Superfluous Stability in Lattice Protein Simulations

May 2011 Matthew Glassman 8.592 Final Project

Naturally occurring proteins are generally found to be marginally stable—that is, they exist with the minimum amount of stability required to perform their function under typical environmental conditions. However, studies on protein evolution in the laboratory setting and observations of rapidly evolving viral populations have provided evidence that stability in excess of the minimum required to fold can aid in the evolution of enzymatic function or the emergence of drug-resistance. Capturing this phenomenon in a computational framework can aid in further study of its relevance and generality. Here, a simple lattice protein model is implemented and the folding of a 27-mer and 36-mer into two motifs is simulated via dynamic Monte Carlo. Efforts are made to examine the minimal number of stabilizing interactions necessary to retain the native folds in each case and thus identify protein sequences that fold near the threshold of stability. Destabilizing interactions are introduced into the fully- and minimally- stabilized structures, and the tolerance of each is investigated. Ultimately, the expected conclusion—that destabilizing interactions can be more readily tolerated in stabilized structures—cannot be reached via the efforts documented in this paper. In part, this is due to critical weaknesses in the simplified model which lead to restricted dynamics. Evidence for this limitation is discussed.

The astonishing robustness with which many natural proteins fold rapidly into unique, well-defined structures has motivated a decades-long pursuit of the mechanisms underlying this universal phenomenon in biology. An important tool in the simulation of protein folding has been the lattice model—a dramatically simplified yet computationally tractable version of the full problem. Despite the removal of many degrees of freedom and most of the chemical complexity of real systems, lattice protein simulations have made remarkable progress in our understanding of the folding process. Early work from Shakhnovich, et al [1] demonstrated that folding begins with collapse from the coil to the unstructured globule, followed by a first-order phase transition to the native state. Furthermore, the authors are able to characterize the ruggedness of the free energy landscape and show that non-folding sequences are plagued by low-energy 'traps' conformationally distant from the native structure. While far from a fully atomistic simulation, the power of this simple model lies in its ability to capture some of the essential features of the folding process, not unlike the transformative work of Flory and Huggins to model polymer solution thermodynamics.

Here, efforts have been made to build a simple lattice model and investigate the minimal energetic interactions among residues that are required to stabilize a desired folded structure. Theoretical studies based on the random energy model establish that foldable heteropolymers must have a sufficiently high energy gap such that the single native structure can compete with the entire spectrum of unfolded globular states. However, specifically how the native energy should be distributed among the possible residue interactions is not clear. In particular, it would be interesting to gain insight on the minimal number of strong interactions required to successfully fold a protein. Utilizing a minimum set of strongly interacting residues in designed sequences may help to avoid kinetic trapping, particularly for the folding of a complex structure. Furthermore, it would be interesting to investigate how the additional stability conferred by attractive interactions in excess of the minimum would play a useful role in buffering destabilizing interactions elsewhere in the structure. While natural proteins are observed to exist with the minimum stability required for folding and functioning in their natural environments [2], recent work has demonstrated the importance of stabilizing mutations in the emergence of antiviral escape mutants of oseltamivir-sensitive influenza [3,4] Investigating this effect on lattice models may provide further insight into the general nature of this phenomenon.

From the long history of lattice protein models, there are a number of different options for building and executing a simulation. In the development of Monte Carlo simulation techniques, various strategies have been employed to rapidly approach the equilibrium state. Dynamic behavior is often investigated using molecular or Brownian dynamics simulations, but an incarnation of the Monte Carlo technique has demonstrated utility as an alternate approach for studying non-equilibrium systems [5, 6]. The level of detail in these simulations can span a broad spectrum of complexity, from a cubic lattice with nearest-neighbor potentials coarse-grained at the residue level, to atomistic simulations and diamond or face-centered cubic lattices. While the detail of the simulation can be tuned to the needed resolution of the study, the manner in which the conformational space is sampled (i.e. how valid steps are chosen) must be chosen carefully to produce a physically relevant simulation. In attempting to account for exclude volume effects, a nonphysical retardation of bond relaxation can occur if the set of elementary motions is improperly defined.

Kolinski et al [7] identified a set of basic moves and properly weighted their frequency during random sampling in order to avoid the restricted dynamics observed in other models.

For this study, a simple freely-jointed chain model was constructed based on the work of Shakhanovich Folding is simulated on a cubic lattice, and the energy of the chain is computed using the following equation:

$$E = P_{ij}\sum_{ij}^{N}\Delta(r_i - r_j) + D_2\sum_{ij}^{N}\delta(r_i - r_j) + D_3\sum_{ijk}^{N}\delta(r_i - r_j)\,\delta(r_i - r_j)$$

where $\Delta(r_i - r_j)$ is 1 if $r_i - r_j = 1$, and is 0 otherwise, and $\delta(r_i - r_j)$ is the usual Kronecker delta function. P_{ij} is the matrix of residue contact energies. In this model, excluded volume interactions are handled implicitly by including the second and third terms in the above equation, accounting for two or three monomers occupying the same position. In practice, a large value is chosen for the two coefficients D_2 and D_3 to introduce a severe penalty for overlap. The authors argue that this formulation captures the greater flexibility of real systems during folding and that conformations violating monomer excluded volume do not appreciably compete with the ground state. Contact energies were chosen to be -2, 0, or -0.2 (in units of k) depending on whether the interaction between neighboring residues was to be strongly or moderately attracting.

Each step in the simulation was executed by the following process. A residue was selected at random and, if allowed by the connectivity constraints of the chain, a move was attempted to the position $r_i' = r_{i-1} + r_{i+1} - r_i$ (i.e. diagonally). The energy of the chain in the new conformation was computed using energy equation discussed previously, and a step was made according to the Metropolis criterion. Parameter choices of $D_2 = 10$, $D_3 = 14$, along with a temperature of 1 yielded simulations that largely stabilized in $10^6 - 10^7$Monte Carlo steps.

Using this model, the folding of 27 and 36-monomer chains was simulated, and each chain was folded into two motifs. Figure 1 shows the contact matrices and the simple heuristic for generating each native fold. The advantage in predefining such simple native structures is the easy interpretation of the evolution of structural patterns in the contact matrices during the simulation. However, the generality of this study could be improved by a broader or more statistical survey of folded structures accessible by these chains.

First, the folding of proteins with both native structures was simulated according to the Go model, where all native interactions contributed an energy of -2 and all non-natives an energy of 0. The extent of folding was examined by the intuitive metric of the fraction (Q) of neighboring interactions matching those of the native state. For the 27-mer protein, successful folding to both native structures was achieved, although not all trajectories reached the ground state during the course of

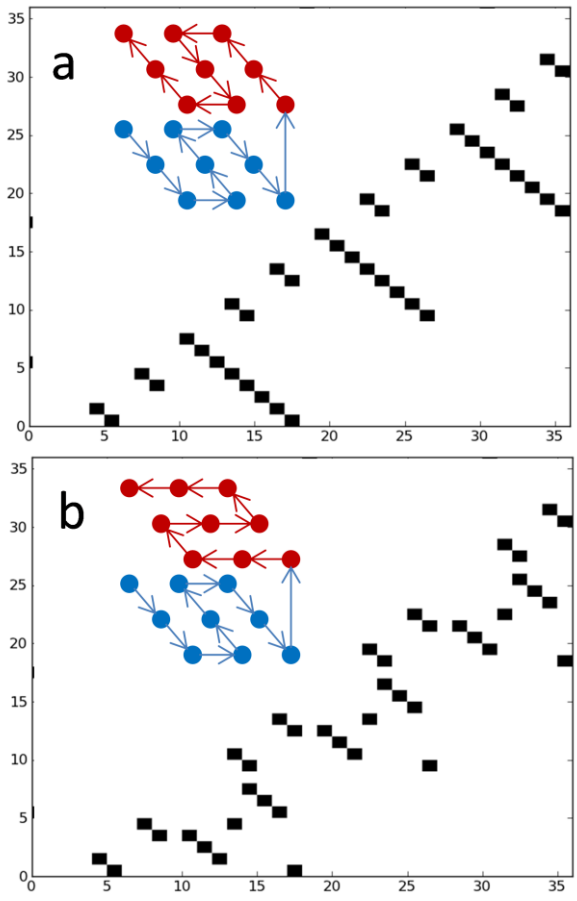

Figure 1: Residue contact matrix for fold 1 (a) and fold 2 (b) for a 36-residue lattice protein investigated in this study. Structure schematic for first two layers shown in inset

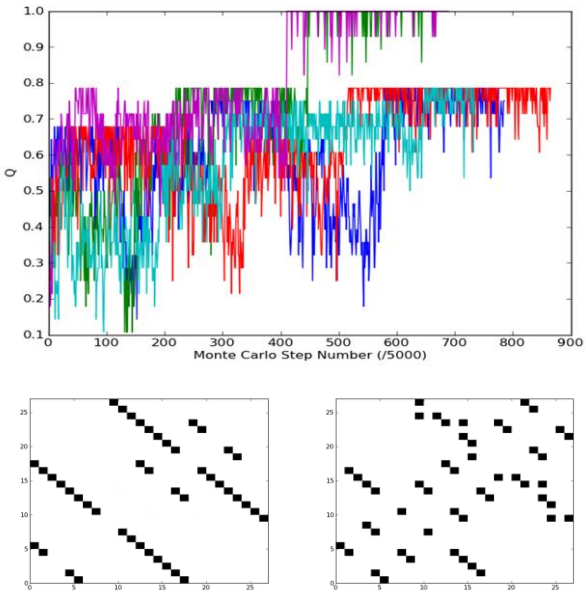

Figure 2: Trajectories of 5 simulations of the folding of the 27-mer protein into motif 1. Snapshots of the neighbor contact maps at the end of the two blue trajectories are shown below.

the simulation. Figure 2 shows typical results from the simulations. The neighbor contact maps are included for two trajectories that did not fold perfectly by the end of the runs.

Due to the simple symmetries of the structure, it is possible to analyze the contact maps to find the reason for incomplete folding. The contact map at right in Figure 2 shows signs of improper stacking of the protein layers as the long stretches of contacts are not intact. More informative is analysis of the contact map at the left of the figure. This snapshot indicates that proper stacking of the 3 layers has been achieved (the long stretches of contacts are intact), but one entire face of the protein is folded open and solvent exposed. While intuitively one might expect an open sheet of residues to snap shut relatively quickly in real systems, simulating the dynamics of closing this structure is a slow process by Monte Carlo. Due to the single elementary step available in the model implemented here, and the fact that chain flexibility can only diffuse from endpoints or turns in the chain, the crankshaft-type motion required to fold this structure is nearly prohibited. This observation hints that the model in this study exhibits the type of restricted dynamics Kolinski sought to avoid by allowing elementary motions involving more than one residue per step. Indeed, Kolinski included a crankshaft motion involving two residues that would be advantageous for this exact situation. This effect is expected to be bigger for larger proteins, and unsuccessful simulations of the folding of the 36-mer protein are likely evidence of this same issue. Figure 3 shows a contact map at the end of one trajectory for the 36-mer, where it is clear that the same open face problem is restricting the protein's ability to fold.

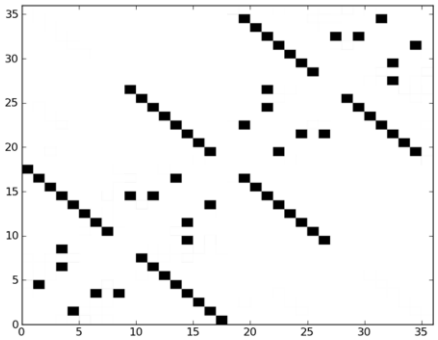

Figure 3: Contact map of the 36-mer protein with possible evidence of the pathological issue of restricted dynamics.

Despite the limitation of this model, efforts continued in order to address the primary objective, namely to observe the contribution of additional stability to mutational tolerance. To start, parameters of the model were adjusted to give all monomer pairs favorable interaction energies ($P_{ij} = -0.2$) to promote globule collapse, a move motivated by the hydrobophic effect observed in real systems. Then, a number of trials were executed whereby increasing numbers of native interactions (with $P_{ij} = -2$) were removed from the matrix of residue contact energies (set to -0.2). While not all simulations were expected to fold successfully on a single attempt, these sweeps provided a general estimate for the number of favorable interactions that could be removed and retain similar folding success. With this estimate in hand, repeated simulations were performed for different sets of 'mutated' interactions to get a better sense of where the transition exists. Here, attention will be directed at the folding of the 27-mer for both motifs because these proteins demonstrated sufficient success to be analyzed at the crude resolution of the aforementioned procedure.

Preliminary sweeps of the 27-mer through the removal of up to 20 randomly chosen native interactions indicated that between 5-7 could be removed and the chain would still adopt the native fold for motif 1, and 7-9 interactions for motif 2. Note that there are 28 total nearest-neighbor interactions in a fully compact globule of a 27-mer. Focusing on this range, 5 simulations on different sets of residue interactions were performed and the success rates determined. For motif 1, successful folding (Q = 1.0 at long times) was achieved for 40%, 80%, and 0% of the simulations for 5, 6, and 7 'mutated' interactions, respectively. The Q value for none of the 7-mutant simulations ever stably survived above 0.5 and exhibited high fluctuations (Figure 4). This observation is in contrast to the results shown in Figure 2, where most of the simulations become stably locked at high Q in a near folded state. This comparison indicates that the long-time behavior of the 7-mutants is characterized by significant sampling of non-native states, rather than delayed folding due to restricted dynamics. For motif 2, successful folding was achieved for 60%, 40%, and 0% of the simulations for 7, 8, and 9 'mutated' interactions. In two of the unsuccessful simulations for the 9-mutant,

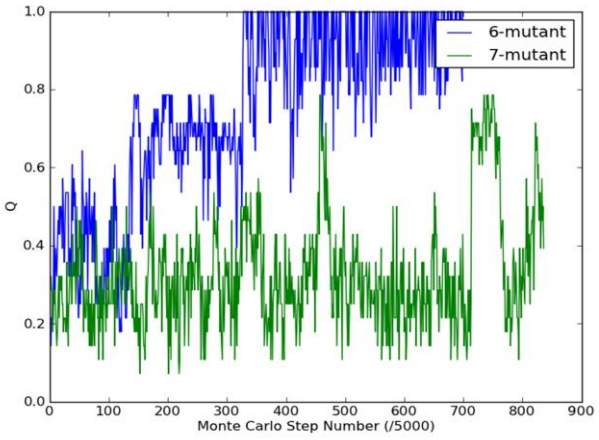

Figure 4: Characteristic folding trajectories for a 27-mer folding to motif 1 with 6 (blue) or 7 (green) nearest-neighbor interactions removed. Note that the typical 'locked' high Q-value seen in earlier restricted trajectories (Figure 2) is absent in the 7-mutant.

a metastable state was observed where the Q value reached 1.0 transiently during the simulation, but fluctuated wildly to as low as 0.35 even at long times. This result is likely indicative of the native state being in equilibrium with a spectrum of other globular states and is consistent with the 9-mutant being at the threshold of tolerable mutations.

Having identified a handful of minimally stabilized 27-mers capable of folding into either motif 1 or 2, the next step was to identify suitable destabilizing interactions to use to introduce into these backgrounds. Screens were initially performed on fully stabilized 27-mers by randomly selecting interaction pairs and setting $P_{ij} = +0.2$ to simulate a moderately destabilized interaction. Candidates were identified as those that did not interfere with the native fold, and a small number of these were then screened over an increased range of destabilizing energies in an attempt to characterize a threshold energy of disruption for the individual interaction. After determining this threshold in the stabilized structure, the same was determined in the 'minimally stabilized' protein structures.

The following results are discussed for the 27-mer adopting fold 1. In the destabilizing screen of up to 15 interactions, only one was observed to disrupt the folding process; the vast majority of simulations ran to nearly folded states. Chosen from these tolerable interactions was one which connected the single buried residue with one of the solvent-exposed face residues. A sweep of destabilizing (> 0) energies was introduced into this position in both the stabilized and a marginally-stable version of motif 1, and a slice of the largely unimpressive results are shown below in Figure 5. The particular marginally-stable shown in the results had a total of five interactions removed from its structure: four connecting surface residues and one connecting to the buried residue at the core. Even in the minimally-stabilized structure with a highly unfavorable interaction, trajectories are found which reach fully folded conformations.

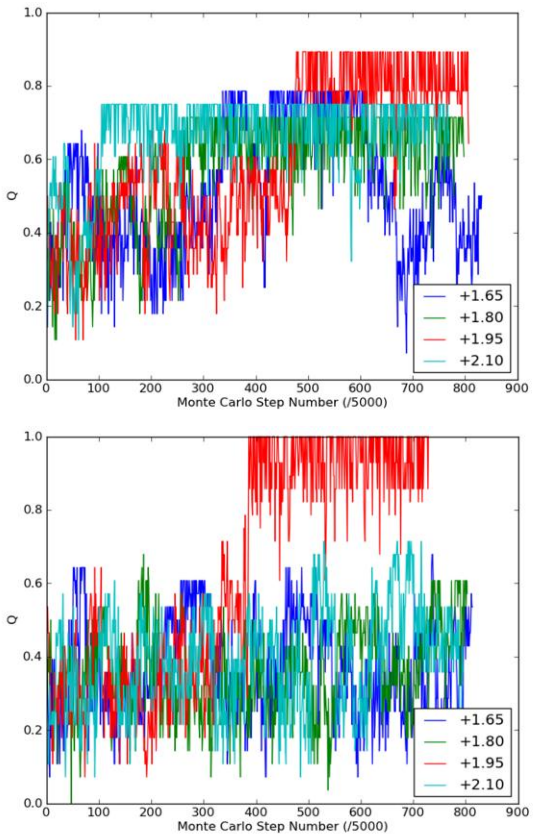

Figure 5: A slice of trajectories for the fully-stabilized (top) and minimally-stabilized (bottom) 27-mer, after introducing an unfavorable interaction with increasing instability.

The failure of this approach was likely due to a number of factors, primary of which being the poor choice of lattice model and configurational sampling strategy that led to the restricted dynamics observed in the early folding attempts. Furthermore, additional effort is needed to invest in exploring how the native structure can be destabilized by unfavorable interactions in the native structure, despite these initially depressing results on this limited scope. While the high symmetry of the chosen folding motifs limits the uniqueness of many interactions in the protein, combinations of unfavorable interactions might be fruitful in distinguishing the tolerance of stabilized structures.

In conclusion, this work discusses the implementation of a simple lattice protein model and explores folding via dynamic Monte Carlo. A simple fold (motif 1) is shown to fold with up to 6 native contacts only mildly attractive, while a slightly more complex fold (motif 2) is shown to tolerate up to 8, of a total of 28 possible interactions. Efforts to investigate the differential tolerance of stabilized versus minimally-stabilized structures to unfavorably interacting residues are embarrassingly unsuccessful, though sufficient exploration has not been undertaken to invalidate the approach pursued here.

[1] E. Shakhnovich, G. Farztdinov, A. M. Gutin and, M. Karplus, Phys. Rev. Let. **67**, 1665 (1991).

[2] F. H. Arnold, L. Giver, A. Gershenson, H. Zhao and K. Miyazaki, Annals New York Academy of Sciences **870**, 400 (1999).

[3] J. D. Bloom, L. I. Gong, and D. Baltimore, Science **4**, 1272 (2010).

[4] J. D. Bloom and M. J. Glassman, PLoS Comp. Bio. **5**, 1 (2009)

[5] U.H.E. Hansmann and Y. Okamoto, Cur. Op. Struct. Bio. **9**, 177 (1999)

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

[7] A. Kolinski, J. Skolnick and R. Yaris, J. Chem. Phys. **86**, 7164 (1987)