations (Cases 2–4).

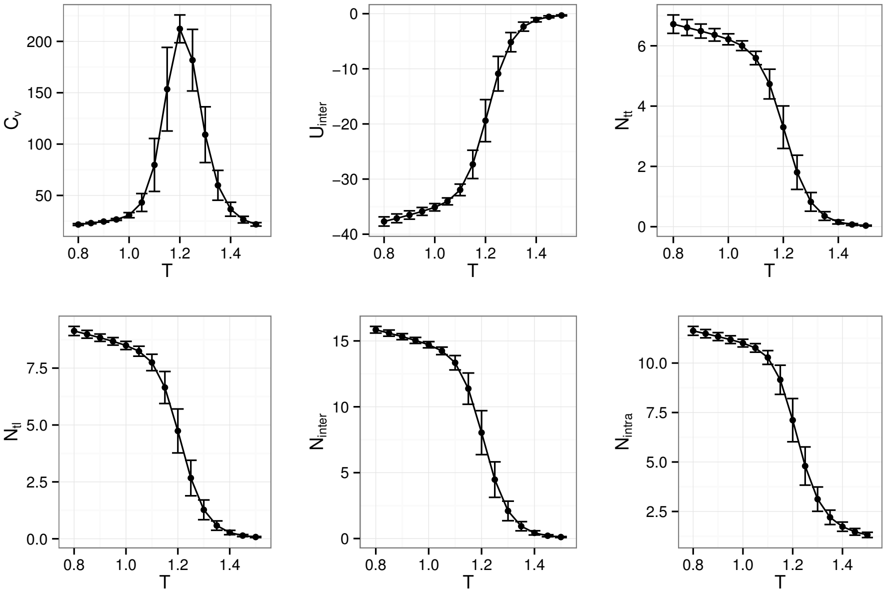


Figure 2: Property estimates for case 1 (no helices). The six properties are the heat capacity (C_v), the inter-polymer contribution to total potential energy (U_{inter}), the number of inter-polymer tail-tail contacts (N_{tt}), the number of inter-polymer tail-loop contacts (N_{tl}), the number of inter-polymer contacts (N_{inter}) and the number of intra-polymer P-P contacts per chain (N_{intra}). The error bar was calculated as one standard deviation of the estimates over 10 independent MUCA runs, each of which has starting configuration chosen from either aggregated or dispersed state.

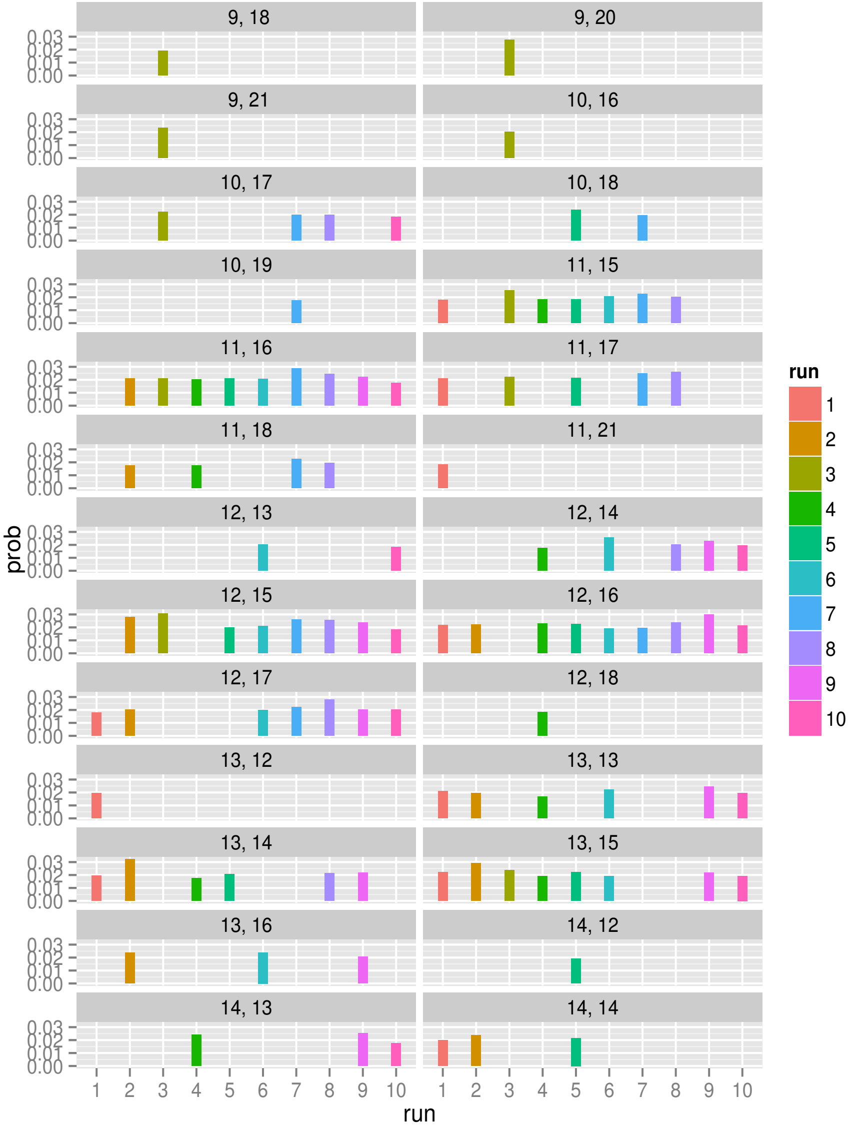


Figure 3: Probability of the pair $(N_{\text{intra}},\ N_{\text{inter}})$ taking on various values across all runs. For each run, the 10 most frequent appearing pairs were extracted; there are in total 26 pairs of contacts in Case 1 (no helices), since many overlaps were observed. The first number on the top of each subplot is N_{intra}, and the second number is N_{inter}.

3.2. Case 2: only TMH is present

Now we switch to Case 2, where the transmembrane helix is present but the amphipathic helix is not. The estimated properties are shown in Figure 4, from which we see that observations we made about the various contacts in Case 1 hold also in Case 2. The transition temperature is around or below $T = 1.15$, since an overlap with the adjacent error bar at $T = 1.1$ was observed. Our results for this case therefore indicate that dimerization occurs at a slightly lower temperature when the transmembrane helix is present.

3.3. Case 3: only APH is present

In the presence of just the APH there is a very clear shift in the equilibrium in favour of dimerization. The temperature range for the simulations was chosen to match that for all other systems (up to $T = 1.6$), and so did not probe high enough temperatures to explore the fully dissociated monomer state. Nevertheless, the properties at $T = 1.6$ are dominated by the monomer state, even though the average inter-polymer energy is negative and the average number of inter-polymer contacts is non-zero, which suggest that some dimers are still present at this temperature. Further, the data show a clear transition temperature at $T = 1.5$, substantially higher than was observed with either no helical structure or just the TMH present.

3.4. Case 4: both TMH and APH are present

The results when both helices are present reveal a balance between a weak destabilisation from the TMH and a stronger stabilisation from the APH. The transition temperature between monomer and dimer occurs near $T = 1.35$, very much midway between cases 1 (no secondary structure) and 3 (only APH). The dispersed state is again not fully formed by $T = 1.6$, but in this case the energy and the number of inter-polymer contacts are both close to zero at that temperature. At $T = 1.4$, there are about 6 inter-polymer contacts; whereas at the same temperature in Case 3, there are still more than 10 inter-polymer

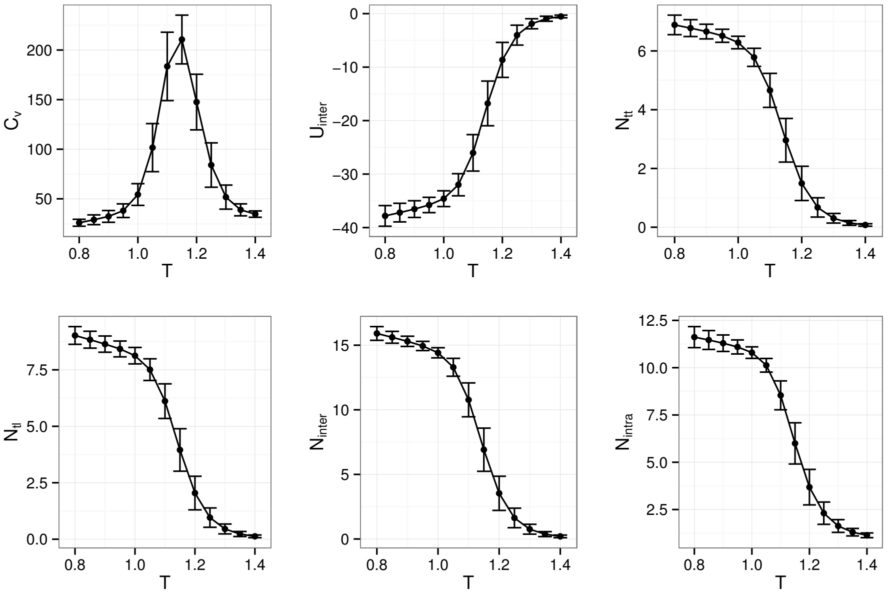


Figure 4: Property estimates for case 2 (only TMH).

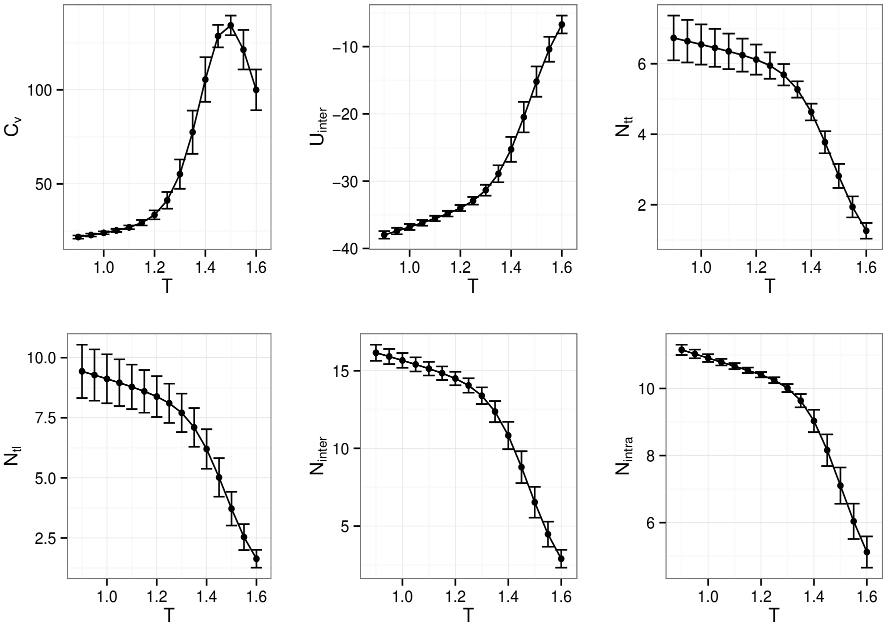


Figure 5: Property estimates for case 3 (only APH).

	σ^2_{tt}	σ^2_{tl}	$\sigma^2_{\mathrm{inter}}$	$\sigma_{\mathrm{tt,tl}}$
Case 1	0.0934	0.0402	0.0635	-0.0351
Case 2	0.1106	0.1542	0.2830	0.0091
Case 3	0.4033	1.2347	0.2697	-0.6842
Case 4	0.4531	1.0330	1.0479	-0.2191

Table 1: Estimated variances of N_{tt}, N_{tl} and N_{inter} at the lowest temperature in each of the four cases. Covariances (last column) have been calculated according to (1).

contacts, and there are almost no inter-polymer contacts at this temperature in Cases 1 and 2.

From all four cases, we can compute estimates of the variances of the contacts. These are listed in Table 1.

Given the equation of variance of a sum

$$\mathrm{Var}(N_{\mathrm{inter}}) = \mathrm{Var}(N_{\mathrm{tt}}) + \mathrm{Var}(N_{\mathrm{tl}}) + 2\mathrm{Cov}(N_{\mathrm{tt}}, N_{\mathrm{tl}}), \tag{1}$$

and substituting into this equation the estimated values in Table 1, we see that except for the TMH case (Case 2), N_{tt} and N_{tl} are negatively correlated ($\mathrm{Cov}(N_{\mathrm{tt}}, N_{\mathrm{tl}}) < 0$) when temperature is low, and this negative correlation is particularly strong in the APH case (Case 3) compared to other cases. This observation implies that when APH is present, the chains "trade" tail-tail contacts for tail-loop contacts when they form a dimer; this can be explained by the fact that the APH creates a barrier for the tail of one chain to reach to the loop of the other chain. It is unclear though why the TMH tends to decorrelate the tail-tail and tail-loop contacts.

Recapitulating all four cases, our results suggest that the amphipathic helix tends to favour dimerization, since a higher temperature is needed to disrupt it; and while the transmembrane helix hinders dimerization (lower transition temperature T_{trans}), the amphipathic helix appears to give the stronger effect, generating a bigger shift in T_{trans} in Case 3 compared with Case 2, and resulting in an increase in T_{trans} when both helices are present.

We also observed that, for Cases 2 and 4, at least 94.7% of transmembrane helical contacts were maintained across the entire simulated temperature range; for Cases 3 and 4, at least 96.3% of amphipathic helical contacts were maintained; for Cases 1 and 3, at least 99.5% of hydrophobic (H) beads were within the membrane region and for Cases 1 and 2, at least 97.3% of amphipathic (H2) beads were within the interface region. These values confirm that the differences in dimer stability were directly related to the differences in the nature of the secondary structure, and not an indirect effect of changes in solubility or helical content with temperature.

4. Results B: four-chain models

The TatA transport channel can adopt variable sizes during substrate translocation, hence it is interesting as well as desirable to study properties of multi-chain systems. As a straightforward extension to Section 3, we doubled the number of proteins while keeping the concentration fixed. In this section we consider the results for the aggregation of four identical protein chains in the membrane. This generates a more complex system than the two-protein systems, as partial aggregates (dimers and trimers) are possible in addition to the fully aggregated (tetramer) and fully dispersed (monomer) states. We have again studied different combinations of TatA secondary structure, looking specifically at zero or one helix (cases 1–3 of Section 3). As will be discussed in Section 5, the dual helix system (case 4) is best left until a more efficient trial move has been identified to use in place of, or in conjunction with, pull moves.

All simulations have again been conducted using MBAR-enhanced Monte Carlo. While 10 independent MUCA runs were performed for each of the three cases, we found that for the TMH case, not all of the ten runs sampled the entire energy space adequately. Specifically, three of the TMH runs generated pull move acceptance ratios of just 0.092, 0.116 and 0.125, while the rest range from 0.16 to 0.19; the data presented in the rest of this paper, including pull move acceptance ratios reported in Table 5, Appendix, have therefore been calculated using just the seven higher acceptance ratio runs for the four-chain TMH system. We suspect that it is more likely to be trapped in entanglements with regular transmembrane helix than with an “amorphous” TMH. The reason for this might be due to the following observation from the tetramer snapshots shown in Figure 7: In Case 2 (Figure 7b), the TMH beads (red) maintain helical shape because of the strong interaction encoded in the helices, the hydrophilic beads (blue) comprising the loop and tail regions (Figure 1) form a compact configuration surrounded by TMH. It is possible that the tetramer becomes trapped in certain collections of such configurations and the only way to get out of it is by deforming one or several helices; on the other hand, when there is no helical interaction, the TMH beads can move more freely (Figure 7a), and the different arrangements of the TMH beads may open pathways to escape from entanglements.

Excluding the 3 runs with the lowest acceptance rates, we plot the heat capacity of the TMH case with error bars in Figure 8, along with the other two cases, i.e. no helices and only APH, for both of which all 10 runs were used to calculate the error bars.

Notice first that a single peak is observed in heat capacity, suggesting that the system is still a two-state system even though partial aggregates can occur. The transition temperatures (T_{trans}) for Cases 1–3 are 1.355, 1.296 and 1.638, respectively; these results are consistent with those for the two protein systems presented in Section 3, namely, that the amphipathic helix (APH) favours aggregation whereas the transmembrane helix (TMH) hinders it, with the former effect being stronger.

To find out, quantitatively, what contributes to the change in heat capacity, we have calculated the fraction of molecules in various states as a function of temperature; these are shown in Figure 9. It can be seen that both the temperature where both tetramer and dispersed states are equally likely, and the temperature where the fraction of partially aggregated state is maximum, are nearly coincident with the transition temperature.

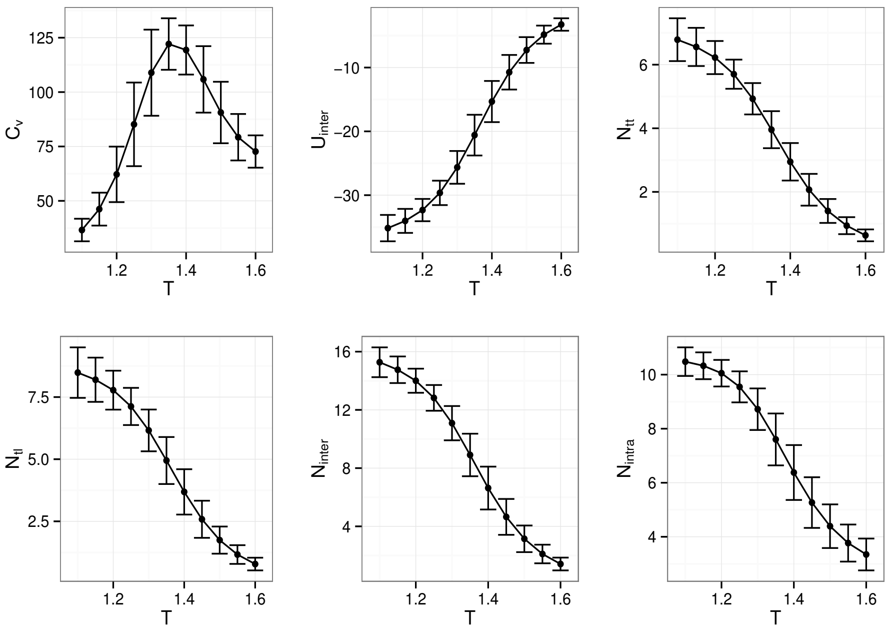
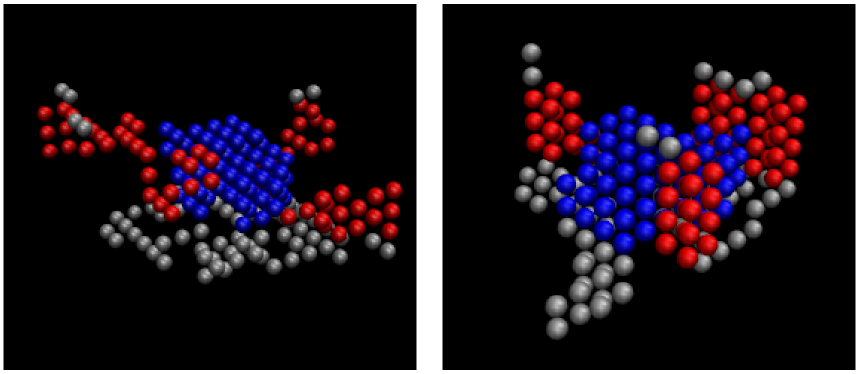


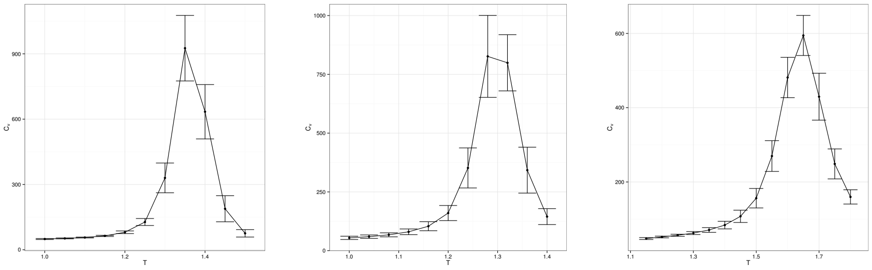
Figure 6: Property estimates for case 4 (both TMH and APH).



(a) no helices

(b) with TMH

Figure 7: Tetramer snapshot for Case 1 (a) and Case 2 (b).

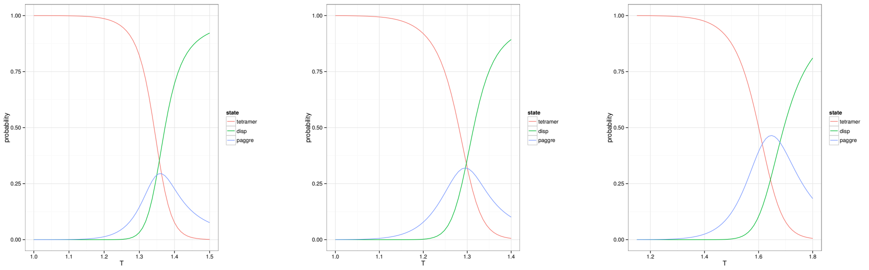


(a) no helices

(b) only TMH

(c) only APH

Figure 8: Heat capacity for three cases with error bars. In plotting Figure 8b, the three runs with the lowest acceptance rates were excluded. The half length of the error bar equals one standard deviation.



(a) no helices

(b) only TMH

(c) only APH

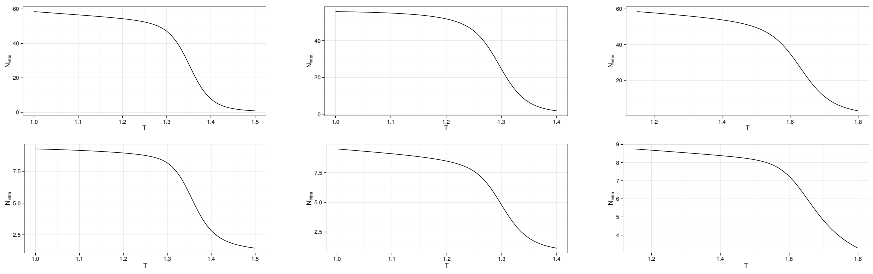
Figure 9: Fractions of tetramer (red), dispersed (green) and partially aggregated (blue) states across temperatures. All three cases are shown—(a): no helices, (b): only transmembrane helix and (c): only amphipathic helix.

	T_{trans}	fraction		
		tetramer	dispersed	par. aggre.
Case 1 (no helices)	1.355	0.375	0.331	0.294
Case 2 (only TMH)	1.296	0.358	0.321	0.320
Case 3 (only APH)	1.638	0.298	0.240	0.462

Table 2: Fractions of tetramer, dispersed and partially aggregated states at the transition temperature for the three cases.

The values of the fractions at the transition temperature are shown in Table 2. It can be seen that the partial aggregates are the most prevalent state when only the APH is present. In contrast, they are the least prevalent for the other two systems (no helix, or just TMH), although the frequency of the partially aggregated and dispersed states are very similar in the presence of the TMH. It can also be seen that presence of either helix increases the population of partial aggregates, and decreases the population of tetramer, with the APH again generating a stronger effect than the TMH.

Next, we show inter- and intra-polymer contacts as functions of temperature, as we have done in the analysis of two-chain models. In Section 3, we decomposed inter-polymer contact into tail-tail and tail-loop contributions; here in four-chain case, we shall consider only the total number of inter-polymer contacts, defined as the sum of contacts from all pairs of polymers. As shown in Figure 10, inter-polymer contacts (N_{inter}) clearly dominate intra-polymer contacts (N_{intra}) for all three cases. This contrasts with what we observed in the two-chain models, where N_{inter} was only marginally greater than N_{intra}. Table 3 shows average values of N_{inter} and N_{intra} at the lowest temperature in the respective cases.



(a) Case 1: no helices (b) Case 2: only TMH (c) Case 3: only APH

Figure 10: Inter-polymer (N_{inter}) and intra-polymer (N_{intra}) contacts as functions of temperature. All three cases are shown—(a): no helices, (b): only transmembrane helix and (c): only amphipathic helix.

	T	N_{inter}	N_{intra}
Case 1	1	58.5	9.3
Case 2	1	55.8	9.5
Case 3	1.15	58.5	8.8

Table 3: Average values of inter- and intra-polymer contacts at the lowest temperature in each case.

5. Discussion

The results presented above clearly show that aggregation of the proteins is strongly favoured by the amphipathic helix. Yet, at the same time, the amphipathic helix tends to lead to smaller, partial aggregates rather than larger, complete aggregates. This latter point can be understood in terms of the limited volume available at the interface, which will make it difficult to form the strongly ordered APH arrays that would be needed in much larger aggregates. While very clearly defined within our continuum membrane model, such volume constraints on the APH are also likely to be found within a fully atomistic model of the protein membrane system. Thus these observations are relevant for the mechanism for pore formation during Tat translocation. The most widely accepted conjecture for the structure of a Tat pore is one in which the APH of TatA folds up into the membrane to become another transmembrane helix and forming a U-shape with the TMH; many of these TatAs are assembled to form a double wall to a cylindrical pore, with the TMH forming the outer wall and the APH forming the inner wall [15]. Our model shows that the APH can, indeed, drive aggregation, but does so with an inefficient packing that is unlikely to support the spontaneous and coordinated upward folding of the TatA APH on a sufficient scale for pore formation. It could, however, create aggregates of APH that were available to interact with a triggering event—perhaps such as the Tat signal sequence interacting with a TatBC complex.

More and larger simulations are needed to determine the relative stability of different possible TatA aggregates. The results presented in this work show that methods such as MBAR-enhanced MC [14] make it viable to model larger aggregates—at least when up to one helix is present. However our simulations showed that the acceptance rate for pull moves decreased substantially with the number of helices in each protein, and became unacceptably low for the double helix with four molecules (see Appendix). The problem lies in the large energy penalty associated with disrupting the helical structure, which means any proposal move which displaces beads out from the helix is likely to be rejected. The pull move starts by randomly choosing a "pull bead" and then using it to pull one end of the molecule through two lattice spacings. This may result in moving any helix it encounters along the protein which, depending on amino acid sequence, is likely to generate a large energy penalty. Thus any pull move which selects a pull bead within a helix, or which pulls an end of the molecule that contains a helix, is likely to be rejected. The effect is manageable with a single helix, but leads to a low sampling efficiency for the loop region when both the TMH and APH are present.

The situation can be alleviated by incorporating a more efficient move set that recognizes, or learns from, the underlying potential. One could, of course, manually encode the sequences of helix blocks into the program and change the trial moves accordingly, but it would be hard to generalize. Even better would be to design moves that can adapt to any given potential. Several existing methods are manifestations of this idea, such as configuration bias Monte Carlo (CBMC) [27], hybrid Monte Carlo [29], and a more recent MC scheme called fragment regrowth via energy-guided sequential sampling (FRESS) [30]. The last method is especially appealing to lattice MC simulations, as it is originally implemented to search for the global minimum energy conformation in hydrophobic-polar protein folding models. The FRESS method resembles CBMC in that regrowing the chain is also involved, as the name suggests. However, it differs from CBMC in two respects: 1) more often an internal segment is regrown, instead of regrowing the chain all the way up to the terminal residue each time; and 2) the segment to be regrown has variable length. These two features equip FRESS with the capability both to explore configurations that are local and to carry out more global moves, which allows the algorithm to jump out of local energy basins. This last point would be particularly useful in exploring the thermodynamic stability of the proposed Tat pore structure.

The MBAR-enhanced Monte Carlo can be used in both on- and off-lattice models, and while we used parallel tempering and multicanonical sampling methods, other choices are available. For example, instead of running MUCA simulations and reweighting the data to obtain property estimates, we could use methods like the Wang-Landau algorithm to further refine the density of states, using those estimates derived from PT and MBAR as a guide.

6. Conclusions

We have constructed a simplified lattice model to study the effect of helices on protein aggregation in membranes. The model is motivated by TatA aggregation in the twin-arginine translocation pathway; TatA is a membrane protein that contains both a transmembrane helix (TMH) and an amphipathic helix (APH). A recently proposed Monte Carlo approach, which exploits the statistical optimality of the MBAR estimator, was used [14]. We observed that for both two- and four-chain models, the TMH tends to disfavor aggregation whereas the APH favors it. The effect of the APH is stronger, and leads to the formation of partial aggregates in a manner that is consistent with the current mechanistic understanding of Tat pore formation. However, a better move set is needed for further improvement of simulation efficiency, since pull moves do not enable adequate sampling of the loop region between the two helices. Nevertheless, we have shown that MBAR-enhanced Monte Carlo is a promising approach for the study of complex, multiscale processes like protein aggregation. Our work also serves to motivate more computational research in TatA aggregation, or the Tat process in general, that calls for computer simulations in addition to experiments to elucidate its mechanism.

Acknowledgement

Computer time for this work was provided by the Scientific Computing Research Technology Platform at Warwick University, and by the MidPlus Regional e-Infrastructure Centre under EPSRC grant ep-K000233. YX would also like to acknowledge partial funding from the Warwick Centre for Scientific Computing.

Appendix

The following list of notations are introduced:

1. n_k: the number of beads in polymer k.
2. x_i: the coordinate of bead i in a polymer. It defines a point in 3-D space.
3. c_i: the color of bead i in a polymer. When there is no interface, it is either hydrophobic (H) or hydrophilic (P); when there is interface, it is one of H, P or H2.
4. d_{ij}: the Euclidean distance between beads i and j.
5. ϵ_{ij}: the strength of interaction between beads i and j, as a linear function of d_{ij},
$$\epsilon_{ij} = \begin{cases} \frac{\epsilon_{\text{max}}(d_{ij}-d_{\text{cut}})}{1-d_{\text{cut}}}, & 1 \leq d_{ij} < d_{\text{cut}} \\ 0, & d_{ij} \geq d_{\text{cut}} \end{cases}$$
where d_{cut} is the cut-off distance and ϵ_{max} defines the maximum interaction when $d_{ij} = 1$.
6. M: the membrane region, $M = \{(x, y, z) \in \mathbf{Z}^3 : 0 < z < h_{\text{memb}}\}$, where $\mathbf{Z}^3$ denotes set of all 3-D integer point and h_{memb} defines the height of membrane.
7. C: the lower (cytoplasmic) side of the interface.
$C = \{(x, y, z) \in \mathbf{Z}^3 : -h_{\text{inter}} \leq z \leq 0\}$, where h_{inter} is the height of interface.
8. P: the upper (periplasmic) side of the interface.
$P = \{(x, y, z) \in \mathbf{Z}^3 : h_{\text{memb}} \leq z \leq h_{\text{memb}} + h_{\text{inter}}\}$
9. W: the water region. When there is no interface, it is everywhere else of M, i.e. $W = M^c$, the complement of set M; when there is interface, it is everywhere else of M, C and P, i.e. $W = (M \cup C \cup P)^c$.
10. δ: delta function. We define the following for membrane region and hydrophobic beads, the sets of notations $\{\delta_{\text{WW}}, \delta_{\text{W}}, \delta_{\text{pp}}, \delta_{\text{p}}\}$, $\{\delta_{\text{CC}}, \delta_{\text{PP}}, \delta_{\text{C}}, \delta_{\text{P}}, \delta_{\text{h2-h2}}, \delta_{\text{h2}}\}$ are defined similarly for water region and hydrophilic beads, and for in-

terface region and H2 beads, respectively.

$$\delta_{\mathrm{MM}}(i,j) = \begin{cases} 1, & \text{if } x_i, x_j \in M \\ 0, & \text{otherwise} \end{cases} \qquad \delta_{\mathrm{M}}(i) = \begin{cases} 1, & \text{if } x_i \in M \\ 0, & \text{otherwise} \end{cases}$$

$$\delta_{\mathrm{hh}}(i,j) = \begin{cases} 1, & \text{if } c_i = c_j = \mathrm{H} \\ 0, & \text{otherwise} \end{cases} \qquad \delta_{\mathrm{h}}(i) = \begin{cases} 1, & \text{if } c_i = \mathrm{H} \\ 0, & \text{otherwise} \end{cases}$$

It is understood and should be clear in the context that the delta functions also depends on polymer index k, s, t.

11. H_{b}: set which defines hydrogen bond condition.

$$H_{\mathrm{b}} = \{(i,j) : j - i = 5,\ j\%4 \neq 0 \text{ or } i\%4 = 0,\ (j+1)\%4 = 0,\ j - i = 3\}$$

where percent sign is the modulo operator. This requires the starting index of a sequence of beads comprising helix being a multiple of 4. The associated delta function for this set is denoted δ_{hbond}.

Now, each individual terms in the potential are given by

$$\begin{aligned} E^k_{\mathrm{intra}} = -\sum_{j-i\geq 3} \quad & \delta_{\mathrm{MM}} \quad (\delta_{\mathrm{hh}}\,\delta_{\mathrm{hbond}}\,(f_1\,\epsilon_{\mathrm{hbond}}) + \delta_{\mathrm{pp}}\,\epsilon_{ij}) + (\delta_{\mathrm{CC}} + \delta_{\mathrm{PP}})\,\delta_{\text{h2-h2}}\,\delta_{\mathrm{hbond}}(f_0\,\epsilon_{\mathrm{hbond}}) \\ + \quad & \delta_{\mathrm{WW}}(\delta_{\mathrm{hh}} + \delta_{\text{h-h2}} + \delta_{\text{h2-h2}})\epsilon_{ij}, \end{aligned}$$

$$E^k_{\mathrm{im}} = -\sum_{i=1}^{n_k} \delta_{\mathrm{M}}\,\delta_{\mathrm{h}}\,\epsilon_{\mathrm{hm}} + \delta_{\mathrm{W}}\,\delta_{\mathrm{p}}\,\epsilon_{\mathrm{pw}} + (\delta_{\mathrm{C}} + \delta_{\mathrm{P}})\delta_{\mathrm{h2}}\,\epsilon_{\mathrm{h2inf}},$$

and

$$E^{s,t}_{\mathrm{inter}} = -\sum_{i=1}^{n_s}\sum_{j=1}^{n_t} \delta_{\mathrm{MM}}\,\delta_{\mathrm{pp}}\,\epsilon_{ij} + \delta_{\mathrm{WW}}\,(\delta_{\mathrm{hh}} + \delta_{\text{h-h2}} + \delta_{\text{h2-h2}})\epsilon_{ij}.$$

The summation in E^k_{intra} is taken over all pairs of beads in polymer k that are separated by at least 3 beads apart, which is the smallest number of beads required to form a contact in a rectangular lattice. In E^k_{intra}, $\epsilon_{\mathrm{hbond}}$ sets the strength of helical contact and f_1, f_0 are factors that control the strength of the transmembrane- and the amphipathic-helix, respectively. The ϵ_{hm}, ϵ_{pw} and $\epsilon_{\mathrm{h2inf}}$ in E^k_{im} reflect, respectively, the tendency of (1) a hydrophobic bead to stay

Parameter	Value			
	case 1	case 2	case 3	case 4
ϵ_{hbond}	0	4	4	4
f_1	0	1	0	1
f_0	0	0	1	1
ϵ_{hm}	4	4	4	4
ϵ_{h2inf}	4	4	4	4
ϵ_{pw}	1.2	1.2	1.2	1.2
ϵ_{max}	1	1	1	1
d_{cut}	1.7	1.7	1.7	1.7
h_{memb}	5	5	5	5
h_{inter}	1	1	1	1

Table 4: Parameters used in the lattice polymer model.

in membrane, (2) a hydrophilic bead to stay in water, and (3) an amphipathic bead to stay in either side of the interface.

Table 4 lists the parameter values used in the lattice polymer model.

	case	temperature ladder	box size	PT iters (per temperature)	MUCA iters	pull move acceptance in MUCA	approx CPU time per 10^7 iters (PT/MUCA)
Two-chain	1	0.8 1 1.2 1.5	90	3×10^7	7×10^7	0.31	7h/1h
	2	0.8 1 1.2 1.4		3×10^7	1.9×10^8	0.20	
	3	0.9 1.1 1.35 1.6		3×10^7	1.9×10^8	0.21	
	4	1.1 1.25 1.4 1.6		5×10^7	3.7×10^8	0.09	
Four-chain	1	1 1.15 1.3 1.5	114	3×10^7	1.9×10^8	0.28	9h/2h
	2	1 1.1 1.25 1.4		5×10^7	2.5×10^8	0.18	
	3	1.15 1.3 1.55 1.8		5×10^7	2.5×10^8	0.20	

Table 5: Additional simulation information: temperature ladder, the size of the cubic simulation box with periodic boundary condition, number of parallel tempering iterations per temperature, number of multicanonical iterations, pull move acceptance rate in MUCA simulations, and approximate CPU time for 10^7 iterations in PT/MUCA. The number of attempted pull moves is 70 per 100 iterations, the remaining 30 is the number of translation move attempts

References

[1] C. M. Dobson, Protein folding and misfolding, Nature 426 (6968) (2003) 884–890.

[2] F. Chiti, C. M. Dobson, Protein misfolding, functional amyloid, and human disease, Annu. Rev. Biochem. 75 (2006) 333–366.

[3] T. P. Knowles, M. Vendruscolo, C. M. Dobson, The amyloid state and its association with protein misfolding diseases, Nat. Rev. Mol. Cell Biol. 15 (6) (2014) 384–396.

[4] A. Morriss-Andrews, J.-E. Shea, Computational studies of protein aggregation: Methods and applications, Annu. Rev. Phys. Chem. 66 (2015) 643–666.

[5] B. C. Berks, The twin-arginine protein translocation pathway, Annu. Rev. Biochem. 84 (2015) 843–864.

[6] U. Gohlke, L. Pullan, C. McDevitt, I. Porcelli, E. de Leeuw, T. Palmer, H. Saibil, B. Berks, The tata component of the twin-arginine protein transport system forms channel complexes of variable diameter, Proc. Natl. Acad. Sci. 102 (30) (2005) 10482–10486".

[7] Y. Sugita, Y. Okamoto, Replica-exchange molecular dynamics method for protein folding, Chem. Phys. Lett. 314 (1999) 141–151.

[8] A. Laio, M. Parrinello, Escaping free-energy minima, Proc. Natl. Acad. Sci. 99 (20) (2002) 1256212566.

[9] C. J. Geyer, Markov chain monte carlo maximum likelihood, in: E. M. Keramidas (Ed.), In Computer Science and Statistics: Proceedings of the 23rd Symposium on the Interface; Keramidas, E. M., Kaufman, S. M., Eds.; Interface Foundation of North America: Fairfax Station, VA, 1991; pp 156-163., Interface Foundation of North America, 1991, pp. 156–163.

[10] K. Hukushima, K. Nemoto, Exchange monte carlo method and application to spin glass simulations, J. Phys. Soc. Jpn. 65 (6) (1996) 1604–1608.

[11] F. Wang, D. P. Landau, Efficient, multiple-range random walk algorithm to calculate the density of states, Phys. Rev. Lett. 86 (10) (2001) 2050.

[12] B. A. Berg, T. Neuhaus, Multicanonical algorithms for first order phase transitions, Phys. Lett. B 267 (2) (1991) 249–253.

[13] M. R. Shirts, J. D. Chodera, Statistically optimal analysis of samples from multiple equilibrium states, J. Chem. Phys. 129 (12) (2008) 124105.

[14] Y. Xu, P. M. Rodger, Improved estimation of density of states for Monte Carlo sampling via mbar, J. Chem. Theory Comput.doi:10.1021/acs.jctc.5b00189.

[15] F. Rodriguez, S. L. Rouse, C. E. Tait, J. Harmer, A. De Riso, C. R. Timmel, M. S. P. Sansom, B. C. Berks, J. R. Schnell, Structural model for the protein-translocating element of the twin-arginine transport system, Proc. Natl. Acad. Sci. 110 (12) (2013) E1092–E1101.

[16] P. Lee, D. Tullman-Ercek, G. Georgiou, The bacterial twin-arginine translocation pathway, Annu. Rev. Microbiol 60:373-395.

[17] Y. Hu, E. Zhao, H. Li, B. Xia, C. Jin, Solution nmr structure of the tata component of the twin-arginine protein transport system from gram-positive bacterium bacillus subtilis, J. Am. Chem. Soc. 132 (45) (2010) 15942–15944.

[18] F. Alcock, M. A. B. Baker, N. P. Greene, T. Palmer, M. I. Wallace, B. C. Berks, Live cell imaging shows reversible assembly of the tata component of the twin-arginine protein transport system, Proc. Natl. Acad. Sci. 110 (38) (2013) E3650–E3659.

[19] B. C. Berks, F. Sargent, T. Palmer, The tat protein export pathway, Mol. Microbiol. 35 (2) (2000) 260–274.

[20] T. Palmer, B. C. Berks, The twin-arginine translocation (tat) protein export pathway, Nat. Rev. Microbiol. 10 (7) (2012) 483–496.

[21] T. Brüser, C. Sanders, An alternative model of the twin arginine translocation system, Microbiol. Res. 158 (1) (2003) 7–17.

[22] K. F. Lau, K. A. Dill, A lattice statistical mechanics model of the conformational and sequence spaces of proteins, Macromolecules 22 (10) (1989) 3986–3997.

[23] B. A. Berg, Algorithmic aspects of multicanonical simulations, Nucl. Phys. B (Proc. Suppl.) 63 (1) (1998) 982–984.

[24] B. Hesselbo, R. B. Stinchcombe, Monte carlo simulation and global optimization without parameters, Phys. Rev. Lett. 74 (12) (1995) 2151–2155.

[25] J.-S. Wang, R. H. Swendsen, Transition matrix monte carlo method, Journal of statistical physics 106 (1-2) (2002) 245–285.

[26] N. Lesh, M. Mitzenmacher, S. Whitesides, A complete and effective move set for simplified protein folding, in: In Proceedings of the 7th Annual International Conference on Research in Computational Molecular Biology, Berlin, Germany, April 10-13, 2003; ACM Press: New York, 2003., 2003, pp. 188–195. doi:10.1145/640075.640099.
URL http://doi.acm.org/10.1145/640075.640099

[27] J. I. Siepmann, D. Frenkel, Configurational bias monte carlo: a new sampling scheme for flexible chains, Mol. Phys. 75 (1) (1992) 59–70.

[28] D. Györffy, P. Závodszky, A. Szilágyi, "pull moves" for rectangular lattice polymer models are not fully reversible, IEEE/ACM Transactions on Computational Biology and Bioinformatics 9 (6) (2012) 1847–1849.

[29] S. Duane, A. D. Kennedy, B. J. Pendleton, D. Roweth, Hybrid monte carlo, Phys. Lett. B 195 (2) (1987) 216–222.

[30] J. Zhang, S. C. Kou, J. S. Liu, Biopolymer structure simulation and optimization via fragment regrowth monte carlo, J. Chem. Phys. 126 (22).