HARVARD UNIVERSITY
Graduate School of Arts and Sciences

DISSERTATION ACCEPTANCE CERTIFICATE

The undersigned, appointed by the

Department of Chemistry & Chemical Biology

have examined a dissertation entitled:

Computer Simulations of Protein Folding and Evolution

presented by : Jiabin Xu

candidate for the degree of Doctor of Philosophy and hereby certify that it is worthy of acceptance.

Signature _____

Typed name: Eugene Shakhnovich

Signature _____

Typed name: Sunney Xie

Signature _____

Typed name: Gregory Verdine

Date: 05 April 2013

Computer Simulations of Protein Folding and Evolution

A dissertation presented

by

Jiabin Xu

to

The Department of Chemistry and Chemical Biology

in partial fulfillment of the requirements

for the degree of

Doctor of Philosophy

in the subject of

Chemical Physics

Harvard University

Cambridge, Massachusetts

April 2013

ProQuest LLC.
789 East Eisenhower Parkway
P.O. Box 1346
Ann Arbor, MI 48106 - 1346

Eugene I. Shakhnovich Jiabin Xu

Computer Simulations of Protein Folding and Evolution

Abstract

Computer simulations for investigating protein folding and evolution are presented. In chapter 1, an all-atom model with a knowledge-based potential is used to study the folding kinetics of Formin-Binding protein. We study the folding kinetics by performing Monte Carlo simulations. We examine the order of formation of two β-hairpins, the folding mechanism of each individual β-hairpin, and transition state ensemble (TSE) and compare our results with experimental data and previous computational studies. Further, a rigorous P_{fold} analysis is used to obtain representative samples of the TSEs showing good quantitative agreement between experimental and simulated Φ values.

In chapter 2, the underlying mechanism of the co-evolution of regulatory and protein coding sequences is studied. Regulatory sequences control the expression of a gene. The protein coding sequence determines the probability of a protein folding correctly through thermodynamic stability. Because organismal fitness is determined by both the total protein products and by the probability of a protein folding correctly, we expect there to be co-evolution between regulatory sequences and protein coding sequences. We provide support for our hypothesis using a molecular-level evolutionary simulation. The results of our simulation are consistent with previous findings demonstrating that highly expressed genes are stable and evolve relatively

slowly. Our simulation also shows that the number of substitutions in a regulatory sequence is positively correlated with the rate of evolution in the coding sequence and that highly expressed genes have low upstream regulatory sequence substitution rates. We then analyze sequence data from yeast; the results of this analysis confirm those of our simulation.

In chapter 3, we study how recombination and mutation act together to shape protein evolution. We use a biophysical model of protein folding with explicit sequences and protein structures. The biophysical model allows us to consider the roles of mutation and recombination in the context of a realistic biophysical fitness landscape. Our model naturally includes epistasis and sequence depletion effects. In addition, our explicit sequence model permits intragenic recombination. We find that mutation and recombination have different effects on the adaptation process, protein stability and the origin and fixation of recombinant alleles during protein evolution.

Contents

Acknowledgement

PhD study is like a marathon. At the moment of reaching the finish line, I am grateful to many individuals who gave me tremendous support and myriad contributions throughout my graduate study. Without them, the work presented in this dissertation would not have been accomplished.

First and foremost, I wish to express my gratitude to my advisor, Prof. Eugene Shakhnovich, who introduced me to the challenges and excitement of protein folding and evolution problems. I have benefited tremendously not only from his broad and profound knowledge but also from his creative big-picture viewpoint in thinking of concrete issues. Through the whole PhD program, I have always been motivated and inspired from his perpetual enthusiasm and curiosity in research. I am really grateful for his generous support and patient guidance on every stage of my progress. What I learned from him is much more than what could be reflected in this dissertation.

I would also like to thank my two other committee members: Prof. Sunney Xie and Prof. Gregory Verdine for the encouragement and discussion of my PhD progress meeting during each academic year.

I am also sincerely grateful to current and past members in Shakhnovich Lab for their help and insightful discussions. Lei Huang worked with me on the protein folding

project when I first joined the lab. Adrian Serohijos and I had many thought-provoking discussions on protein evolutions. Peter Kutchukian, Lee Wei Yang and Amy Gilson read my manuscripts and offered many useful suggestions. I am also grateful to other current and past Shakhnovich Lab members, including Shimon Bershtein, Jingshan Zhang, Muyoung Heo, Peiqiu Chen, Scott Wylie, Jeong-Mo Choi, Murat Centibas, Nicolas Cheron, Jaie Woodard, Orit Peleg, Ariel Weinberg, Bharat Adkar, Tatyana Kuznetsova. Many thanks also to Judy Morrison for her administrative assistance.

Lastly, I must thank my family for their never ending love and support, regardless of where I am in the world. Without their encouragement and spiritual support, I can never arrive at this stage.

Chapter 1.

The ensemble folding kinetics of the FBP28 WW domain revealed by an all-atom Monte Carlo simulation in a knowledge-based potential.

1.1 Abstract:

In this work, we apply a detailed all-atom model with a transferable knowledge-based potential to study the folding kinetics of Formin-Binding protein, FBP28, which is a canonical three-stranded β-sheet WW domain. Replica exchange Monte Carlo (REMC) simulations starting from random coils find native-like (Cα RMSD of 2.68Å) lowest energy structure. We also study the folding kinetics of FBP28 WW domain by performing a large number of *ab initio* Monte Carlo folding simulations. Using these trajectories, we examine the order of formation of two β–hairpins, the folding mechanism of each individual β–hairpin, and transition state ensemble (TSE) of FBP28 WW domain and compare our results with experimental data and previous computational studies. To obtain detailed structural information on the folding dynamics viewed as an ensemble process, we perform a clustering analysis procedure based on graph theory. Further, a rigorous P_{fold} analysis is used to obtain representative samples of the TSEs showing good quantitative agreement between experimental and simulated Φ values. Our analysis shows that the turn structure between first and second β strands is a partially stable structural motif that gets formed before entering the TSE in FBP28 WW domain and there exist two major

pathways for the folding of FBP28 WW domain, which differ in the order and mechanism of hairpin formation.

1.2 Introduction:

Understanding the folding mechanism of β–structure is crucial for general and comprehensive understanding of protein folding kinetics. Compared to α-helical proteins, structure prediction and study of folding kinetics of β-proteins is more computationally challenging because β–hairpin is an extended structure with a large number of long-range contacts, making it more difficult to reach its correct structure in an atomistic computer simulation.[1] Therefore, most simulation studies on folding of β-proteins are limited to small β-sheet domain, for example, the WW domain.[2-8] Formin-binding protein 28 WW domain (FBP28) is one member of the WW domain family. FBP28 is a small three-stranded β-sheet protein with high content of hydrophobic and aromatic residues. The characteristic features of WW domain are that this family of proteins has two highly conserved tryptophan residues and a strictly conserved proline residue. The native structure of FBP28 has been resolved by NMR.[9-10] The FBP28 makes interactions with many signaling and regulatory proteins,[11] and can also form complexes which have been implicated in a number of diseases such as Alzheimer's and Huntington's disease.[12] FBP28 unfolds reversibly in both denaturant and thermal denaturation experiments[10,13-17], but it can also form amyloids at elevated temperature.[18] Temperature jump experiment showed that folding of FBP 28 is a cooperative, two-state process without any intermediate state

detected.[13] Another laser-temperature jump experiment suggests that there are two decay phases for wild-type FBP28, the fast one is about 30 μs and the slow one is >900 μs at low temperature.[19] The heterogeneity suggests that a third state has to be considered in the folding process. Moreover, a large number of Φ values have been obtained experimentally by mutational analysis on FBP28, which may serve as a benchmark for simulation studies.[8]

Many computational studies have been performed on FBP28 or other family members of WW domain to gain insight into the formation of β structures. These studies can be grouped into the following categories: the first type of simulations employ high temperature unfolding.[8] The drawback of this approach is that the reconstructed high-T folding pathways do not necessarily dynamically coincide with ones at ambient temperature.[20-21] The second type of simulations used replica exchange method, e.g. REMD (Replica Exchange Molecular Dynamics)[22] and multiplexed Q-replica molecular dynamics,[3] to study equilibrium thermodynamics of the protein and derive the folding pathway indirectly from the free energy landscape. However the issue of how to derive dynamics from low-dimensional projections of energy landscape remains unresolved.[23-24] The third type of simulations used the structure-based Gō model to directly study the folding dynamics at a fixed temperature from extended random coils.[2,5] There are no attractive non-native interactions in the Gō model which may be unphysical – several studies showed the importance of transient stabilizing non-native interactions at various stages of folding.[25-27] Recent simulation used "physics-based" force field to study folding

dynamics of WW domains at fixed temperature.[6,28] However, this method, while highly desirable, is still too computationally costly to produce sufficient number of folding events for detailed statistical anlysis.

Recently, we developed an all-atom knowledge-based potential, which succeeded in folding a diverse set of proteins to their near-native conformations.[29] In addition, our potential, combined with dynamic Metropolis Monte Carlo (MC) simulation methods, has been used to study folding dynamics of α-helical proteins directly from extended random coils at a fixed temperature.[30-31] Our group used structural kinetics cluster analysis in combination with transition state ensemble analysis and Φ value calculation to analyze folding pathways of α-helical proteins.[30,32] Good agreement with experiment suggests that this approach can reproduce folding dynamics of proteins efficiently and with good accuracy. The key feature of our approach is that it uses an all-atom model to provide an atomistically resolved picture of the folding process. However, it is somewhat coarse-grained dynamically making it efficient enough to generate a large number of long-time trajectories to glean statistically significant robust features of the folding process. Here we apply this approach to get insights into folding mechanism(s) of β–proteins using FBP28 as our model. There are several fundamental questions concerning folding of FBP28 as a prototypical β–protein. For instance, in what order are two β hairpins formed in FBP28? What's the folding mechanism(s) of individual β hairpins? Are they the same or different? What's the TSE (transition state ensemble) and nucleation center during the folding process? The purpose of this chapter is to

address these questions by direct all-atom folding simulation.

1.3 Models and Methods:

The detailed description of the simulation model could be found elsewhere[29-30]. Here we give a brief summary of the model and simulation technique. First, all heavy-atom positions of the FBP28 WW domain were acquired from the NMR structure (residues 6-32 of Protein Data Bank id 1e0l), with the unstructured tails truncated.[10] The truncation of the N-terminal residues had no observable effect on the stability of the domain.[19] The truncation of the C-terminal residues decreases the stability of the protein because Trp-8, Tyr-20 and Pro-33 form a hydrophobic core in the wild-type native state. Nevertheless, the truncation does not result in significant structural change of the native state.[19,33] There are 27 residues and 238 atoms in total. In our model, Tyr-11, Tyr-19, Tyr-21 and Trp-30 form main hydrophobic core. Trp-8 and Tyr-20 form another hydrophobic core. The all-atom "knowledge-based" transferable energy function takes the form as:

$$E = w_{con} \times E_{con} + w_{trp} \times E_{trp} + w_{hb} \times E_{hb} + w_{sct} \times E_{sct} \qquad [1]$$

where E_{con} is the pairwise atom-atom contact potential, E_{hb} is the hydrogen-bonding potential, E_{trp} is the sequence-dependent local torsional potential based on the statistics of sequential amino acid triplets, and E_{sct} is the side-chain torsional angle potential.

To test the ability of the potential to identify near-native state as lowest energy one, we use the REMC simulation to sample the conformation space with 32 replicas

at different temperatures, ranging from 0.15 to 1.50. In the REMC simulation, we can move ψ and χ angles of all residues and ϕ except in proline and we use three different move sets to increase the sampling efficiency: backbone moves, side-chain moves, and "knowledge-based" moves. The backbone move has two types with equal probability: global move and local move. A global move is to rotate the dihedral angle (ϕ or ψ) of a randomly selected residue. A local move moves seven successive torsional angles with other residues unchanged. The step sizes of the global and local moves for the backbone are drawn from a normal distribution with zero mean and standard deviation of 2^o and 60^o, respectively. A side-chain move consists of rotating all χ angles in a randomly selected nonproline residue. The step size of the side-chain rotation is drawn from a normal distribution with zero mean and standard deviation of 10^o. The knowledge-based moves were discussed in details elsewhere.[34] A knowledge-based move of a residue during simulation entails setting the dihedral angles of the residue randomly to one of the clustered ϕ/ψ angles. The knowledge-based move can efficiently sample low energy states. For folding kinetics study, we perform 2304 independent Monte Carlo simulations, starting from different random coil configurations at T = 0.50 for 10^8 steps. The ensemble of initial random coil conformations is obtained by first running 5×10^5 MC steps at very high temperature, T = 1000 for each trajectory. Snapshots were stored at every 5×10^5 MC steps. Backbone moves and side-chains moves are still used in folding kinetics simulation. To satisfy the detailed balance condition, a knowledge-based move used in REMC simulation was not used, and the local move set was modified.[35] A new

sampling method rather than the conventional Metropolis rule is used to conserve detailed balance. The probability of accepting a move from the old state o to the new state n for the local move set is given by

$$P(o \to n) = \min\left[1, \frac{N^{(n)} \exp(-U(n)/T)J(n)}{N^{(o)} \exp(-U(o)/T)J(o)}\right],$$

where N is the number of solutions, U is the potential energy, T is temperature, and J is the Jacobian determinant.

Not all of the 2304 trajectories contain native-like low-energy structures. Therefore, before turning to the folding kinetics, we make an initial objective selection of a set of "representative" trajectories. There is one minimum energy structure in each of the 2304 trajectories and we select 100 trajectories whose minimum energy structures have the lowest energies. To better quantify the structure similarity between the simulation structure and native structure, we use fraction of nonlocal native contacts (|i-j|>2) as our order parameter to monitor the folding process. Two residues are in contact if any two of their heavy atoms are in contact. Two heavy atoms are defined to be in contact if the distance between them is less than $\lambda(r_A + r_B)$, where r_A and r_B are their van der Waals radii and $\lambda = 1.8$.[29]

A simulated Φ value is defined according to Vendruscolo and co-workers [36] as

$$\Phi_I^{sim} = \frac{N_I^{TS}}{N_I^{NS}}$$

where N_I^{TS} is the average number of native contacts made by residue I in the transition state ensemble, and N_I^{NS} is the number of native contacts made by residue I in the native state.

1.4 Results:

First, we check whether our potential can identify a set of near-native conformations of FBP 28 as global energy minimum. To that end, we performed replica exchange Monte Carlo (REMC) simulation with our energy function, starting from random coils. We obtained a total of 14719 structures and the energy landscape is shown in Figure 1.1(A). The minimum energy structure (Figure 1.1(B)) has the correct topology with three β strands correctly folded and a Cα RMSD of 2.68Å. Some differences between the simulated lowest energy structure and the experimental structure are: first, Ser-6 has no contacts with other residues in the experimental structure, while it has contacts with Asn-23 and Arg-24 in the simulated structure. Second, Trp-8 has several contacts with Glu-27, Ser-28 and Thr-29 in the experimental structure, while such contacts are not observed in the simulated structure. Third, the β strand 3 in our simulated minimum energy structure is longer than that in the native structure. Importantly, our simulation correctly predicts two hydrophobic cores and side-chains belonging to these two hydrophobic cores are in the correct position. The results show the power of our knowledge-based potential to discriminate between near-native conformations and misfolded ones.

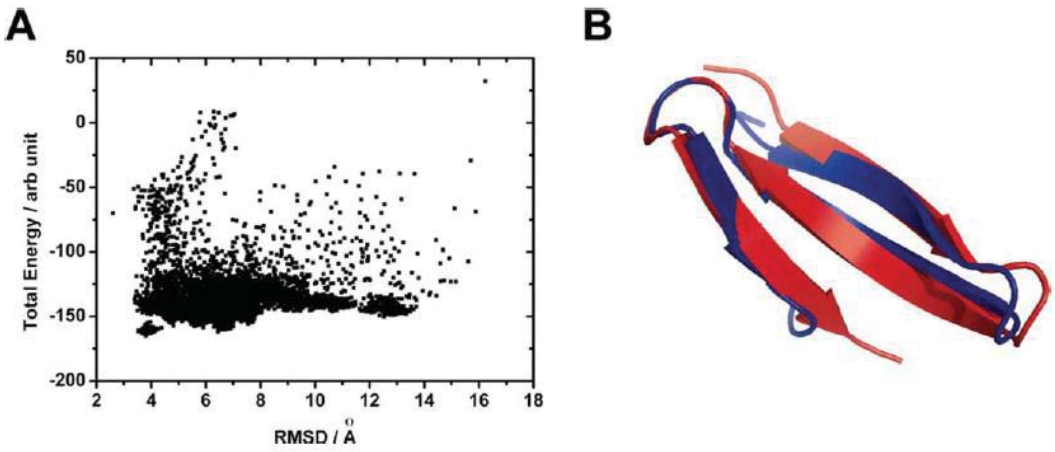


Figure 1.1(A) The energy landscape for FBP 28 WW domain in ab initio REMC simulations as projected onto RMSD axis.

Figure 1.1(B) Superposition of the backbones of the native structure (in blue) and minimum energy structure (in red) obtained through the REMC simulations. The RMSD and CαRMSD between the minimum energy structure and the native structure are 3.79 Å and 2.68Å, respectively. Structures were created by using PyMOL.[44]

1.4.1 Folding dynamics and secondary structure formation

We selected 100 trajectories out of total 2304 for detailed analysis of folding dynamics. The temperature used in our dynamic Monte Carlo simulation is 0.5 in arbitrary units of temperature used in our simulations. We relate our temperature units to real temperature using the simulated melting curve simulation (Figure 1.2), which shows mid-transition at ~ 0.6, while the experimental folding temperature of FBP28 is 337K.[8] Therefore, our simulation temperature of 0.5 corresponds to real temperature of ~281K.

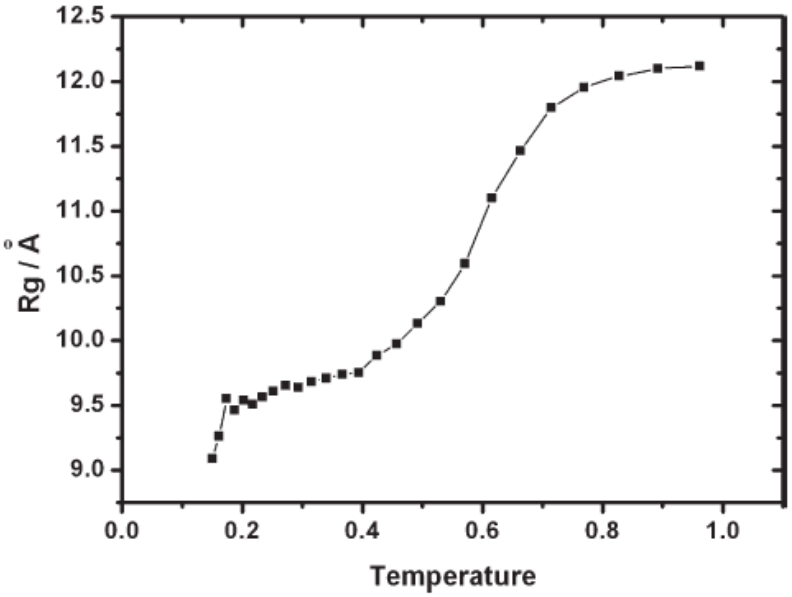


Figure 1.2 Simulated melting curve for the FBP28 WW domain in terms of average size of the molecule (Rg) vs temperature.

The average fraction of total native contacts Q and the native contacts between β1 and β2, between β2 and β3, within loop 1 and between loop 1 and other residues, and within loop 2 and between loop 2 and other residues (averaged over all of 100 folding trajectories) are shown as a function MC time-steps in Figure 1.3(A). The formation of the native contacts between β1 and β2 is faster than the formation of the native contacts between β2 and β3. Also, there is a rapid formation of these structures at early stages of folding.

To further understand the details of the folding process, we plot the probabilities of contacts at different stages of folding (Q between 0.0 and 0.5) in Figure 1.3(B-F). All 100 trajectories are used to make the plot in Figure 1.3. At $0.0<Q<0.1$, the contact pair between the Tyr-21 and Arg-24 has the highest contact probability (0.225). The majority of the contacts are neighboring contacts, indicating that the structures are

still in the random coil state. At 0.1<Q<0.2, the highest contact probability locates at loop 1 region for β hairpin 1 and the contact probability decreases outward from the turn to the end of the hairpin. The highest contact probability for β hairpin 2 locates at loop 2 region for β hairpin 2. At 0.2<Q<0.3, the contact probability for β hairpin 1 continues to increase over 0.40 and the contact probability for β hairpin 2 has little change compared to 0.1<Q<0.2, indicating that the formation of β hairpin 1 could occur earlier than the formation of β hairpin 2. At 0.3<Q<0.4, the contact probability in the loop1 region increases to over 0.50 along with increased probabilities of other contacts within loop1 and between β1 and β2. For β hairpin 2, the contact probability increases to about 0.4 for two regions and in between these two green regions there is a blue region with a lower contact probability. At 0.4<Q<0.5, the contact probabilities for pair residues in β hairpin 1 reach over 0.7 and the contact probabilities for pair of residues in β hairpin 2 are over 0.3. The above results suggest that statistically there are more folding pathways whereby the β hairpin 1 forms first and β hairpin 2 forms later. For the folding mechanism for each β hairpin, we observed that the contacts are first formed near the turn and then propagate outward for β hairpin 1. For β hairpin 2, the contacts first formed at two separate regions in the hairpin and later the whole β hairpin is formed.

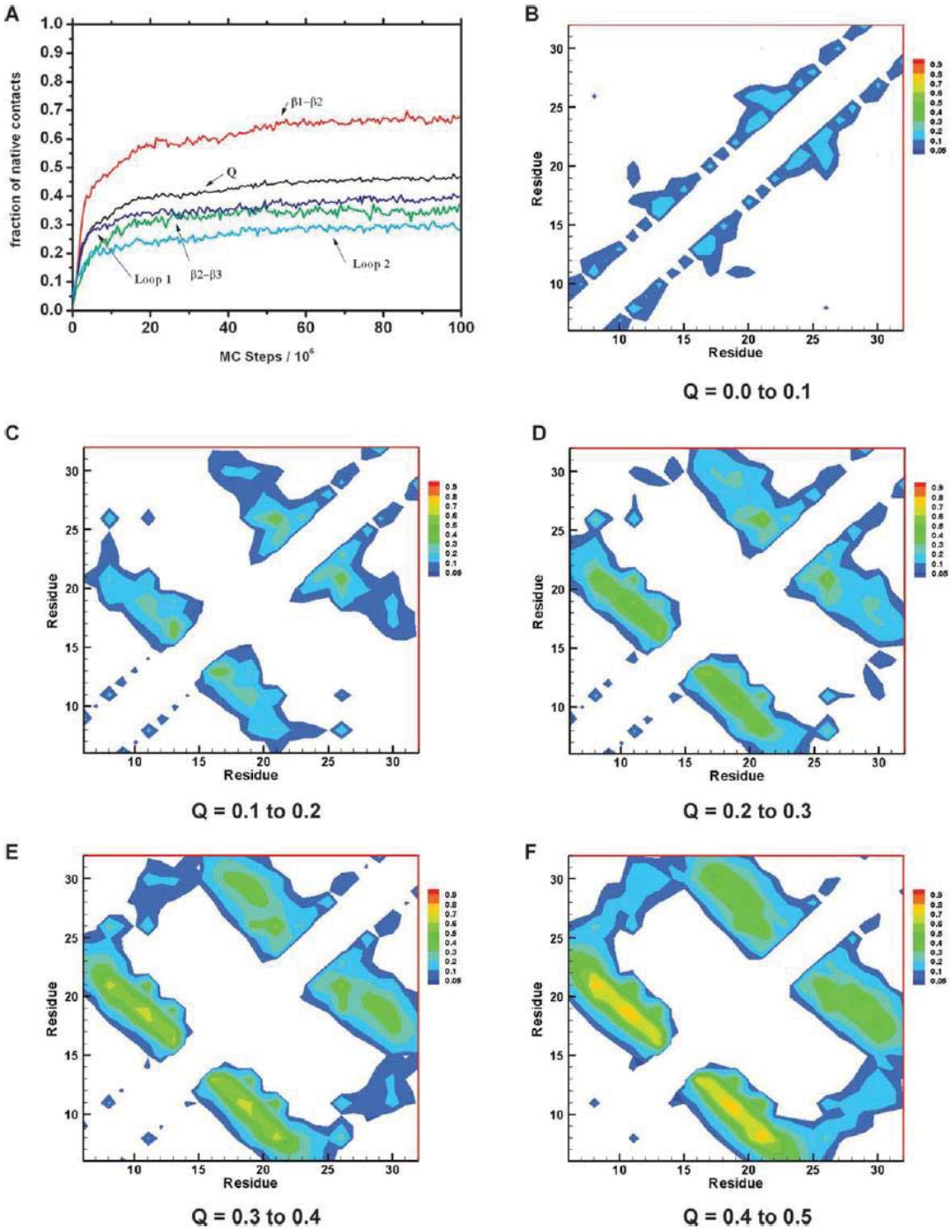


Figure 1.3(A) Fractions of native contacts averaged over all 100 trajectories as a function of MC time steps at T=0.50. The total fraction of native contacts (Q) is shown in black. The fraction of native contacts between β1 and β2 is in red, between β2 and β3 is in green, within loop 1 and between loop 1 and other residue is in blue, and within loop 2 and between loop 2 and other residues is in cyan.

Figure 1.3(B-F) (Continued) Probabilities of native residue-residue contact at various stages of folding according to the Q values. The folding temperature is 0.50.

1.4.2 Folding mechanism for individual β hairpin formation

It is worth noticing that a contact between two residues does not necessarily imply that there is a hydrogen bond between them. In order to see the formation of hydrogen bonds in both hairpins, we monitor eleven main chain H-bond contacts at different folding stages (Q values) shown in Figure 1.4(A) and Table 1.1. Since we use a heavy atom model, we measure the distance between the N atom and O atom of two residues to determine formation of a hydrogen bond. If the distance between the N atom and O atom is smaller than 3.5 Å, then we define that there is a hydrogen bond between these two residues. From Figure 1.4(B), we can see that the probability of H1 in β hairpin 1 is always highest from $0.0<Q<0.4$ and the probability decreases outward from the turn region to the end of the β hairpin 1, indicating that the formation of hydrogen bonds starts from the turn region to the end of the hairpin for β hairpin 1. For β hairpin 2, the probability is different, where the probability is lowest near the turn and it increases outward from the turn region to the end of the hairpin, indicating that the formation of hydrogen bond starts from the end of the hairpin to the turn region for β hairpin 2.

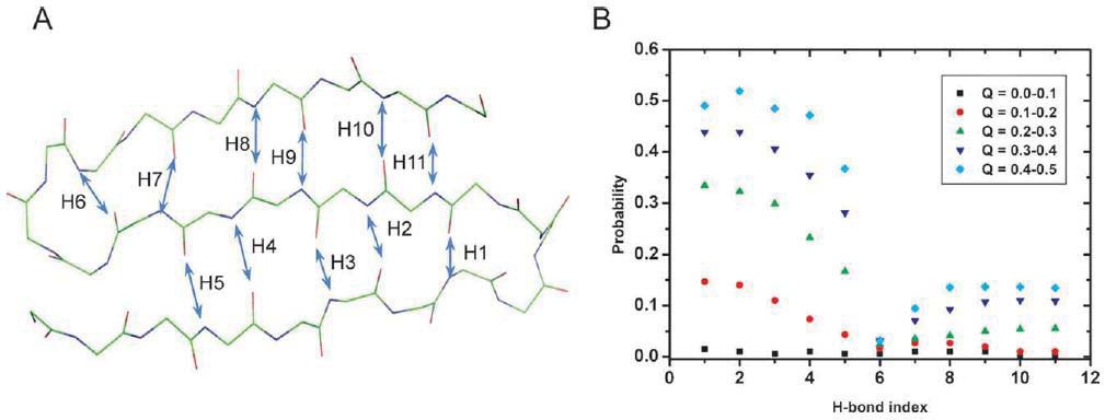


Figure 1.4(A) Eleven hydrogen bonds for β1 and β2 which are monitored during MC simulations.

Figure 1.4(B) Probabilities of 11 H-bonds at various stages of folding categorized according to the Q values of 100 folding trajectories at T=0.50. The H-bond indices are defined in the text.

Table 1.1 Eleven hydrogen bonds monitored during folding simulation

H1	Thr-13-N --- Lys-17-O
H2	Tyr-19-N --- Tyr-11-O
H3	Tyr-11-N --- Tyr-19-O
H4	Tyr-21-N --- Thr-9-O
H5	Thr-9-N --- Tyr-21-O
H6	Glu-27-N --- Asn-22-O
H7	Asn-22-N --- Glu-27-O
H8	Thr-29-N --- Tyr-20-O
H9	Tyr-20-N --- Thr-29-O
H10	Glu-31-N --- Thr-18-O
H11	Thr-18-N --- Glu-31-O

1.4.3 Structural kinetic cluster analysis

In order to identify possible obligatory intermediates during the folding process,

we use a structural cluster procedure developed before[32]. The cluster procedure uses a "structural graph" of geometrically clustered conformations to provide a coarse-grained structural and kinetic information during folding process, which is shown in Figure 1.5. The structural clustering procedure is different from kinetic clustering employed by several authors[37-39] and is carried out in two steps. In the first step, all snapshots from 100 representative trajectories are clustered in a single-link graph. Each node in this graph represents a conformation. Two nodes are linked together by an edge if their structural similarity distance measure d is smaller than the cutoff value d_c. Therefore, we will get several clusters in our "structural graph" after the first step. The largest cluster, which contains near-native conformations, is called the Giant Component (GC). In the second step, an important quantity flux, F, which is defined as the fraction of all trajectories passing through the cluster, is introduced to characterize the clusters kinetically. Therefore, the clusters with high F constitute major folding intermediates (on or off-pathway). Clusters with F=1 are the set of conformations constituting obligatory intermediate states. In addition, we also calculate the mean-first passage time (MFPT) and the mean least-exit time (MLET) for each cluster. Finally, one representative structure, defined as the structure with the highest number of edges, is extracted from each cluster. These quantities, together with the representative structure from each cluster, provide a detailed picture of folding process from an ensemble perspective.

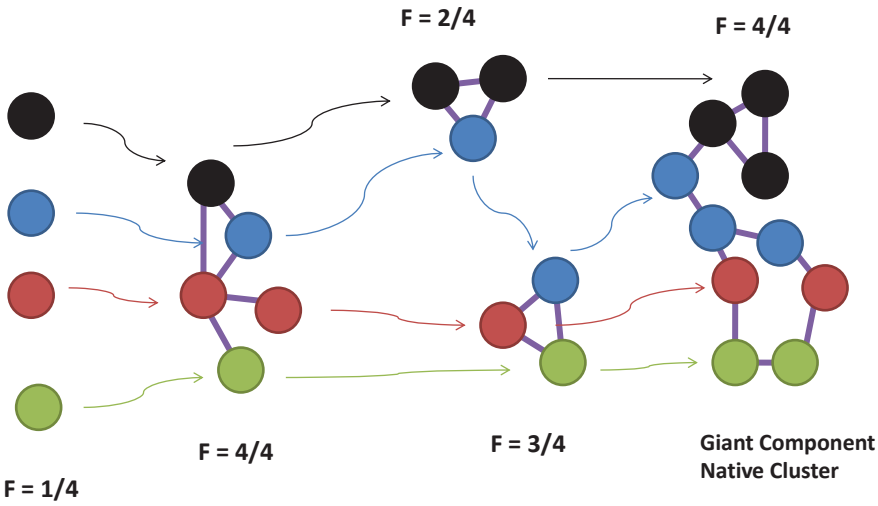


Figure 1.5 An illustration of the structural kinetic cluster analysis. Different colors represent different simulation trajectories. The arrows represent the direction of time steps in the simulation trajectories. Each node is a specific protein conformation. The line connecting two nodes represent connection based on structural similarity measure. All simulation trajectories end up in the right most Giant Component Native Cluster. The F value represents the fraction of simulation trajectories included in the cluster. All clusters with F = 1 preceding the Giant Component Native Cluster is an obligate intermediate.

In this work, we follow Hubner et al.[32] and use rmsd, distance rmsd(drms) and Rg as our order parameters for clustering. Each order parameter provides different complementary perspectives on the folding process. Table 1.2-1.4 show the results of structural kinetic cluster analysis for FBP 28 WW domain. When we use drms and

rmsd as order parameters, we find only one single dominating cluster with high flux, which is the native state cluster. The absence of high flux clusters at early time of the folding process in the drms and rmsd structural cluster result means that at the initial stage of folding, there is no accumulation of a structurally well-defined folding intermediate. When we use Rg as order parameter, we observe a large number of clusters at early time of folding process with large variation of Rg. The largest cluster (GC) is a low-Rg cluster with MFPT $\approx 4\times10^{6}$ MC steps. However, the GC in the Rg cluster must contain not only conformations that are part of the native conformational ensemble but also pre-TSE low Rg conformations. We observe some representative structures with folded β hairpin 2, but fragments of β1 form a small α helix.(Figure 1.6) This type of structures is observed in a recent unfolding simulation using explicit solvent[7] and high temperature unfolding MD simulation.[8]

Table 1.2: Summary of the clusters identified by order parameter Rg. Values are averaged over the entire cluster. MFPT and MLET are in $5*10^5$ MC steps

CLUSTER INDEX	FLUX	MFPT	MLET	<RG>
5	1.00	7.81	197.14	9.08
32	0.53	20.51	43.91	10.54
19	0.52	21.63	44.23	10.69
28	0.45	31.84	55.64	10.31
53	0.37	14.76	39.27	13.31
22	0.36	27.78	47.06	10.92
31	0.35	21.31	50.57	11.13
23	0.34	23.71	50.56	11.81
96	0.34	16.50	25.18	13.72
6	0.33	25.06	42.42	13.50
100	0.33	28.30	42.79	12.05
12	0.32	30.88	40.31	13.14
29	0.32	27.66	50.88	11.35
42	0.32	29.25	40.59	12.57
41	0.31	36.42	49.61	10.39
77	0.31	31.87	48.97	11.96
16	0.28	28.29	43.50	11.27
40	0.28	24.43	45.50	11.05
47	0.28	32.39	52.50	11.64
54	0.28	31.57	48.96	12.92
14	0.27	22.93	30.11	12.35
50	0.27	20.81	40.85	12.71
45	0.24	29.79	45.63	12.13
49	0.24	26.71	49.71	13.59
59	0.23	38.96	62.57	11.90
20	0.22	24.55	49.55	11.71
85	0.22	26.77	37.00	14.25
15	0.21	36.81	49.62	10.98
17	0.20	42.05	52.30	13.85
21	0.20	34.30	49.60	11.48
24	0.20	36.15	51.60	10.79
33	0.20	36.60	57.50	11.22
101	0.20	42.00	52.90	12.50

Table 1.3: Summary of the clusters identified by order parameter drms. Values are averaged over the entire cluster. MFPT and MLET are in $5*10^5$ MC steps

CLUSTER NO.	FLUX	MFPT	MLET	<RG>
5	0.98	9.59	199.00	9.46
159	0.06	21.83	37.17	10.54
14	0.04	0.50	0.50	18.47
55	0.04	5.00	5.00	10.51
57	0.03	56.67	85.33	11.09
154	0.03	28.33	63.00	12.41
342	0.03	8.67	8.67	11.85
362	0.03	14.33	19.33	17.71

Table 1.4: Summary of the clusters identified by order parameter RMSD. Values are averaged over the entire cluster. MFPT and MLET are in $5*10^5$ MC steps

CLUSTER NUMBER	FLUX	MFPT	MLET	<RG>
5	0.81	25.01	196.73	9.17
226	0.06	37.67	68.00	12.85
245	0.05	53.00	55.20	11.17
20	0.03	34.33	37.67	11.76
111	0.03	36.00	47.00	13.00
541	0.03	52.33	66.33	13.29
1408	0.03	43.67	75.67	10.09

.

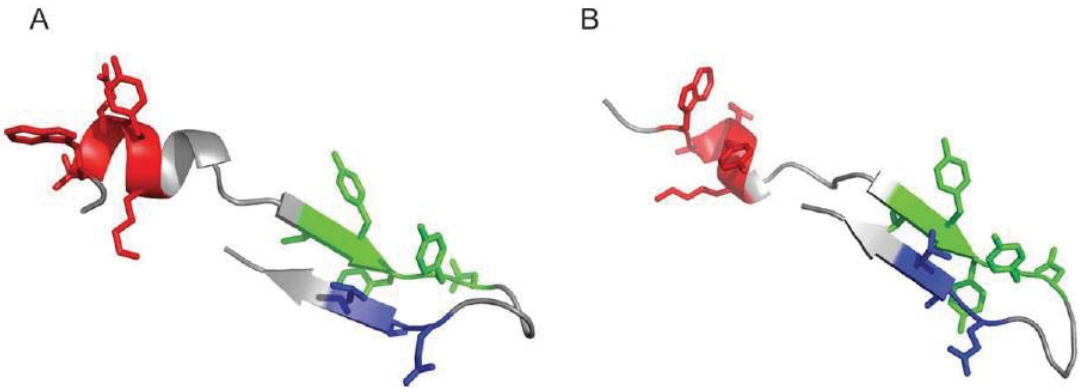


Figure 1.6 Two representative structures from two large Rg clusters with α-helical structures at N-terminus.

1.4.4 Transition state ensembles

Transition state ensembles (TSEs) are key to understand the folding pathways. We use P_{fold} analysis to construct the TSEs from putative TSEs.[23] The P_{fold} analysis is based on the fact that simulations starting from a transition state conformation have equal probability of reaching the native state and a conformation belonging to the unfolded state because TSEs have higher free energy as illustrated by Figure 1.7. The way we get the putative TSEs is to select structures that immediately precede entry into the Giant Component (GC), which is the largest cluster in the RMSD structural cluster graph, which gives us 239 putative transition-state structures. For each conformation in the putative TSEs, we perform 256 independent short MC simulations with 10^6 MC steps. If the trajectory contains at least one structure whose RMSD to the minimum energy structure obtained from the REMC simulation is smaller than 3.5 Å, then we count this trajectory as a trajectory that reached the native state ensemble. Conformations with $0.4 < P_{fold} < 0.6$ constitute the TSE. This

procedure generates a set of 15 “true” transition state structures for FBP 28 WW domain. (Figure 1.8)

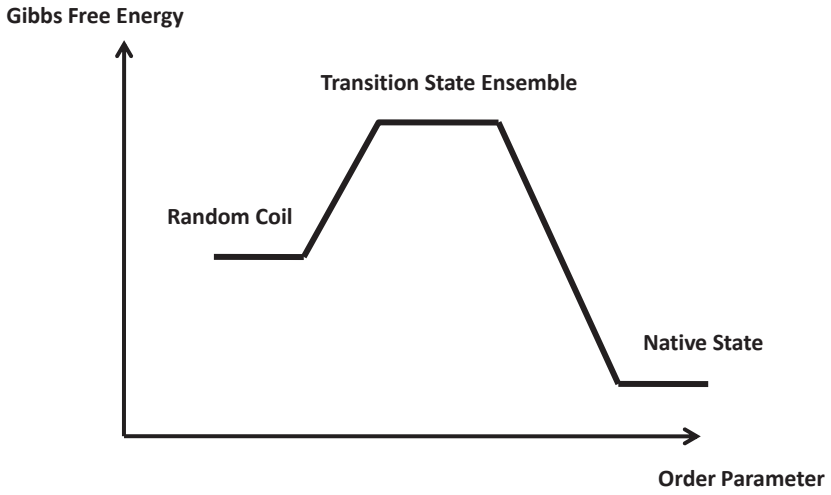


Figure 1.7 Energy landscape of Transition State Ensemble, Random Coil and Native State.

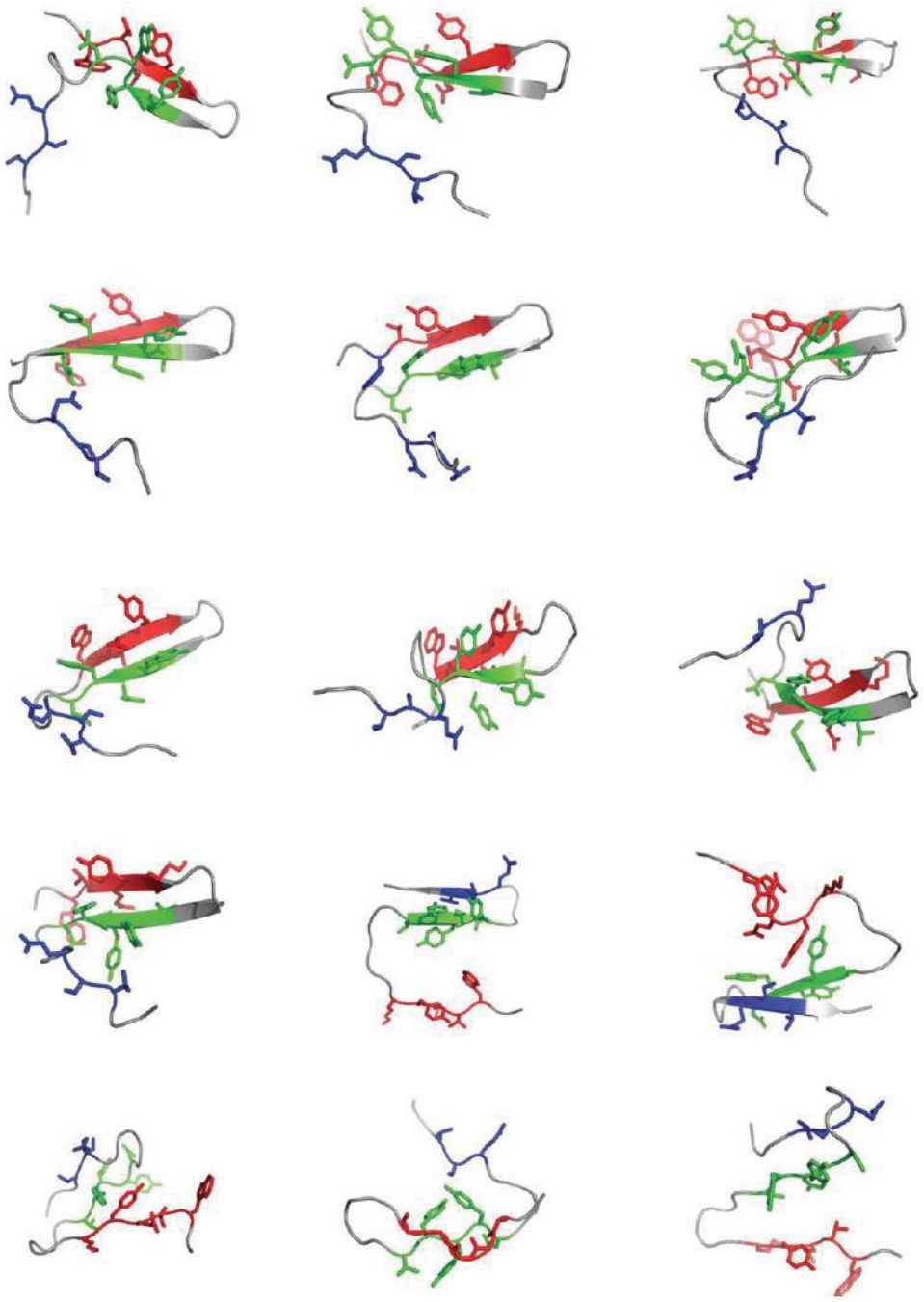


Figure 1.8 The transition state ensemble of 15 structures determined by the P_{fold} analysis.

There are 10 transition state structures having formed hairpin 1 of β1 and β2 with an unformed hairpin 2 of β2 and β3. Two transition state structures have a well-formed hairpin 2 of β2 and β3 with an unformed hairpin 1 of β1 and β2. The

remaining 3 transition state structures do not have secondary structures formed. This type of transition state structures with no secondary structures formed are also observed in the previous study of high temperature unfolding MD simulations.[8] The structural analysis of the transition state ensemble demonstrates that the dominant folding pathway is that first β hairpin forms first and second β hairpin forms later. The minor folding pathway is that second β hairpin forms first and first β hairpin forms later.

Having obtained the true TSEs by P_{fold} analysis, we are now ready to use these structures to calculate the theoretical Φ values for FBP 28 WW domain. Following previous conventions[36], Φ_i for a residue i is interpreted as the number of contacts present in the TSE for residue i divided by the number of native contacts(of the same residue i). The simulation version of Φ values is consistent with the experiment version of Φ values, which is illustrated in Figure 1.9. Simulation Φ values with their standard deviations, averaged over all TSE conformations are given in Figure 1.10. Experimental Φ values have been obtained previously for FBP 28 WW domain.[8] The agreement between theory and experiment is good. Exceptions are Trp-8, Thr-9, Glu-10 and Ser-28 in our protein model, where the simulated Φ values are much higher than the experimental Φ values. The reason for the discrepancy is that there are very few native contacts for these residues in native structures so the simulated Φ values are not reliable – they can be very high and have large standard deviation. Another important reason is that our model uses implicit solvent. In reality these residues will form hydrogen bonds with water molecules while in simulation they will

form other intramolecular contacts, resulting in apparently high Φ values. We find that, the most structured regions in the TSE is the turn between β1 and β2, as indicated by high Φ values, which forms a native-like β hairpin turn. There is another peak of Φ values in the region between β2 and β3, which suggests that the hairpin structure between β2 and β3 is also weakly formed. This picture is in good agreement with the result obtained by detailed all-atom high temperature unfolding molecular dynamics simulation.[8]

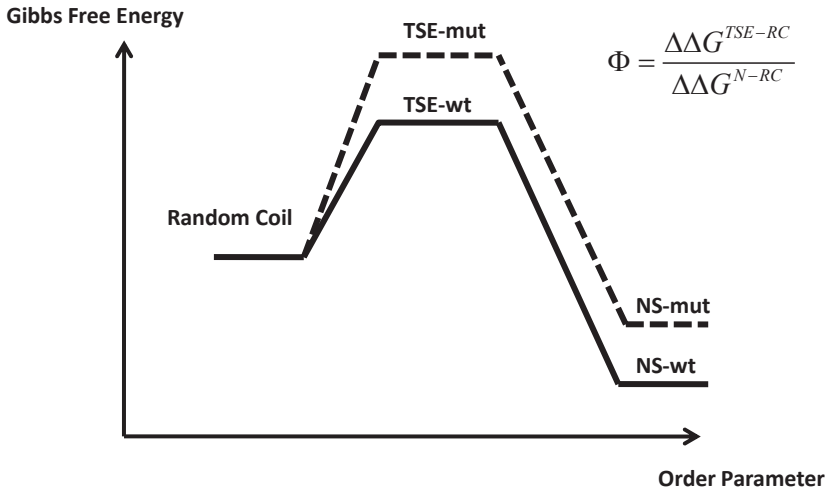


Figure 1.9 An illustration of the experimental measurement of Φ values for a residue. The Φ value is calculated as the change of free energy between the transition state and the random coil state after the mutation divided by the change of free energy between the native state and the random coil state after the mutation.

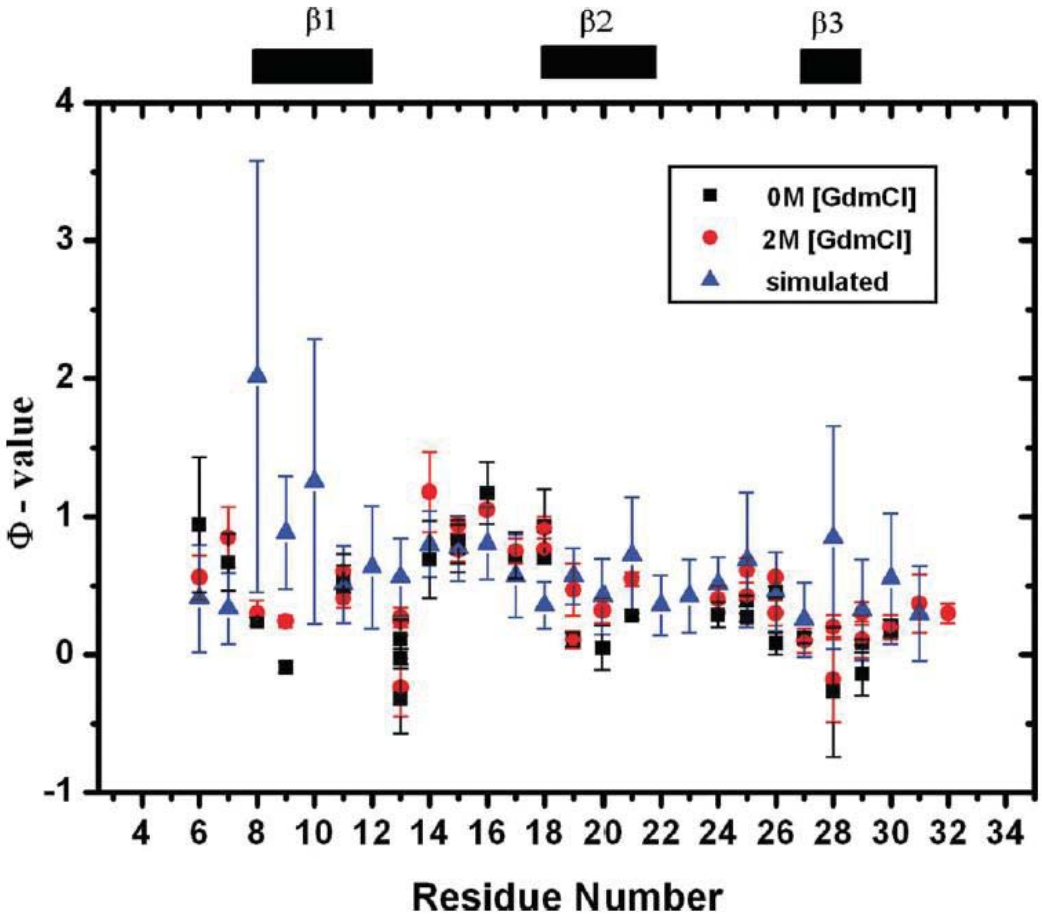


Figure 1.10 Comparison between simulated and experimental Φ values. Error bars denote the standard deviation σ of 15 Φ values calculated from 15 structures of transition state ensemble.

1.5 Discussion

1.5.1 Most probable folding pathways:

Using a relatively simple transferable knowledge-based all-atom model, we performed a large number of *ab initio* protein folding runs for FBP28 WW domain that provided us with necessary data to study the folding kinetics as an ensemble process. By combining our results, we obtained a detailed picture of the folding dynamics of the three β-strand FBP28 WW domain. The dominant folding pathway

includes first formation of β hairpin 1 which consists of β1 and β2, followed by formation of β hairpin 2 which consists of β2 and β3. The other non-dominant folding pathway is formation of β hairpin 2 first, followed by the formation of β hairpin 1. This non-dominant folding pathway was found earlier in improved Gō model simulations[40] and in a recent study using multiple rare event simulations.[7] Our finding of the propensity of hairpin 1 to form first during folding for FBP28 WW domain agrees with the result of Juraszek et al,[7] who found that the free energy barrier between unfolded states and intermediate state with only hairpin 1 formed is much lower than the free energy barrier between unfolded states and intermediate state with only hairpin 2 formed. In addition, our simulations qualitatively agree with the results by Luo et al. on Pin1 WW domain using the Gō model Molecular Dynamics simulation, which showed that Pin1 WW domain also has two folding pathways that differ by sequence in which hairpins are formed.[5] Our findings are also consistent with the simulation results by Ensign and Pande on the Fip35 in implicit solvent, in which it was found that the mechanism is heterogeneous, but that the larger hairpin (first) is more likely to form first.[41] Previous high-temperature unfolding simulation has shown that the contacts of the first β hairpin forming early in the folding process is the dominant folding pathway[8] and our result showed that this dominant pathway is still the same at ambient condition. Moreover, we also observed a structure with α-helix in the N-terminus with a relatively large Rg, which has been reported before in high temperature unfolding simulation[8] and bias-exchange metadynamics unfolding simulation[7]. A possible reason for the observation that dominant folding pathway

involves an early formation of hairpin 1 is that hairpin 1 has more aromatic residues, which belong to the hydrophobic core of the native protein, than hairpin 2. There are five aromatic hydrophobic residues (Trp-8, Tyr-11, Tyr-19, Tyr-20 and Tyr-21) involved in stabilizing β hairpin 1, while there are only two aromatic hydrophobic residues (Tyr-19 and Trp-30) involved in stabilizing β hairpin 2. Therefore, the assembly of β hairpin 1 is enthalpically more favorable than β hairpin 2. In addition, the lengths for loop 1 and loop 2 are almost the same so the entropic contributions are almost the same for two loops. Taken together these factors indicate that, - it is more likely that β hairpin 1 will get formed first.

1.5.2 Folding mechanism of two β hairpins:

There are two proposed mechanisms for β hairpin folding. The first mechanism is the "zipper" model proposed by Munoz et al,[42] which involves the initial folding of the turn structure and following formation of hydrogen bonds zipping from the turn to the end of the hairpin. The other mechanism is the hydrophobic collapse mechanism proposed by Dinner et al, stating that the hydrophobic collapse nucleates the hairpin formation.[43] Our simulations show that the folding mechanism for β hairpin 1 follows the "zipper" model while the folding mechanism for β hairpin 2 follows the hydrophobic collapse mechanism. Previous study using high temperature unfolding method showed that the folding mechanism for β hairpin 1 was hydrophobic collapse[8]. There are several possible explanations for the discrepancy between our results and the previous results. First, the previous simulation study was performed at high

temperature (373K) and the native contacts for the hydrophobic interactions are more stable to withstand the thermal fluctuations than the native contacts at the turn area for β hairpin 1. Therefore, previous high temperature unfolding simulation probably favored the hydrophobic collapse mechanism. Our simulation is performed at low temperature (~281K) and the first formation of hydrogen bonds near the turn is entropically more favorable because these contacts are spatially closer. It is therefore possible that the folding mechanism of β hairpin 1 is temperature dependent. At high temperature, the folding mechanism for β hairpin 1 may involve hydrophobic collapse and at low temperature, the folding mechanism for β hairpin 1 may follow the "zipper" model. Luo et al. used Gō model to fold Pin1 WW domain and found that β1–β2 hairpin folded via a turn zipper mechanism at low temperatures but a hydrophobic collapse mechanism at the folding-transition temperature.[2] The difference of the folding mechanism for the two hairpins can be understood as follows. For β1–β2 hairpin, the closest hydrogen bond to the turn region is between Thr-13 and Gly-16, which are only two residues apart. Therefore, it is relatively easy to get this hydrogen bond formed first due to spatial proximity. For β2–β3 hairpin, the closest hydrogen bond to the turn region is between Asn-22 and Glu-27, which are four residues apart. Therefore, it is relatively hard for this hydrogen bond to form first at low temperature. In this case, the hydrophobic interaction is the major driving force to form β2–β3 hairpin.

1.5.3 Transition state ensemble and nucleation center

Our simulation suggests that the β-turn structure in the first β hairpin is the most structured region according to the result of Φ value analysis. We also observed relatively high Φ values in the β-turn region in the second β hairpin, which corresponds to transition state conformations with only β2 and β3 formed. Our prediction from simulated Φ values analysis agrees with the previous REMD simulation by Mu et al,[22] who predict the turn-1 formation as the transition state. However, we also get other types of "non-dominant" transition state structures in our simulation, for example, transition states with no β structures formed, which were also observed in high temperature unfolding MD simulation.[8]

1.6 Conclusion:

We use our transferrable knowledge-based energy potential to perform multiple folding trajectories, which allows us to get a complete picture of the folding kinetics from an ensemble perspective. Further, we use the most reliable method, the p_{fold} analysis, to identify the transition state ensembles and calculate simulated Φ values for all residues of the FBP28. The statistically significant number of folding events, combined with the structural cluster analysis technique, provides a complete and detailed outline of the ensemble pathway of the FBP28 WW domain folding, and possibly an insight into general features of kinetics of β-sheet formation. The conclusion we get from this study is that, first, there are two folding pathways for FBP28 WW domain. The dominant folding pathway involves formation of β hairpin 1

first, followed by the formation of β hairpin 2. The other non-dominant folding pathway is the first formation of β hairpin 2, followed by the formation of β hairpin 1; second, at low temperature, the folding mechanism for the two β hairpins are different. β hairpin 1 follows the "zipper" folding mechanism and β hairpin 2 follows hydrophobic collapse folding mechanism. Third, Φ-value analysis suggests that the turn region in β hairpin 1 is the nucleation center and the transition state ensembles can be categorized as three types of conformations: (1) structures with β1 and β2; (2) structures with β2 and β3; (3) structures without secondary structures.

1.7 Reference

1. Kubelka J, Hofrichter J, Eaton WA. The protein folding 'speed limit' Curr Opin Struct Biol 2004;14:76-88.

2. Luo Z, Ding J, Zhou Y. Folding mechanisms of individual b-hairpins in a Gō model of Pin1 WW domain by all-atom molecular dynamics simulations. The Journal Of Chemical Physics 2008;128:225103-225110.

3. Kim E, Jang S, Lim M, Pak Y. Free Energy Landscape of the FBP28 WW Domain by All-Atom Direct Folding Simulation. J Phys Chem B 2010;114:7686–7691.

4. Sharpe T, Jonsson AL, Rutherford TJ, Daggett V, Fersht AR. The role of the turn in b-hairpin formation during WW domain folding. Protein Science 2007;16:2233-2239.

5. Luo Z, Ding J, Zhou Y. Temperature-Dependent Folding Pathways of Pin1 WW

Domain: An All-Atom Molecular Dynamics Simulation of a Go Model. Biophysical Journal 2007;93:2152-2161.

6. Freddolino PL, Liu F, Gruebele M, Schulten K. Ten-Microsecond Molecular Dynamics Simulation of a Fast-Folding WW Domain. Biophysical Journal: Biophysical Letters 2008;94:L75-L77.

7. Juraszek J, Bolhuis PG. (Un)Folding Mechanisms of the FBP28 WW Domain in Explicit Solvent Revealed by Multiple Rare Event Simulation Methods. Biophysical Journal 2010;98:646–656.

8. Petrovich M, Jonsson AL, Ferguson N, Daggett V, Fersht AR. Φ-Analysis at the Experimental Limits: Mechanism of β-Hairpin Formation. J Mol Biol 2006;360:865-881.

9. Chan DC, Bedford MT, Leder P. Formin binding proteins bear WWP/WW domains that bind proline-rich peptides and functionally resemble SH3 domains. EMBO J 1996;15:1045–1054.

10. Macias MJ, Gervais V, Civera C, Oschkinat H. Structural analysis of WW domains and design of a WW prototype. Nat Struct Biol 2000;7:375–379.

11. Ton-Lo W, Dongzhou H, Mohsen S, Kaizan H, Sudol M. Structure and function of the WW domain. Prog Biophys Mol Biol 1996;65:113–132.

12. Hu H, Columbus J, Zhang Y, Wu D, Lian L, Yang S, Goodwin J, Luczak C, Carter M, Chen L, James M, Davis R, Sudol M, Rodwell J, Herrero JJ. A map of WW domain family interactions. Proteomics 2004;4:643–655.

13. Ferguson N, Johnson CM, Macias M, Oschkinat H, Fersht A. Ultrafast folding of

WW domains without structured aromatic clusters in the denatured state. Proc Natl Acad Sci USA 2001;98:13002–13007.

14. Jäger M, Nguyen H, Crane JC, Kelly JW, Gruebele M. The folding mechanism of a β-sheet: the WW domain. J Mol Biol 2001;311:373-393.

15. Crane JC, Koepf EK, Kelly JW, Gruebele M. Mapping the transition state of the WW domain β-sheet. J Mol Biol 2000;298:283-292.

16. Ferguson N, Pires JR, Toepert F, Johnson CM, Pan YP, Volkmer-Engert R, Schneider-Mergener J, Daggett V, Oschkinat H, Fersht A. Using flexible loop mimetics to extend Φ-value analysis to secondary structure interactions Proc Natl Acad Sci USA 2001;98:13008-13013.

17. Sudol M, Hunter T. NeW wrinkles for an old domain. Cell 2000;103:1001-1004.

18. Ferguson N, Berriman J, Petrovich M, Sharpe TD, T FJ, Fersht AR. Rapid amyloid fiber formation from the fast-folding WW domain FBP28. Proc Natl Acad Sci USA 2003;100:9814-9819.

19. Nguyen H, Jager M, Moretto A, Gruebele M, Kelly JW. Tuning the free-energy landscape of a WW domain by temperature, mutation, and truncation. Proc Natl Acad Sci USA 2003;100:3948-3953.

20. Finkelstein AV. Can protein unfolding simulate protein folding? Protein Eng 1997;10(8):843-845.

21. Dinner AR, Karplus M. Is protein unfolding the reverse of protein folding? A lattice simulation analysis. J Mol Biol 1999;292(2):403-419.

22. Mu YG, Nordenskiold L, Tam JP. Folding, misfolding, and amyloid protofibril

formation of WW domain FBP28. Biophys J 2006;90:3983-3992.

23. Du R, Pande V, Grosberg A, Tanaka T, Shakhnovich EI. On the transition coordinate for protein folding. Journal of Chemical Physics 1998;108(1):334-350.

24. Rao F, Settanni G, Guarnera E, Caflisch A. Estimation of protein folding probability from equilibrium simulations. J Chem Phys 2005;122(18):184901.

25. Li L, Mirny LA, Shakhnovich EI. Kinetics, thermodynamics and evolution of non-native interactions in a protein folding nucleus. Nat Struct Biol 2000;7(4):336-342.

26. Clementi C, Plotkin SS. The effects of nonnative interactions on protein folding rates: theory and simulation. Protein Sci 2004;13(7):1750-1766.

27. Bowman GR, Pande VS. Protein folded states are kinetic hubs. Proc Natl Acad Sci U S A;107(24):10890-10895.

28. Maisuradze GG, Liwo A, Scheraga HA. Principal Component Analysis for Protein Folding Dynamics. J Mol Biol 2009;385:312-329.

29. Yang JS, Chen WW, Skolnick J, Shakhnovich EI. All-Atom Ab Initio Folding of a Diverse Set of Proteins. Structure (London) 2007;15:53–63.

30. Yang JS, Wallin S, Shakhnovich EI. Universality and diversity of folding mechanics for three-helix bundle proteins. Proc Natl Acad Sci USA 2008;105:895-900.

31. Kutchukian PS, Yang JS, Verdine GL, Shakhnovich EI. All-Atom Model for Stabilization of a-Helical Structure in Peptides by Hydrocarbon Staples. J Am Chem Soc 2009;131:4623-4627.

32. Hubner IA, Deeds EJ, Shakhnovich EI. Understanding ensemble protein folding at atomic detail. Proc Natl Acad Sci USA 2006;103:17747–17752.

33. Periole X, Allen LR, Tamiola K, Mark AE, Paci E. Probing the free energy landscape of the FBP28WW domain using multiple techniques. J Comput Chem 2009;30:1059-1068.

34. Chen WW, Yang JS, Shakhnovich EI. A Knowledge-Based Move Set for Protein Folding. Proteins: Structure, Function, and Bioinformatics 2007;66:682-688.

35. Dodd LR, Boone TD, Theodorou DN. A concerted rotation algorithm for atomistic Monte Carlo simulation of polymer melts and glasses. Mol Phys 1993;78:961-996.

36. Paci E, Vendruscolo M, Dobson CM, Karplus M. Determination of a transition state at atomic resolution from protein engineering data. J Mol Biol 2002;324:151-163.

37. Rao F, Caflisch A. The protein folding network. J Mol Biol 2004;342:299-306.

38. Bowman GR, Huang X, Pande VS. Using generalized ensemble simulations and Markov state models to identify conformational states. Methods 2009;49:197-201.

39. Karpen ME, Tobias DJ, Brooks CLr. Statistical clustering techniques for the analysis of long molecular dynamics trajectories: analysis of 2.2-ns trajectories of YPGDV. Biochemistry 1993;32:412-420.

40. Karanicolas J, Brooks CL. Improved Go-like models demonstrate the robustness of protein folding mechanisms towards non-native interactions. J Mol Biol 2003;334:309–325.

41. Ensign DL, Pande VS. The Fip35 WW domain folds with structural and mechanistic heterogeneity in molecular dynamics simulations. Biophys J 2009;96:L53-L55.

42. Munoz V, Henry ER, Hofrichter J, Eaton WA. A statistical mechanical model for b-hairpin kinetics. Proc Natl Acad Sci USA 1998;95:5872-5879.

43. Dinner AR, Lazaridis T, Karplus M. Understanding beta-hairpin formation. Proc Natl Acad Sci USA 1999;96:9068-9073.

44. DeLano WL. The PYMOL Molecular Graphics System. (DeLano, San Carlos, CA) 2002.

Chapter 2

Co-evolution of Regulatory and Protein Coding Sequences

2.1 Abstract

Identifying the factors that determine the rate of protein coding sequence evolution and the rate of regulatory sequence evolution is a central goal in the study of molecular evolution. Little is known about the mechanism underlying the co-evolution of regulatory sequence and protein coding sequence, although previous studies have suggested that there is a correlation between regulatory sequence evolution and protein coding sequence evolution. We propose the following model to explain the constraints on the co-evolution of regulatory sequence and protein coding sequence: On one hand, an organism requires a specific number of correctly folded proteins to perform biological functions, meaning that the birth rate is related to the number of correctly folded proteins. The number of correctly folded proteins is determined by the product of two factors: one is the total number of proteins produced by transcription and translation; and the other is the probability that a protein is folded correctly. The product of these two factors gives the number of correctly folded (functional) proteins. On the other hand, because misfolded proteins are toxic to the organism, the death rate is related to the number of misfolded proteins. The regulatory sequence controls the expression of a gene and therefore has an effect on the total protein products. The protein coding sequence determines the probability that a protein will be folded correctly through thermodynamic stability. Because an

organism's fitness is determined by both the total protein products and the probability of a correct folding, we expect to observe co-evolution between regulatory sequences and protein coding sequences. Here, we test this hypothesis using a molecular-level evolutionary simulation. The results of our simulation are consistent with previous demonstrations that highly expressed genes are more stable and evolve slowly. Our simulation also shows that the number of substitutions in the regulatory sequence is positively correlated with the evolution rate of the coding sequence and that highly expressed genes have low upstream regulatory sequence substitution rates. We then analyze sequence data from yeast; the results of these bioinformatic analyses show a positive correlation between coding sequence evolution rate and divergence of the upstream regulatory sequence and a negative correlation between gene expression level/protein abundance and the divergence of upstream regulatory sequence, supporting the results of our simulation.

2.2 Introduction

What is the mechanism underlying the interplay between coding and regulatory sequence evolution? To answer this question, it is important to first understand the evolution of protein coding sequences and regulatory sequences separately. Therefore, previous studies of molecular evolution can be divided into two categories: one for investigations of protein coding sequence and the other for investigations of regulatory sequences.

The first studies of protein coding sequence evolution were performed by

Zuckerkandl and Pauling over 40 years ago.[1] They found that the number of nonsynonymous substitutions per site (dN) for orthologous proteins is proportional to the divergence time of the two orthologous proteins, indicating neutral evolution rather than adaptive evolution. This finding indicates that dN is a rough measure of the fixation rate of amino acid substitution. One striking property of dN is that it has large variance for different genes; for example, Drummond et al. showed that, in yeast, the most rapidly evolving gene has a dN 1000-fold larger than that of the most slowly evolving gene.[2] It is now well-accepted that most fixed mutations within genes are neutral, and therefore, the variance of dN among different genes results from the different selection constraints on different genes.[1, 3] Thus, genes with more stringent selection constraints will have larger dN values. The availability of genomic data has allowed numerous investigators to study the underlying constraint on coding-sequence evolution. Recent bioinformatics studies have found that many properties of genes correlate with dN, including the dispensability or the essentiality of the coding gene,[4-7] its number of protein interaction partners,[8, 9] its length,[9, 10] its centrality in the protein interaction network,[11] its expression level,[2, 12, 13] its designability,[14] its relative solvent accessibility,[15] and its surface-core association.[16] Among these properties, the dominant determinant of dN is expression level.[13] This finding leads to the accepted hypothesis that protein misfolding is a dominant constraint on coding sequence evolution.[17, 18]

Compared to protein coding sequence evolution, far less is known about regulatory sequence evolution, in part because regulatory sequences are difficult to

identify.[19, 20] In addition, it is less clear how changes in regulatory sequence contribute to the fitness of an organism. Previous studies have shown that certain gene regulatory functions can be maintained despite variance in the cis-regulatory sequence.[21-23] Nevertheless, it is generally accepted that binding specificity and binding affinity are the major constraints on regulatory sequence evolution.[24, 25] Previous studies have shown that sites within the transcription factor (TF) binding sites that are involved in protein-DNA complex formation evolve more slowly than other sites and that TF binding sequences as a whole evolve more slower than the surrounding background sequence, suggesting a purifying selection mechanism.[24] In addition, comparative genomics studies of regulatory sequences in 12 *Drosophila* species found that the turnover rate of TF binding sites follows a molecular clock pattern rather than lineage-specific pattern, indicating that the evolutionary mode of these sites is neutral rather than adaptive.[26] Therefore, computational modeling of TF binding site evolution generally uses binding energy as the phenotypic trait of fitness.[25, 27]

A comparative genomic study of *Caenorhabditis elegans* and *C. briggsae* showed that there is a positive coupling effect of coding and regulatory sequence evolution, indicating that natural selection acts on genes and their upstream regulatory regions as integrated units.[28] Therefore, an understanding of the mechanism underlying the interplay between coding sequence and regulatory sequence requires an integrated approach that considers the fitness effects of coding sequence and regulatory sequence simultaneously rather than the traditional approach, which treats coding sequence and regulatory sequence separately. In this study, we hope to bridge

this gap and unite the studies of coding sequence and regulatory sequence evolution by proposing a biophysical model with which to study the co-evolution of regulatory sequence and protein coding sequence. We assume that an organism's fitness is a function of the expression levels and stabilities of its proteins. Because the regulatory sequences determine the proteins' expression levels and the coding sequences determine the proteins' stabilities, we expect that there will be some correlation between the evolution rates of regulatory sequences and coding sequences. We merge the biophysical model with monoclonal population genetics simulations that include selection and mutation. In our simulation, mutations occur in both the model protein coding region and the upstream regulatory region. Mutations in the coding region change the stability of the protein, and mutations in the regulatory region change the binding affinity of the regulatory sequence, leading to a change in protein expression level. The dynamics observed in our asexual population model recapitulate the previous findings that highly expressed genes are stable and evolve relatively slowly.[2, 17] In addition, our simulation shows a positive correlation between coding sequence evolution rate and regulatory sequence evolution rate and that gene expression and protein abundance are negatively correlated with regulatory sequence evolution rate. Finally, we perform a bioinformatics analysis of yeast genomic data to provide empirical evidence that coding sequence evolution rate and regulatory sequence evolution rate are positively correlated and that regulatory sequence evolution rate and gene expression level are negatively correlated. Our model proposes that "selection for correctly folded proteins and selection against misfolded proteins" is

the critical biophysical constraint that shapes the co-evolution of regulatory sequence and protein coding sequence.

2.3 Methods

The simulations are initiated with a cell carrying five genes with base protein abundance levels of 10, 100, 1000, 10,000 and 100,000. Each of the five genes is associated with an upstream cis-regulatory sequence that determines the abundance of the protein encoded by that gene. We run twenty monoclonal parallel simulations from the common ancestor for 150,000,000 generations. We record a population snapshot every 500,000 generations. At each generation, a random mutation occurs with equal probability at each site of the genome. Therefore, the random mutation will have 70% probability of occurring in the coding region and a 30% probability of occurring in the regulatory region; these numbers were chosen based on the ratio of coding to non-coding sequence in *S. cerevisiae*[29]. For mutations within the coding sequence, 25% are synonymous.[30] Of the nonsynonymous mutations, 10% are unconditionally lethal, meaning that these mutations abolish activity and lead to a birth rate of 0.[31] The remaining 90% of nonsynonymous mutations change the thermodynamic stability of the proteins. The change in stability ($\Delta\Delta G$) is selected at random from a Gaussian distribution with a mean of $-0.13\times\Delta G+0.23$ kcal/mol and a standard deviation of 1.7 kcal/mol. The numerical value is obtained from the linear regression of the data from the ProTherm database.[32] The dependence of $\Delta\Delta G$ on ΔG is referred to as the "sequence depletion" effect. Mutations that result in $\Delta G > 0$ are

considered lethal. Mutations within a cis-regulatory region change the binding energy $\Delta\varepsilon_{pd}$. The change in binding energy $\Delta\Delta\varepsilon_{pd}$ is drawn from a Gaussian distribution with a mean of 1 kcal/mol and a standard deviation of 2 kcal/mol. This change in binding energy is associated with a change in protein expression level. Then, the selection coefficient is calculated, and the mutation is accepted or rejected depending on the substitution probability. The effective population size is defined as 10,000. Figure 2.1 shows a flow chart of the simulation.

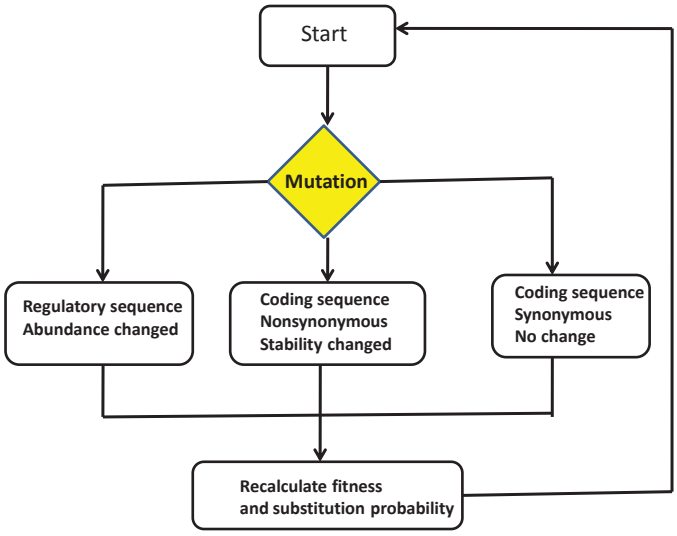


Figure 2.1 Flow chart of the monoclonal simulation

For the empirical bioinformatics study, we obtained the dS and dN data from previous studies in yeast.[4] We used the gene expression level data measured by Holstege et al.[33] We used the protein abundance data measured in yeast by Ghaemmaghami et al.[34] To characterize the evolution rate of the cis-regulatory sequence, we used two approaches.

The first approach is the calculation of TF binding site evolution rates that was previously used by Moses et al.[24] Experimentally verified binding sites for Abf1p, Gal4p, Gcn4p, Mcm1p, Rap1p, Reb1p, and Tbp1p were extracted from the Promoter database of *Saccharomyces cerevisiae* (SCPD).[35] We first used a classical parsimony algorithm to calculate the minimum number of nucleotide changes for each column of the alignment.[36] The alignment used the accepted species tree (Sbay, Smik, (Spar, Scer)).[37, 38] The evolution rate of a binding site is the sum of the number of changes at each position divided by the length of the binding site. The second method used to characterize the evolution rate of the cis-regulatory region is the Shared Motif Method (SMM), which was used previously by Castillo-Davis et al.[28] The SMM can discover regions with local similarity between two DNA sequences without respect to their order, orientation, or spacing. The fraction of shared motifs is defined as the length of regions with local similarity divided by the DNA sequence length being compared. The shared motif divergence (dSM) is defined as one minus the fraction of shared motifs. Figure 2.2 shows a sample dSM calculation. In our study, we first identify the regions upstream of the yeast open reading frames (ORFs) from the global alignment of *S. cerevisiae* and *S. paradoxus*.[37] Then we use upstream sequence of 500 bp for analysis. using 10 bp as the minimum length of a perfect stretch with no mismatches; this is equivalent to setting the minimum score of 40 (because a perfect match has score of +4). The scoring system for the algorithm is as follows: if there is a match of A,T,G,C, the score is +4, if there is a non-A,T,C,G match, the score is +1, if there is an unalignable X, the score is -100,000, and if there is a mismatch, the score is -4.

Then, the dSM for each ORF of yeast is obtained.

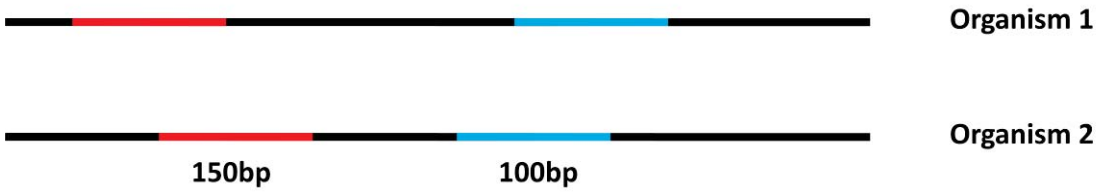


Figure 2.2 An example of the calculation of the shared motif method (SMM). In the example, suppose the two regulatory sequences have 500bp in length. After applying the local iterative alignment algorithm, two segments of sequences with significant local similarity are discovered. One region is 150bp long and the other is 100bp long and they have been translocated. The fraction of "shared motifs" between these sequences is (150+100)/500, or 0.5. The shared motif divergence (dSM) is one minus the fraction of "shared motifs". Thus, dSM quantifies the fraction of regions between two sequences that does not have significantly similar local segments.

2.4 Results

A schematic diagram of the model cell is shown in Figure 2.3. Each model cell has a genome of Γ genes, each coding an essential protein characterized by a free energy of folding (ΔG_i). Many proteins fold in a two-state manner. The fraction of time spent in the native state is given by Equation 2.1:

$$P^{nat} = \frac{e^{-\Delta G_i / kT}}{1 + e^{-\Delta G_i / kT}} \tag{2.1}$$

where k is Boltzmann's constant, and T is temperature. We assume that proteins must be in their native states to be functional.

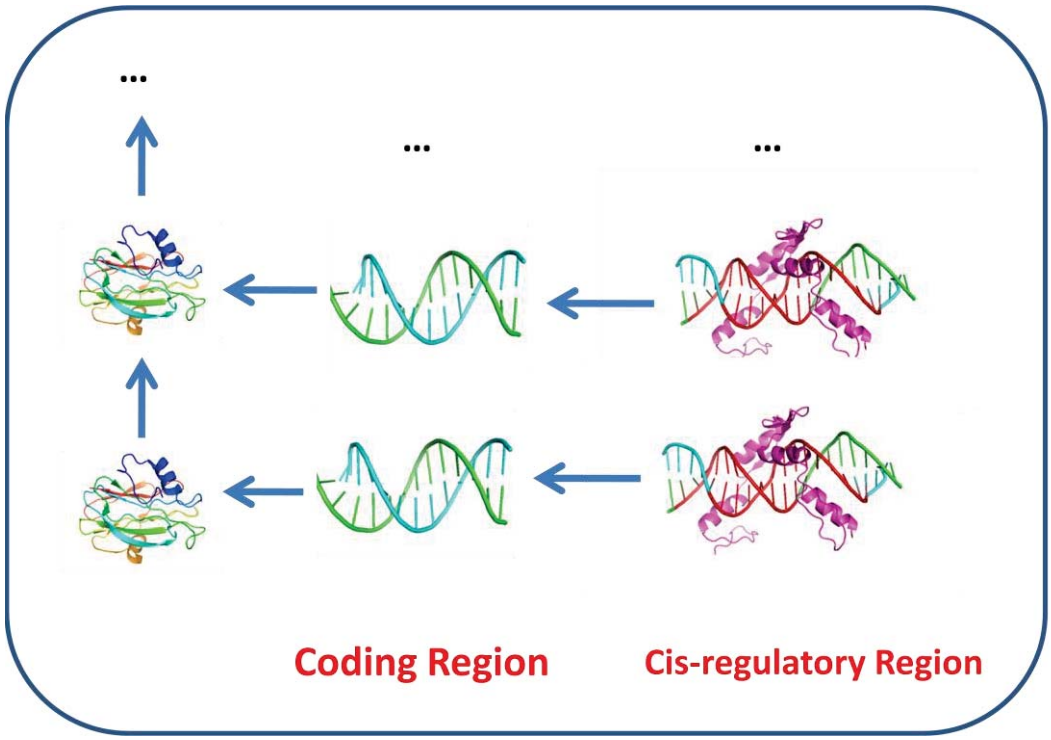


Figure 2.3 A schematic representation of the model. A model organism has five genes, which are expressed into multiple copies of proteins. Each gene has a corresponding upstream cis-regulatory region which determines the expression level of a corresponding gene.

The fate of the cell is described by two quantities: the birth rate and death rate. The birth rate is given by Equation 2.2:

$$b = \prod_{i=1}^{\Gamma} \min\left(C_i P_i^{nat}, C_{0i}\right) \tag{2.2}$$

where Γ is the number of essential genes, C_i denotes the total abundance of the protein encoded by gene i, P_i^{nat} is the probability that protein i is in its native state, and C_{0i} is

the protein abundance required to maintain biological function. The rationale behind this birth rate function is as follows: first, a cell needs to have a specific minimal number of functional proteins to perform its biological functions, and all essential proteins need to be correctly folded for the organism to survive. Therefore, the birth function takes a multiplicative form. Second, the contribution of protein abundance to the birth rate is saturable;[39] therefore, we take the minimum of the real, correctly folded protein abundance and the required protein abundance. This type of diminishing benefit behavior is often observed in enzymes.[39, 40]

In vivo, unfolded proteins may aggregate and poison the organism. To capture this effect, we also include a death rate function, which is given by Equation 2.3:

$$d = (\sum_i C_i(1 - P_i^{nat})) + d_0 \tag{2.3}$$

where C_i denotes the protein abundance coded by gene i. $1 - P_i^{nat}$ denotes the probability that protein i in its unfolded state, and d_0 is the natural death rate, which is caused by environmental factors other than protein stability.

In our model, the protein abundance levels are also evolvable. Each gene in our model is associated with an upstream regulatory sequence. Different regulatory sequences exhibit different binding affinities to regulatory proteins. We assume that the protein abundance levels are determined by the affinities of the binding sites in their regulatory sequences. If the affinities of the binding sites in the regulatory sequences increase, then more mRNAs will be expressed, and the protein abundance will increase. This quantitative relationship has been discussed thoroughly in previous studies. [41] The protein abundance is given by Equation 2.4:

$$C_i = C_{0i} f_i \tag{2.4}$$

where C_i is the abundance of the protein encoded by gene i. C_{0i} is the base protein abundance for gene i. and f_i is the fraction of expression relative to the required protein abundance due to different binding affinity. f is given by Equation 2.5:[41]

$$f = \frac{1 + \frac{N_{NS}}{P} e^{\Delta\varepsilon_{pd}/k_B T}}{1 + \frac{N_{NS}}{P} e^{(\Delta\varepsilon_{pd} + \Delta\Delta\varepsilon_{pd})/k_B T}} \tag{2.5}$$

where N_{NS} is the number of non-specific binding sites for DNA-binding proteins, e.g., TF, P is the number of DNA-binding protein molecules, k is the Boltzmann's constant, T is the temperature, $\Delta\varepsilon_{pd}$ is the wild-type binding affinity of DNA-binding protein to the DNA, and $\Delta\Delta\varepsilon_{pd}$ is the change in the binding affinity of the DNA-binding protein to DNA due to mutation in the regulatory sequence.

Mutations can occur in both coding and regulatory sequences. The details of the coding sequence mutation are explained elsewhere.[31] Briefly, the coding sequence mutations can be divided into two groups: synonymous mutations and nonsynonymous mutations. The nonsynonymous mutations can be further divided into two types, namely, unconditionally lethal mutations, which introduce STOP codons or disrupt critical residues, and other nonsynonymous mutations, which change proteins' thermodynamic stability by a value ($\Delta\Delta G$) that is drawn from a Gaussian distribution.[42] The mean value of the $\Delta\Delta G$ distribution is dependent on ΔG, i.e., there is a sequence depletion effect.

Mutations can also occur in the regulatory sequence, which will change the

affinities of the binding sites in the regulatory sequence for the corresponding regulatory factors and therefore change the abundance of the protein encoded by the gene. The change in binding energy, $\Delta\Delta\varepsilon_{pd}$, is drawn from a Gaussian distribution with a mean of +1 kcal/mol and an SD of 2 kcal/mol.[43] The range of binding energies used in this study allows the expression level to vary across several orders of magnitude due to mutations. A large variation in expression level due solely to the effects of mutation is observed in mutation accumulation experiment.[44]

The selection coefficient for the evolution model, incorporating both birth rate and death rate, is given by Equation 2.6:[45]

$$\nu = \frac{b_f / d_f}{b_i / d_i} = 1 + s \tag{2.6}$$

where s is the selection coefficient, and i and f denote the initial state and the final state. The substitution probability that mutations fixed in the population in a monoclonal regime from state i to state j is given by Equation 2.7:[45]

$$\pi(i \to j) = \frac{1 - \nu^{-1}}{1 - \nu^{-N}} \tag{2.7}$$

where N is the effective population size. Figure 2.4 shows the dependence of the fixation probability on different selection coefficients in populations of different sizes.

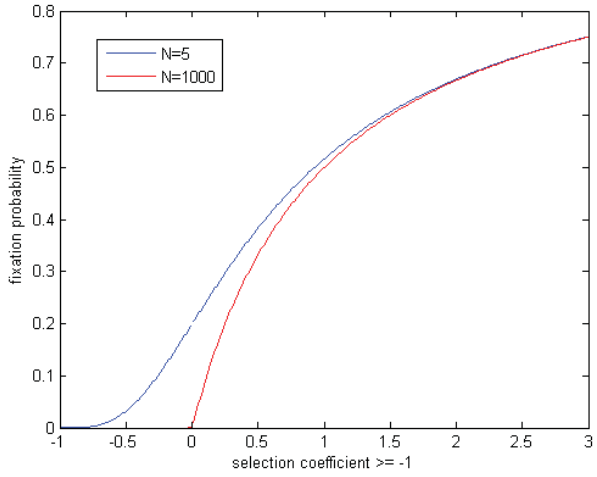


Figure 2.4 The dependence of probability of a mutation to fix in the population on selection coefficient for two population sizes N = 5 and N = 1000.

We generate twenty parallel evolution trajectories starting from a common ancestor, determine the final states of each evolution trajectory after mutation-selection equilibration and plot each protein's ΔG and abundance. The results of the simulation recapitulate well-known previous results and are consistent with the protein-misfolding hypothesis, demonstrating that highly expressed proteins are stable and evolve relatively slowly[2, 17] (Figure 2.5A and Figure 2.5B). This result suggests that our model is consistent with the "misfolding avoidance hypothesis". Figure 2.6 shows the evolution of protein abundances for different genes. The protein abundances for all genes quickly reach plateaus at approximately one order of magnitude above the required level, after which they fluctuate around that level. This demonstrates that the birth rate plays a more important role than the death rate in the early stage of evolution when insufficient amounts of functional proteins are produced. Therefore, selection acts mainly on the amount of correctly folded proteins. After

equilibrium, the organism has sufficient amount of functional protein, and the death rate becomes the dominant factor determining evolutionary dynamics. Because selection acts mainly on misfolded proteins, highly expressed proteins are under stronger selection against misfolding; therefore, they are more stable and evolve more slowly than do proteins with lower levels of expression. The simulation shows that all essential proteins are expressed above the required levels. This phenomenon was observed in previous experiments on enzymes.[46, 47]

A

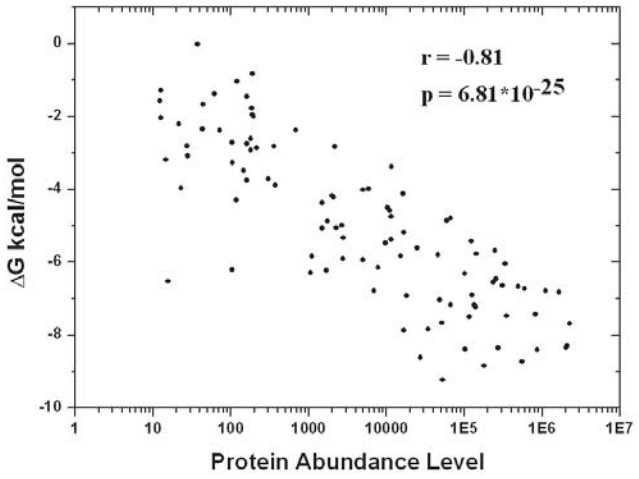


Figure 2.5A is the simulation result of the negative correlation between protein abundance level and ΔG.

B

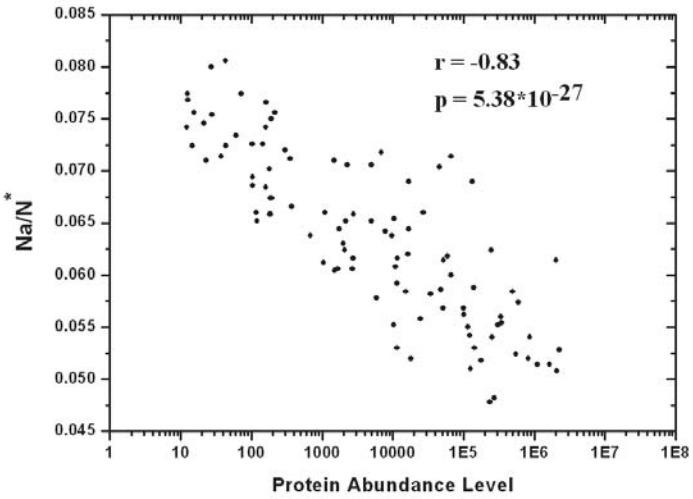


Figure 2.5B is the simulation result of the negative correlation between the number of accepted nonsynonymous mutations (Na) normalized by the expected number of accepted substitutions under neutral case (N^*) with acceptance probability of 1/N, where N is the effective population size. Correlation coefficients and significance levels are determined by Spearman's rank correlation test.

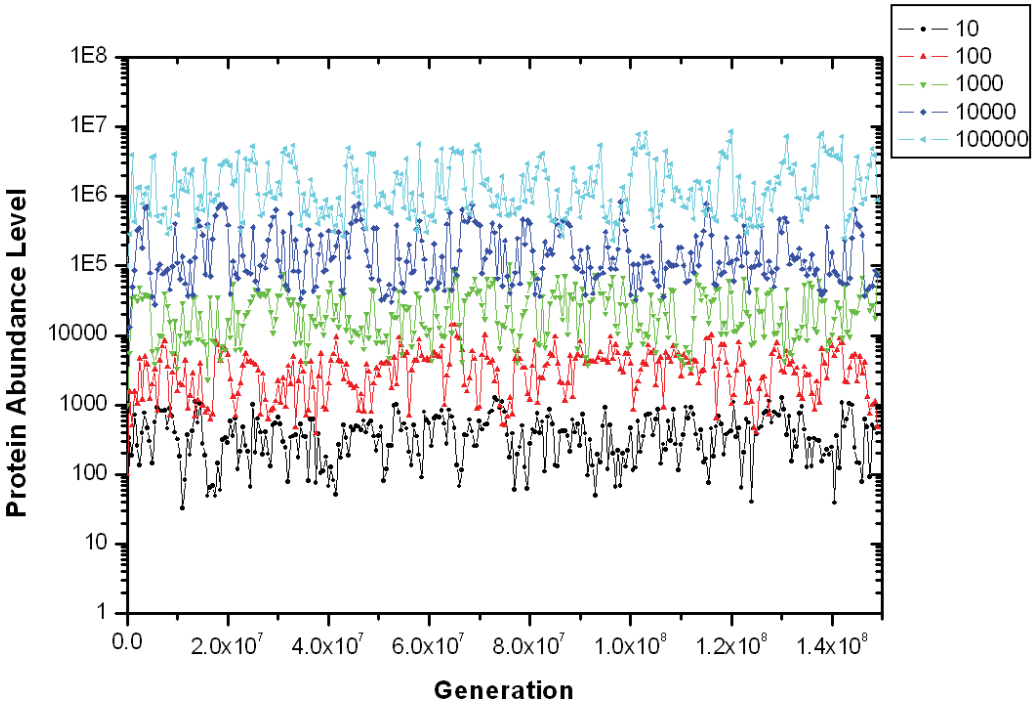
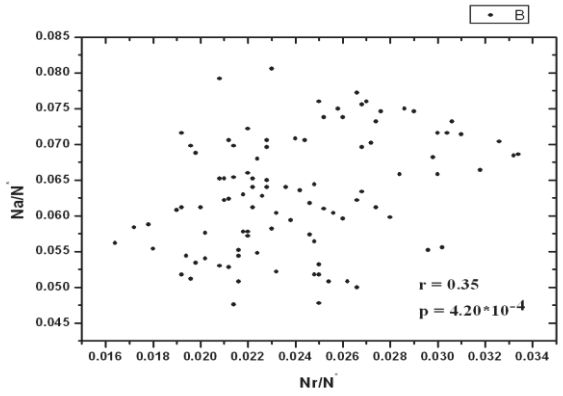


Figure 2.6 Evolution of protein abundance levels for five genes. On the top right of the plot shows the required protein abundance for each gene.

After reproducing the previously reported results, our next goal is to study the co-evolution of regulatory and protein coding sequences. Figure 2.7A shows that there is a positive correlation between the number of accepted nonsynonymous substitutions (normalized to the expected number of accepted neutral substitutions in the coding sequence) and the number of accepted substitutions in the regulatory sequence (normalized to the expected number of accepted neutral substitutions in the regulatory sequence). In addition, Figure 2.7B shows that protein abundance is negatively correlated with the normalized number of accepted substitutions in the regulatory sequence in our simulation.

A



B

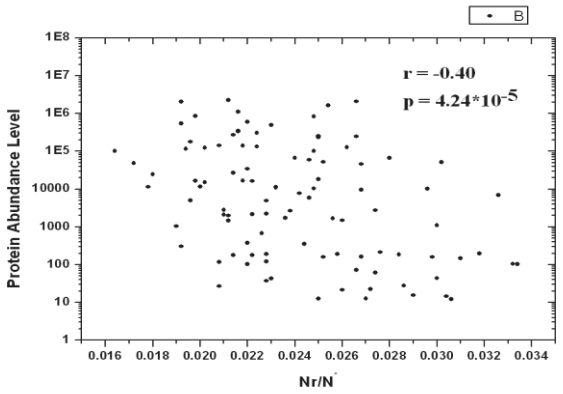


Figure 2.7A The simulation result of the positive correlation between the normalized number of accepted nonsynonymous mutations (Na/N^*) and the normalized number of accepted mutations in the regulatory sequence (Nr/N^*). The normalizing quantity is the expected number of accepted mutations (N^*) under neutral case with acceptance probability 1/N, where N is the effective population size.

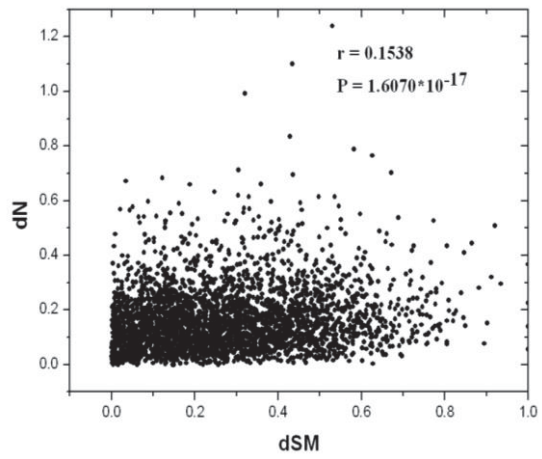
Figure 2.7B The simulation result of the negative correlation between protein abundance level and the normalized number of accepted mutations in the regulatory sequence (Nr/N^*).Correlation coefficients and significance levels are determined by Spearman's rank correlation test.

We used a first-principle biophysical model of living cells to investigate the relationship between regulatory sequence evolution and coding sequence evolution. Our model does not make any a priori phenomenological assumptions about optimal gene expression levels and is based on molecular biophysics. Additionally, our model takes into account the fitness effects of both regulatory and coding sequences. Finally, our model incorporates the evolvability of protein abundance levels.

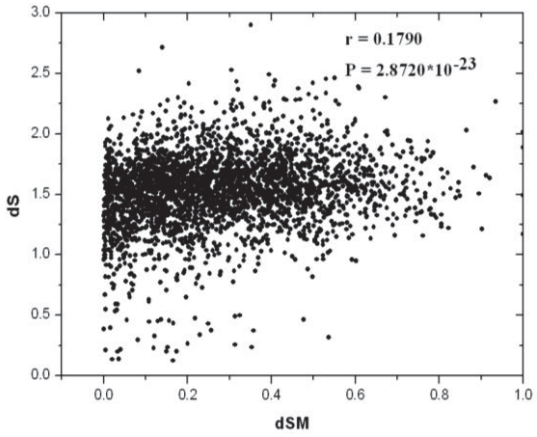
We next used yeast genomic data to test the predictions made by our model. First, we investigate the relationship between gene expression level and regulatory sequence evolution. For regulatory sequence evolution, we use the shared motif divergence (dSM) to characterize the evolution of regulatory sequence.[28] The definition of dSM is provided in the Methods. dSM ranges from zero to one, with higher dSM values indicating greater divergence between the two sequences. We use comparative genomics data from four yeast species (*S. cerevisiae, S. paradoxus, S. mikatae* and *S. bayanus*) to calculate the dN, the number of synonymous substitutions per site (dS) and the dSM of the regulatory sequence.[37] Figure 2.8A shows that there is a statistically significant correlation between dN and dSM (Spearman's correlation coefficient = 0.1538; P value = 1.6070×10^{-17}). The statistically significant positive correlation of dN and dSM provides strong empirical evidence to support our simulation finding that the regulatory sequence evolution rate and coding sequence evolution rate are correlated. Furthermore, we investigated the correlation between dS and dSM. Figure 2.8B shows that there is a positive correlation between dS and dSM (Spearman's correlation coefficient = 0.1790; P value = 2.8720×10^{-23}). This is not

surprising because previous studies have suggested that codon usage bias plays an important role in coding sequence evolution.[2, 18] Figure 2.8C shows that there is a strong positive correlation between dN/dS and dSM (Spearman's correlation coefficient = 0.1177; P value = 7.8178×10^{-11}). This observation suggests that a strong correlation between coding sequence evolution rate and regulatory sequence evolution rate persists even when we normalize the number of nonsynonymous substitutions per site to the number of synonymous substitutions per site and control the difference in mutation rate among different coding sequences, consistent with our simulation results. Our model also predicts that there should be a negative correlation between protein abundance and the number of accepted substitutions in the regulatory sequence; Figure 2.8D shows that there is a negative correlation between gene expression level and dSM (Spearman's correlation coefficient = -0.0966; P value = 4.1125×10^{-10}). These bioinformatics results support the conclusions from our model simulations.

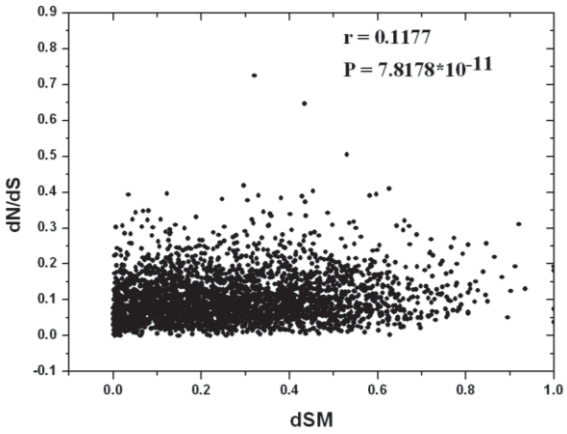
A



B



C



D

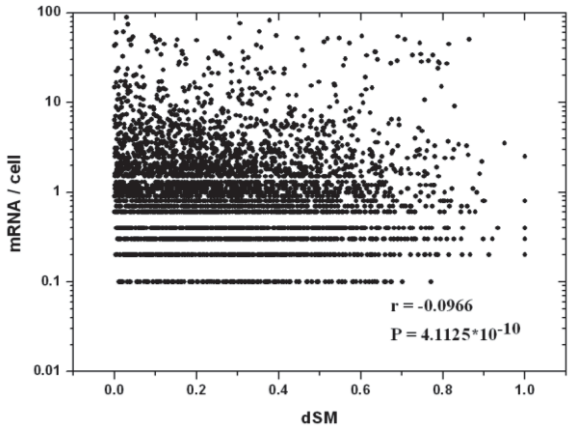


Figure 2.8A is the yeast empirical result of the correlation between shared motif divergence (dSM) and nonsynonymous substitution rate per site (dN). Figure 2.8B is the yeast empirical result of the correlation between shared motif divergence (dSM) and synonymous substitution rate per site (dS). Figure 2.8C is the yeast empirical result of the correlation between shared motif divergence (dSM) and dN/dS. Figure 2.8D is the yeast empirical result of the correlation between gene expression levels and the shared motif divergence (dSM). Correlation coefficients and significance (Continued) levels are determined by Spearman's rank correlation test.

The quantity that we use to measure the evolution of regulatory sequences (dSM) considers the entire regulatory sequence as well as the orientation and translocation effects of the binding sites. Our next question is whether the negative correlation between expression level and regulatory sequence evolution rate will be preserved if we focus on individual TF binding sites. We use several experimentally validated TF binding sites from the *Saccharomyces cerevisiae* promoter database to calculate the rate of evolution within these binding sites.[35] The rate of evolution for each site in the TF binding sites is calculated using the classical parsimony algorithm and the accepted species tree (Sbay, Smik, (Spar, Scer)).[37,38] The evolution rate of a binding site is the sum of the evolution rate for each position in the binding site divided by the total length of the binding site. Figure 2.9A shows the scatter plot of the average evolution rate of a TF binding site and the expression level of the downstream (regulated) gene. Our results show that there is no significant correlation between the gene expression level and the average evolution rate of the TF binding site (Spearman's correlation coefficient = 0.0957; P value = 0.3317). Figure 2.9B shows the scatter plot of the average evolution rate of a TF binding site and the protein abundance of the downstream gene. This result also shows that there is no significant correlation between the protein abundance level and the average evolution rate of the TF binding sites (Spearman's correlation coefficient = -0.0284; P value = 0.8022). Although our empirical results show that there is no significant correlation between TF binding site evolution rates and gene expression levels or protein abundance levels,

the number of experimentally verified TF binding sites is relatively small. Therefore, it is possible that a different conclusion might be reached when more experimentally verified TF binding sites are available for analysis.

A B

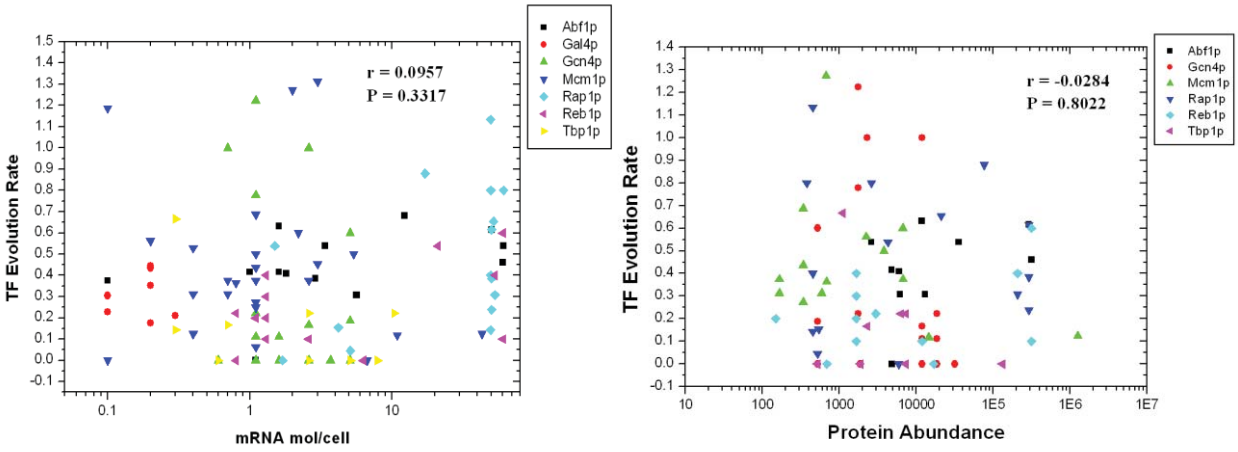


Figure 2.9A shows the correlation between the Transcription Factor binding site evolution rate and the gene expression level from yeast. Figure 2.9B shows the lack of statistically significant correlation between the Transcription Factor binding site evolution rate and the protein abundance level from yeast. Correlation coefficients and significance levels are determined by Spearman's rank correlation test.

2.5 Discussion

In this work, we used a first-principle biophysical model of living cells to investigate the relationship between regulatory sequence evolution and coding sequence evolution. Previous studies using phenomenological modeling generally

make a priori assumptions about the optimal levels of gene expression.[27, 48, 49] Other studies using biophysical models consider the fitness of either regulatory sequences or coding sequences separately.[17, 18, 25] Our model does not make any a priori phenomenological assumptions about optimal gene expression levels and is based on molecular biophysics. Additionally, our model takes into account the fitness effects of both regulatory and coding sequences. Finally, our model incorporates the evolvability of protein abundance levels through the binding energies of the sites in the regulatory sequence. Our model reproduces the important result that highly expressed proteins are stable and evolve relatively slowly.[2, 13, 17, 18] In addition, our model predicts that the coding sequence evolution rate and regulatory sequence evolution rate are positively correlated and that protein expression level is negatively correlated with regulatory sequence evolution rate. Furthermore, the results of an empirical study using comparative yeast genomic data were consistent with these predictions.

Our results show that, during evolution, protein abundance will exceed the biologically required level, leading to a surplus of protein. The increase in protein abundance will maximize organismal fitness because the birth rate is multiplicative while the death rate is additive. The "over-expression" of enzymes has been observed in studies on metabolic flux and enzyme amount,[46, 47] metabolic flux in the arginine biosynthetic pathway in *Neurospora crassa*[50, 51] and ethanol flux in relation to alcohol dehydrogenase activity in *Drosophila melanogaster*.[52] In addition, our simulation shows that protein abundance will reach a plateau after equilibrium. This is consistent

with previous studies showing that binding sites are generally under stabilizing selection to produce the correct expression level; both stronger and weaker binding affinities produce less optimal functionality.[53]

Our model is consistent with previous results showing that highly expressed proteins are more stable and that highly expressed genes evolve relatively slowly.[2, 13, 17, 18] Previous studies have proposed that highly expressed genes evolve slowly due to the toxic effects of protein misfolding due to transcriptional and translational errors. [18] Another study showed that the above two phenomena are also consequences of transcriptional and translational error-free misfolding, supporting a general misfolding avoidance hypothesis.[17] In our model, we do not explicitly consider the translational process, but we still observe a strong correlation between expression levels and number of nonsynonymous substitution, which is consistent with the results produced by the transcriptional and translational error-free misfolding model. Furthermore, our model is distinct from previous models in that we explicitly consider the benefits produced by having a certain number of functional proteins, while previous studies focused only on the toxic effects of misfolded proteins. Additionally, our model allows protein abundance to fluctuate, while previous studies have examined only fixed protein abundances. In previous studies, the misfolded protein was always derived from a single source, either mistranslation-induced misfolding or translational error-free misfolding due to protein thermodynamic stability; in contrast, our model introduces two sources of misfolded proteins (the protein thermodynamic stability and the raw protein abundance level). Therefore, our model shows that the above two

relationships still hold under a more general assumption of evolvable protein abundance. Our simulation demonstrates the trade-off effect between the beneficial effects of functional proteins and the deleterious effects of toxic misfolded proteins.

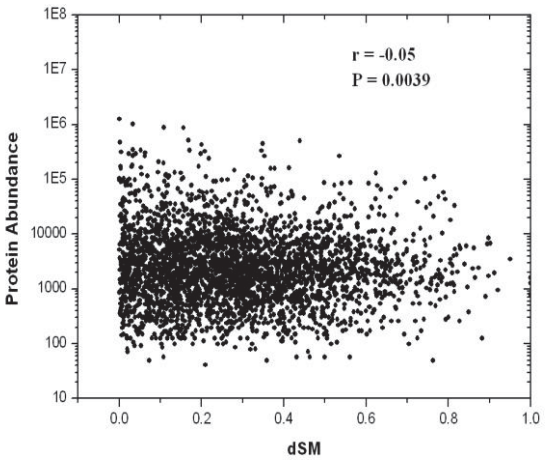
The major finding of our simulations is that the number of substitutions in the regulatory region and the number of non-synonymous substitutions in the protein coding region are correlated. Previous studies on cis-regulatory and protein evolution in orthologous and duplicate genes of *Caenorhabditis elegans* and *C. briggsae* showed that there is a positive correlation between functional regulatory evolution and protein evolution in orthologous genes.[28] Our empirical study of data from four yeast species also shows that there is a significant correlation between regulatory evolution and protein evolution in orthologous genes. This suggests that natural selection acts on a gene and its upstream regulatory sequence as an integrated unit. Although our model accurately demonstrates the correlation between protein evolution rate and regulatory sequence evolution, the correlation coefficient in our model is stronger than the empirical result obtained from the comparative genomic studies of four yeast species. One possible reason that a stronger correlation is observed in our model is that we do not consider the trans-effects of regulation. In reality, when mutations occur in cis-regulatory sequences, they are often accompanied by compensatory mutations in the TF binding sites to maintain the binding affinity. Previous experiments have demonstrated that in at least two genetic loci in *Drosophila*, bicoid and even-skipped, TFs and their DNA binding sites in cis-regulatory sequences have co-evolved, and promoter structure rearrangement has

occurred to maintain a stable gene expression pattern.[22, 54] Another possible factor that might contribute to the difference in the magnitude of the correlation coefficient between our model and empirical study is that different cis-regulatory sequences might have the same binding affinities for a TF even in the absence of a compensatory mutational effect. We refer to this phenomenon as a "sequence degeneracy" effect. Recent high-throughput experiments that systematically measured the binding energy landscape of transcription factors showed that different TF binding site sequences within a cis-regulatory sequence can have similar binding energies.[43] Furthermore, Tirosh et al. found that the majority of the previously identified differences in TF-binding sequences between yeasts and mammals have no detectable effect on gene expression.[55] Because our phenomenological model considers only the effect of the cis-regulatory sequence on binding energy, there might be a discrepancy between model prediction and empirical data.

The final prediction made by our model is that highly expressed proteins should have slowly evolving regulatory regions. Empirical studies have shown a significant correlation between the mRNA level of a gene in *S. cerevisiae* cells and its dSM. We also compare the protein abundance level versus the dSM in our model (Figure 2.10A). The correlation between the protein abundance and dSM remains significant, but the correlation is weaker and less significant than that between the gene expression level and dSM. There are several potential reasons for this difference. First, gene expression data is more readily available than protein abundance data. This lack of data might cause the lower significance and lower correlation. Second, there are

many complex biological processes involved in the process of mRNA translation to protein; these are not captured in our simple, minimalist model, which might cause a difference in the correlations of protein abundance and gene expression level with dSM. Finally, in our model, all proteins are essential proteins. Therefore, our model predicts that essential protein abundance, rather than all protein abundance, should negatively correlate with dSM. When we plot the correlation between dSM and essential protein abundance in *S. cerevisiae*, there is an increased and more significant correlation than when all proteins are considered (Spearman's correlation coefficient = -0.1898; P value = 5.4945×10^{-7}) (Figure 2.10B). Thus, the empirical results are in good agreement with the predictions of our model.

A



B

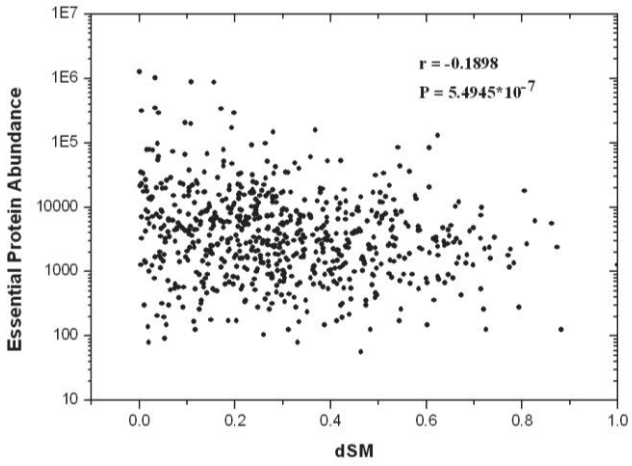


Figure 2.10A shows the yeast empirical result of the correlation between protein abundance level and the shared motif divergence (dSM). Figure 2.10B shows the yeast empirical result of the correlation between essential protein abundance level and the shared motif divergence (dSM). Correlation coefficients and significance levels are determined by Spearman's rank correlation test.

Although our model captures many realistic biophysical aspects of protein coding sequence and cis-regulatory sequence evolution, it remains a minimalistic model and can be improved in the future. First, our model focuses on cis-regulatory region evolution, while real cells can also have compensatory mutations in DNA-binding proteins that will maintain the corresponding gene expression levels and mitigate the deleterious effects caused by mutations in the cis-regulatory region. Second, our model does not explicitly model the translation process. Therefore, our model cannot capture the effects of synonymous substitutions on evolution. In vivo,

because of the effects of codon usage bias, synonymous substitutions also influence the protein abundance.[18] Third, we do not explicitly consider the effect of chaperones in mitigating the deleterious effects on fitness caused by destabilizing mutations in protein coding sequences. Fourth, we do not include protein-protein interactions in our model.[56, 57] Fifth, our model does not include the epistatic effects of regulatory sequence evolution due to a lack of experimentally verified quantitative data. Recent studies have shown that the different positions within a binding site do not evolve independently.[58-60] The effect of a change in binding affinity on fitness depends on the initial binding affinity of the binding site. This is the same as the "sequence depletion" effect that we used for the coding sequence. We were able to use ProTherm data to derive a quantitative relationship between $\Delta\Delta G$ and ΔG, but we did not have enough experimental data to derive the same relationship for the regulatory sequences.[32] Nevertheless, our model is able to reproduce important correlations between regulatory sequence and coding sequence evolution.

2.6 Conclusion

In this study, we developed a biophysically realistic model of protein folding and the binding of protein to DNA. Using this model and population genetics simulations, we investigated the co-evolution of protein coding and regulatory sequences. First, our simulation showed that there is a negative correlation between protein abundance and protein stability, which is consistent with previous simulation results. Additionally, we reproduced the well-known finding that highly expressed proteins evolve slowly,

showing that the accepted number of nonsynonymous mutations is negatively correlated with protein abundance level. Furthermore, our simulation shows that there is a positive correlation between the normalized number of accepted nonsynonymous mutations and the normalized number of accepted mutations in the regulatory sequence, demonstrating that there is a positive correlation between the evolution rate of a coding sequence and the evolution rate of its regulatory sequence. In addition, our simulation showed that there is a negative correlation between the protein abundance and the normalized number of accepted mutations in the regulatory sequence, similar to the behavior exhibited by the coding sequence. We also performed a bioinformatic analysis using empirical sequence data from yeast, which showed that there is a positive correlation between the shared motif divergence and the nonsynonymous substitution rate per site. Additionally, we found a negative correlation between the essential protein abundance level and the shared motif divergence, demonstrating that the evolution rate of the regulatory sequence is negatively correlated with protein abundance level, as is the evolutionary rate of the coding sequence.

2.7 Reference

1. Zuckerkandl E & Pauling L (1965) Evolutionary divergence and convergence in proteins (Academic Press, New York).

2. Drummond DA, Bloom JD, Adami C, Wilke CO, & Arnold FH (2005) Why highly expressed proteins evolve slowly. Proc Natl Acad Sci 102(40):14338-14343.

3. Ohta T & Kimura M (1971) On the constancy of the evolutionary rate of cistrons.

J. Mol. Evol. 1:18-25.

4. Wall DP, et al. (2005) Functional genomic analysis of the rates of protein evolution. Proc Natl Acad Sci 102(15):5483-5488.

5. Hirsh AE & Fraser HB (2001) Protein dispensability and rate of evolution. Nature 411:1046-1049.

6. Jordan IK, Rogozin IB, Wolf YI, & Koonin EV (2002) Essential genes are more evolutionarily conserved than are nonessential genes in bacteria. Genome Res. 12:962-968.

7. Zhang J & He X (2005) Significant impact of protein dispensability on the instantaneous rate of protein evolution. Mol Biol Evol 22:1147-1155.

8. Fraser HB, Hirsh AE, Steinmetz LM, Scharfe C, & Feldman MW (2002) Evolutionary rate in the protein interaction network. Science 296:750-752.

9. Lemos B, Bettencourt BR, Meiklejohn CD, & Hartl DL (2005) Evolution of Proteins and Gene Expression Levels are Coupled in Drosophila and are Independently Associated with mRNA Abundance, Protein Length, and Number of Protein-Protein Interactions. Mol. Biol. Evol. 22:1345-1354.

10. Marais G & Duret L (2001) Synonymous codon usage, accuracy of translation, and gene length in Caenorhabditis elegans. J. Mol. Evol. 52:275-280.

11. Hahn MW & Kern AD (2005) Comparative genomics of centrality and essentiality in three eukaryotic protein-interaction networks. Mol Biol Evol 22:803-806.

12. Pal C, Papp B, & Hurst LD (2001) Highly expressed genes in yeast evolve slowly.

Genetics 158:927-931.

13. Drummond DA, Raval A, & Wilke CO (2006) A single determinant dominates the rate of yeast protein evolution. Mol Biol Evol 23:327-337.

14. Bloom JD, Drummond DA, Arnold FH, & Wilke CO (2006) Structural Determinants of the Rate of Protein Evolution in Yeast. Mol Biol Evol 23(9):1751-1761.

15. Ramsey DC, Scherrer MP, Zhou T, & Wilke CO (2011) The relationship between relative solvent accessibility and evolutionary rate in protein evolution. Genetics 188:479-488.

16. Toth-Petroczy A & Tawfik DS (2011) Slow protein evolutionary rates are dictated by surface-core association. Proc Natl Acad Sci 108:11151-11156.

17. Yang J-R, Zhuang S-M, & Zhang J (2010) Impact of translational error-induced and error-free misfolding on the rate of protein evolution. Mol Syst Biol 5:421.

18. Drummond DA & Wilke CO (2008) Mistranslation-induced protein misfolding as a dominant constraint on coding-sequence evolution. Cell 134(2):341-352.

19. Moses AM, Chiang DY, Pollard DA, Iyer VN, & Eisen MB (2004) MONKEY: identifying conserved transcription-factor binding sites in multiple alignments using a binding site-specific evolutionary model. Genome Biol. 5:R98.

20. Gasch AP, et al. (2004) Conservation and Evolution of Cis-Regulatory Systems in Ascomycete Fungi. PLoS Biol. 2:e398.

21. Dickinson WJ (1988) On the architecture of regulatory systems: evolutionary insights and implications. BioEssays 8:204-208.

22. Ludwig MZ, Bergman C, Patel NH, & Kreitman M (2000) Evidence for stabilizing selection in a eukaryotic enhancer element. Nature 403:564-567.

23. Ludwig MZ, et al. (2005) Functional evolution of a cis-regulatory module. PLoS Biol. 3:e93.

24. Moses AM, Chiang DY, Kellis M, Lander ES, & Eisen MB (2003) Position specific variation in the rate of evolution in transcription factor binding sites. BMC Evol Biol 3:1471-2148.

25. Mustonen V, Kinney J, Jr. CGC, & Lassig M (2008) Energy-dependent fitness: a quantitative model for the evolution of yeast transcription factor binding sites. Proc. Natl. Acad. Sci. 105:12376-12381.

26. Kim J, He X, & Sinha S (2009) Evolution of Regulatory Sequences in 12 Drosophila Species. PLos Genet. 5(1):e1000330.

27. Berg J, Willmann S, & Lassig M (2004) Adaptive evolution of transcription factor binding sites. BMC Evol. Biol. 4(42):1471-2148.

28. Castillo-Davis CI, Hartl DL, & Achaz G (2004) cis-Regulatory and protein evolution in orthologous and duplicate genes. Genome Research 14:1530-1536.

29. Xu L, et al. (2006) Average Gene Length Is Highly Conserved in Prokaryotes and Eukaryotes and Diverges Only Between the Two Kingdoms. Mol. Biol. Evol. 23(6):1107-1108.

30. Patthy L (2008) Protein Evolution (Blackwell, Oxford) 2nd Ed.

31. Wylie CS & Shakhnovich EI (2011) A biophysical protein folding model account for most mutational fitness effects in viruses. Proc Natl Acad Sci 108(24):9916-9921.

32. Kumar MD, et al. (2006) ProTherm and ProNIT: thermodynamic databases for proteins and protein-nucleic acid interactions. Nucleic Acids Res. 34:D204-206, Database issue.

33. Holstege FCP, et al. (1998) Dissecting the Regulatory Circuitry of a Eukaryotic Genome. Cell 95:717-728.

34. Ghaemmaghami S, et al. (2003) Global analysis of protein expression in yeast. Nature 425:737-741.

35. Zhu J & Zhang MQ (1999) SCPD: a promoter database of the yeast Saccharomyces cerevisiae. Bioinformatics I 5(7-8):871-877.

36. Durbin R, Eddy S, Krogh A, & Mitchison G (1998) Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids (Cambridge University Press, Cambridge, UK).

37. Kellis M, Patterson N, Endrizzi M, Birren B, & Lander ES (2003) Sequencing and comparison of yeast species to identify genes and regulatory elements. Nature 423:241-254.

38. Cliften PF, et al. (2001) Surveying Saccharomyces genomes to identify functional elements by comparative DNA sequence analysis. Genome Res. 11(7):1175-1186.

39. Bershtein S, Mu W, Serohijos AWR, Zhou J, & Shakhnovich EI (2013) Protein Quality Control Acts on Folding Intermediates to Shape the Effects of Mutations on Organismal Fitness. Mol. Cell. 49:133-144.

40. Hartl DL, Dykhuizen DE, & Dean AM (1985) Limits of adaptation: the evolution of selective neutrality. Genetics 111:655-674.

41. Bintu L, et al. (2005) Transcriptional regulation by the numbers: models. Curr. Opin. Genet. Dev. 15:116-124.

42. Zeldovich KB, Chen P, & Shakhnovich EI (2007) Protein stability imposes limits on organism complexity and speed of molecular evolution. Proc Natl Acad Sci 104:16152-16157.

43. Maerkl SJ & Quake SR (2007) A Systems Approach to Measuring the Binding Energy Landscapes of Transcription Factors. Science 315:233-237.

44. Landry CR, Lemos B, Rifkin SA, Dickinson WJ, & Hartl DL (2007) Genetic properties influencing the evolvability of gene expression. Science 317:118-121.

45. Parsons TL & Quince C (2007) Fixation in haploid populations exhibiting density dependence I: The non-neutral case. Theoretical Population Biology 72:121-135.

46. Kacser H & Burns JA (1973) The control of flux. Symp. Soc. Exp. Biol. 32:65-104.

47. Kacser H & Burns JA (1981) The molecular basis of dominance. Genetics 97:639-666.

48. Gerland U & Hwa T (2002) On the selection and evolution of regulatory DNA motifs. J. Mol. Evol. 55:386-400.

49. Gout J-F, Kahn D, & Duret L (2010) The relationship among gene expression, the evolution of gene dosage, and the rate of protein evolution. PLoS Genet 6(5):e1000944-e1000952.

50. Flint HJ, Porteous DJ, & Kacser H (1980) Control of flux in the arginine pathway of Neurospora crassa: the flux from citrulline to arginine. Biochem. J. 190:1-15.

51. Flint HJ, et al. (1981) Control of flux in the arginine pathway of Neurospora crassa: modulations of enzyme activity and concentration. Biochem. J. 200:231-246.

52. Middleton RJ & Kacser H (1983) Enzyme variation, metabolic flux and fitness: alcohol dehydrogenase in Drosophila melanogaster. Genetics 105:633-650.

53. Ochoa-Espinosa A, et al. (2005) The role of binding site cluster strength in Bicoid-dependent patterning in Drosophila. Proc Natl Acad Sci 102:4960-4965.

54. Shaw PJ, Wratten NS, McGregor AP, & Dover GA (2002) Coevolution in bicoid-dependent promoters and the inception of regulatory incompatibilities among species of higher Diptera. Evol. Dev. 4:265-277.

55. Tirosh I, Weinberger A, Bezalel D, Kaganovich M, & Barkai N (2008) On the relation between promoter divergence and gene expression evolution. Mol Syst Biol 4:159.

56. Heo M, Kang L, & Shakhnovich EI (2009) Emergence of species in evolutionary "simulated annealing.". Proc Natl Acad Sci 106:1869-1874.

57. Heo M, Maslov S, & Shakhnovich EI (2011) Topology of protein interaction network shapes protein abundances and strengths of their functional and nonspecific interactions. Proc Natl Acad Sci 108(10):4258-4263.

58. Halpern A & Bruno W (1998) Evolutionary distances for protein-coding sequences: modeling site-specific residue frequencies. Mol Biol Evol 15:910-917.

59. Sinha S, van Nimwegen E, & Siggia ED (2003) A probabilistic method to detect regulatory modules. Bioinformatics 19 Suppl(1):i292-301.

60. Siddharthan R, Siggia ED, & van Nimwegen E (2005) PhyloGibbs: a Gibbs

sampling motif finder that incorporates phylogeny. PLoS Comput Biol 1:e67.

Chapter 3

A Biophysical Model for Mutation and Recombination

3.1 Abstract

We present a comprehensive study of the effects of mutation and recombination on protein evolution. We use two biophysical protein folding models to quantify the genotype-phenotype relationship, i.e., between protein sequence and protein folding stability. The two protein folding models used are the Eris model and the lattice model. The Eris model uses realistic protein structures and physical force fields to calculate the protein folding stability. The lattice model takes into account the long-range interactions between amino acids. In addition, our realistic biophysical representation of protein sequences and structures naturally include the fitness changes caused by mutation and recombination. For example, epistatic effects and the sequence depletion effect are naturally incorporated in our model. Furthermore, our model allows intragenic recombination as well as intergenic recombination, which is a significant improvement compared to traditional population genetics models. We assess the effects of mutation and recombination on the adaptation dynamics, the protein equilibrium thermodynamic effects and fixation of recombinant alleles by focusing on three parameters: mutation rate, population size and recombination strength. First, we find that recombination increases the adaptation speed and final equilibrium fitness level compared to mutation. Additionally, a high mutation rate can lead to greater adaptation speed but a lower final equilibrium fitness level. Our realistic protein

models also allow us to examine the sequence space; we find that high recombination and mutation rates will increase the sequence entropy. We also observe the "synchronous" rise and fall of sequence entropy when there is only mutation, due to the hitchhiking effect. The time interval between each rise and fall is inversely proportional to the mutation rate. If recombination occurs, we do not observe any synchronous movement. Second, we investigate the equilibrium protein thermodynamic effect. Our model reproduces the previously published result that protein stability decreases with increased mutation rate and increases with population size in the pure-mutation process. In addition, our model shows that recombination increases protein stability and widens the distribution of protein stability compared to the pure-mutation process. The dependence of protein stability on the mutation rate in the presence of recombination can be divided into two categories. When the mutation rate is low, protein stability increases with the increase of mutation rate. However, when mutation rate is high, protein stability decreases with the increase in mutation rate. Finally, our model shows that high recombination strength and large population size increase the fixation probability of the recombinant allele, while a high mutation rate decreases the fixation probability of the recombinant allele.

3.2 Introduction

Mutation and recombination are the two major sources of protein evolution. Studying how mutation and recombination affect the evolution of protein sequence and structure can shed light on protein evolutionary history and the evolution of

evolvability, and may be useful for protein engineering.[1-3] Mutation introduces new variants of the original protein sequences, providing the raw material for evolution. Recombination creates different combinations of segments of protein sequences, providing greater sequence diversity. Almost all forms of life are susceptible to mutations and recombination. Thus, identifying the roles of mutation and recombination in protein evolution is a central goal of evolutionary biology research. Although the empirical bioinformatics analysis of sequence data can shed some light on the effects of mutation and recombination on protein evolution, a complete understanding cannot be gained from empirical analysis because data is available only for sequences that have been affected by natural selection. Therefore, the computer simulation of the evolutionary process is important because evolution itself is dynamic.

Although mutation and recombination both contribute to protein evolution, most previous studies have focused only on the effects of mutation on protein evolution.[4-8] In addition, previous theoretical studies on the evolutionary process using population genetics models have had several disadvantages.[9, 10] First, the traditional population genetics approaches usually assume a fitness landscape in which a single genotype has the highest fitness value, and any genotype deviating from the optimal genotype has lower fitness.[11] This single-peaked fitness landscape does not reflect reality. Indeed, previous studies employing a lattice protein structure model have shown that the fitness landscape might be glass-like, with multiple local maxima.[12] Second, traditional population genetics models use pre-specified parameters to quantify the

selection coefficients for deleterious mutations.[13] However, different mutations generally have different fitness effects in real organisms, and the same mutation might have different fitness effects depending on the genetic background (epistasis).[14] Third, when modeling recombination events, traditional population genetics approaches can only model events in which recombination takes place between different loci or intergenic recombination.[15] This corresponds to a situation in which recombination can only occur between different genes. However, in reality, eukaryotic genes are discontinuous segments of coding DNA that are interrupted by introns, which can be quite long.[16] This suggests that recombination can also take place within genes in real organisms. In fact, one possible function for introns is to increase the rate of recombination within protein-coding genes.[17] In viruses, the chance of intragenic recombination is even higher because viral genomes, especially those of RNA viruses, are extremely compact and contain many overlapping open reading frames. Therefore, a single crossover might actually result in the recombination of multiple proteins. Finally, some other studies have used a two-letter amino-acid alphabet (HP) to represent the interactions between hydrophobic and polar residues.[18, 19] While this reduced representation of proteins saves computational time, it grossly oversimplifies the complexity of the amino acid interaction force fields in a real protein.

Therefore, to overcome the disadvantages of traditional population genetics models, we will use a biophysical model of protein folding with explicit sequences and protein structures to study the effects of mutation and recombination on protein evolution. The biophysical model allows us to consider the roles of mutation and

recombination in the context of a realistic biophysical fitness landscape. Additionally, each mutation in the genome has its own fitness effect, determined by protein folding thermodynamics, so we do not require any a priori assumptions about the fitness effect of a single mutation or epistasis effect. In addition, our use of an explicit sequence enables the analysis of recombination within genes, thus reflecting a more realistic picture of recombination. We want to study the adaptation process, the effect of equilibrium thermodynamics, and the origin and fixation of recombinant alleles during protein evolution, as well as the effects of mutation rate, population size and recombination strength on those processes.

3.3 Models and Methods

Our goal is to bridge the gap between structural biology and population genetics to study problems in evolution. We employ a sequence-based model with explicit protein structure to study the evolution of mutation and recombination. We have a total number of N cells in our model. Each model cell has a genome of Γ genes with explicit protein sequences, each encoding an essential protein with a specific amino acid sequence and a free energy of folding (ΔG_i). In addition, each cell has a recombination modifier allele that determines whether the organism undergoes recombination during reproduction. If the recombination modifier allele is inactive, the organism undergoes only mutation during reproduction; if the recombination modifier allele is active, the organism undergoes both mutation and recombination. Many proteins fold in a two-state manner. The fraction of time spent in the native

state is given by Equation 3.1:

$$P^{nat} = \frac{e^{-\Delta G_i / kT}}{1 + e^{-\Delta G_i / kT}} \tag{3.1}$$

where k is Boltzmann's constant, and T is temperature. We assume that proteins must be in native states to be functional.

The fate of the cell is captured by the birth rate. A cell requires a specific number of functional proteins to perform the biological function. Our birth rate function is given by Equation 3.2:

$$b = \left(\prod_{i=1}^{\Gamma} p_{nat}^{i} \right) \tag{3.2}$$

where Γ is the number of essential genes and P_i^{nat} is the probability that protein i is in its native state.

We consider two different types of evolutionary process: a pure-mutation process and a mutation-recombination process, which are shown in Figure 3.1 and Figure 3.2, respectively. A flow chart summarizing the simulation is shown in Figure 3.3. In the pure-mutation process, we start with a pure-mutation population of haploid organisms of population size N. We evolve the population according to Wright-Fisher model, with no overlapping generations. For each generation, we select individual organisms with a probability of replication that is proportional to their fitness. Each replicating organism undergoes a number of mutations selected from a Poisson distribution. We then evolve the population to reach mutation-selection equilibrium. In the mutation-recombination process, we start with a population of haploid organisms of

population size N. Each organism carries a recombination modifier allele with recombination strength r. For each generation, we choose two distinct organisms with probabilities proportional to their fitness. This process is performed with N/2 replacements to form N new organisms for the next generation. The two chosen organisms undergo both mutation and recombination. The recombination process takes place with probability r. We call r the "recombination strength" because it quantifies the organism's ability to perform recombination. During the process of producing offspring, a random position in the genome is selected, and two parent organisms produce two offspring with exchanged genomes. For example: before recombination we have genome A with segment 1 and segment 2 and genome B with segment 1 and segment 2; then, after recombination, we have two new organisms, with genome A' and B', where A' is composed of segment 1 of genome A and segment 2 of genome B, and B' is composed of segment 1 of genome B and segment 2 of genome A. Then, each offspring genome acquires a random number of mutations drawn from a Poisson distribution with a mean of m. We estimate ΔΔG for proteins with all possible single point mutations to their wild type sequences using Eris.[20, 21] We made a simple assumption of the additivity of the mutational effect of ΔΔG. The ΔG for mutant proteins are the sum of all corresponding single point mutation ΔΔGs and the wild type ΔG, as shown in Equation 3.3:

$$\Delta G_{mut} = \Delta G_{wild} + \sum \Delta\Delta G \tag{3.3}$$

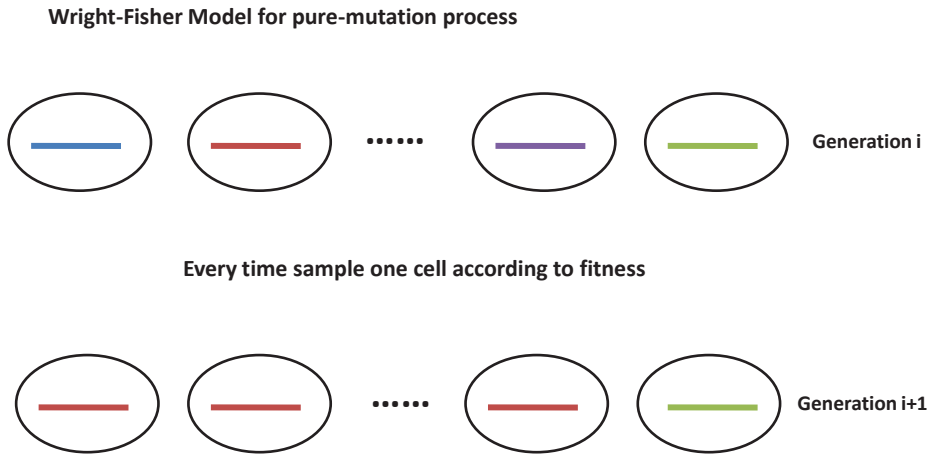


Figure 3.1 An illustration of the Wright-Fisher model for pure-mutation process. In this example, the cell with brown genome has relatively high fitness. The cell thus has more chance to get selected and produces more offsprings in the next generation.

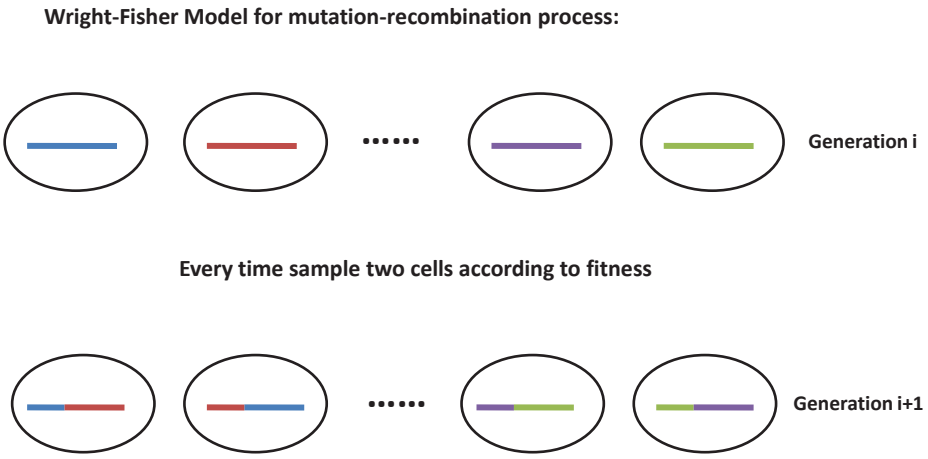


Figure 3.2 An illustration of the Wright-Fisher model for mutation-recombination process. In this example, two cells with blue and brown genomes are sampled according to their fitness level. A random crossover point is located in the genome. Their two offspring have different genotypes. The first offspring cell has its left part of the genome coming from the blue cell and the right part of the genome coming from the brown cell. The second offspring cell has its right part of the genome coming from the brown cell and the left part of the genome coming from the blue cell.

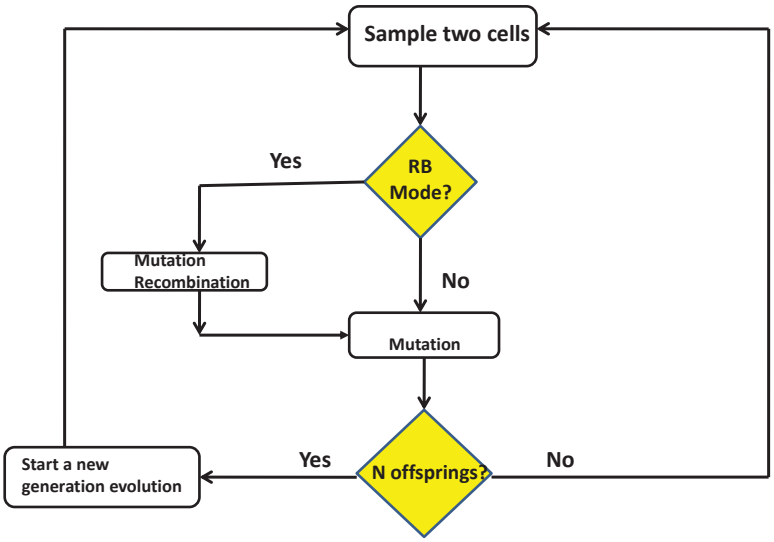


Figure 3.3 Flow chart of the Wright-Fisher model simulation

We also run a control simulation using lattice protein model. First, we perform all single amino acid substitutions on the sequence to obtain the ΔΔG distribution using the Miyazawa-Jernigan (MJ) potential.[22] Then, we rescale the ΔΔG distribution to the empirical ΔΔG distribution by scale transformation. The empirical ΔΔG distribution is obtained from the ProTherm database.[23] Thus, we have a one-one mapping from sequence to free energy. We add the lattice model simulation to examine the effect of epistasis in a protein on the linear additive assumption of ΔG in the Eris-based model.

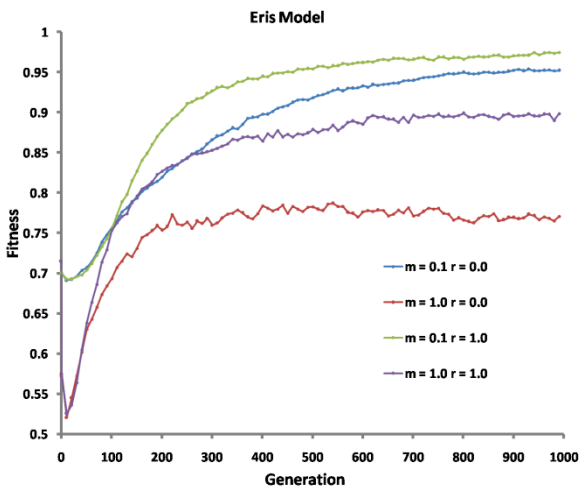
3.4 Results

3.4.1 Adaptation dynamics

We first compare the different adaptation dynamics between the pure-mutation and mutation-recombination processes using a real protein model. Figure 3.4 (A) shows the adaptation of fitness under the Eris-based model. There are several interesting differences in the dynamics between the pure-mutation process and the mutation-recombination process. First, we observed that the mutation-recombination fitness curves have steeper slopes than the pure-mutation fitness curves do, indicating that recombination increases the speed of adaptation. This result is consistent with the generally accepted view that recombination will speed up the adaptation process by exploring the sequence space more efficiently than the pure-mutation process can. [24-26] Beneficial mutations may be united in the same lineage by recombination; while in the pure-mutation process, different beneficial mutations will compete with each other. Second, we found that the mutation-recombination process has higher equilibrium fitness levels than the pure-mutation process under the same mutation rate. This is also consistent with the current view that recombination is beneficial to organisms in a new environment because they will be more fit than their pure-mutation counterparts. Additionally, when the mutation rate increases, the final fitness levels for both pure-mutation and mutation-recombination processes decreases. Previous computational studies have shown that a high mutation rate is detrimental to the fitness of an organism in the absence of recombination effects. [14, 27] Here, we show that a high mutation rate is also detrimental to the fitness level in the presence of recombination. This is because although recombination can remove deleterious mutations, it cannot do so effectively if the mutation rate is very high; therefore, the

final fitness level of an organism will be reduced. Moreover, there is a larger initial decrease in fitness for both the mutation-recombination and pure-mutation processes evolving under high mutation rates. The larger initial decrease for the high mutation rate process occurs because the adaptation process is essentially a global optimization of the fitness landscape. The adaptation process can thus be viewed as an organism "walking" on the fitness landscape, searching for the global maximum. The fitness landscape is very rugged, meaning that it has many local peaks. We select an initial condition that is in the vicinity of a local peak so that any subsequent sequence space search will cause the organism to leave the local peak, resulting in an initial decrease in fitness level. A high mutation rate leads to the accumulation of many deleterious mutations, and the fitness of the population initially drops. Finally, although high mutation rate lowers the final fitness level of the population, we observe that a high mutation rate can increase the speed of adaptation for both pure-mutation and mutation-recombination populations. This is because a high mutation rate also increases the likelihood that a beneficial mutation will occur, therefore increasing the speed of adaptation. Figure 3.4 (B) shows the results of a control experiment using a lattice model. The features appearing in the Eris model also appear in the lattice model, showing there is no large discrepancy between the two models. Therefore, the epistasis between different protein residues does not significantly affect the linear additive assumption of protein free energy.

A



B

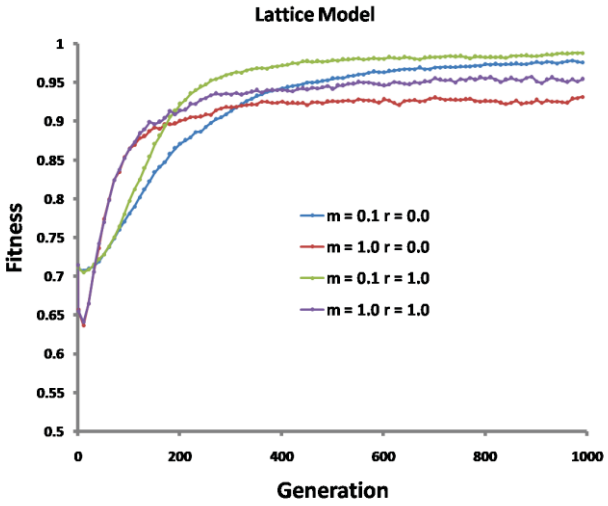


Figure 3.4(A) Evolution of fitness using Eris Model with different population size N, mutation rate m and recombination strength r.

Figure 3.4(B) Evolution of fitness using Lattice Model with different population size N, mutation rate m and recombination strength r.

Our realistic protein model allows us to study protein sequence evolution in the context of the adaptation process. We therefore consider amino acid changes in all proteins in our model. We calculate the sequence entropy S(p) of these proteins to analyze the degree of diversity of the proteins of the organism. We then align all sequences in the population for each of the proteins to obtain S(p). The sequence entropy of a residue in the kth position is defined as follows: [28]

$$S_k = -\sum_{i=1}^{20} P_i^k \log P_i^k$$

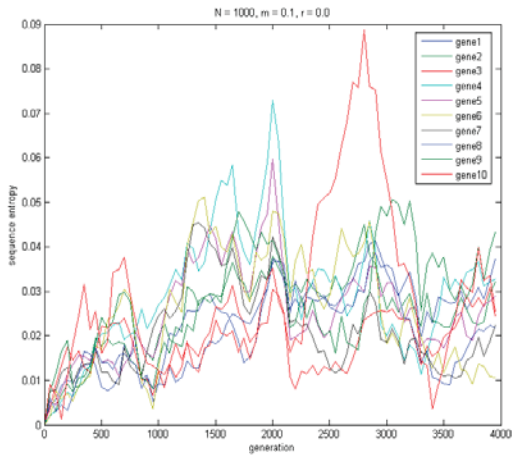
where P_i^k is the frequency of amino acid i in the kth position in a multiple alignment of sequences from all organisms in the population. The sequence entropy for a whole

protein is obtained by averaging the entropy over all positions in its sequence.

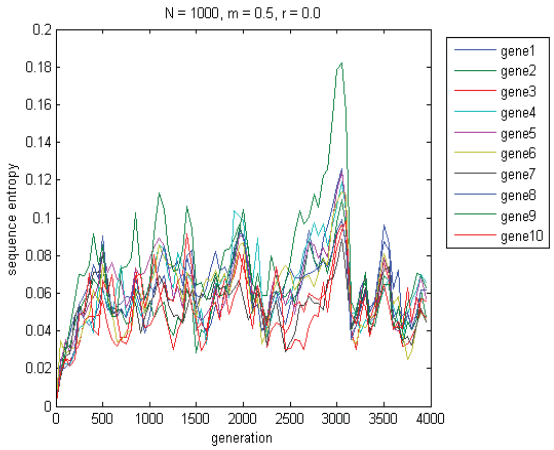
Figure 3.5 (A-J) shows the evolutionary time dependence of sequence entropy [S(p)] for each of the 10 proteins in our model. Low S indicates that all proteins of a given locus in the population have very similar sequences, whereas high S suggests substantial sequence heterogeneity in the population. There are several important features of the sequence entropy. First, we find that a high mutation rate increases the sequence entropies in both pure-mutation and mutation-recombination populations. This is because a high mutation rate will increase the sequence diversity and therefore increase the sequence entropy. Second, we notice that the incorporation of recombination increases the sequence entropy. This is because recombination explores the sequence space more thoroughly than a pure-mutation process alone, and therefore, the population carries more polymorphic loci. Third, we find that the sequence entropies for different genes move "synchronously" in the pure-mutation process. The synchronous movement of sequence entropies in different proteins is due to epistatic events in which a beneficial mutation occurs in one gene and the lineage that carries this beneficial mutation quickly takes over the population. Because of the effect of linkage, other genes in that organism will also become fixed, resulting in the synchronous movement of sequence entropy. Additionally, we observe that when the mutation rate increases, the synchronous movement effect is stronger, and the time interval between each move becomes smaller. This is because the chance that a beneficial mutation will occur increase as the mutation rate increases. Therefore, more beneficial mutations may become fixed. However, in the mutation-recombination

population, there is no such synchronous movement. This is because recombination disrupts the linkage of different genes. Therefore, the occurrence of a beneficial mutation in one gene has a weaker effect on the fixation of other genes within the same organism. The sequence entropy evolution in a control experiment using lattice protein is shown in Figure 3.6 (A-J), which exhibits similar behaviors.

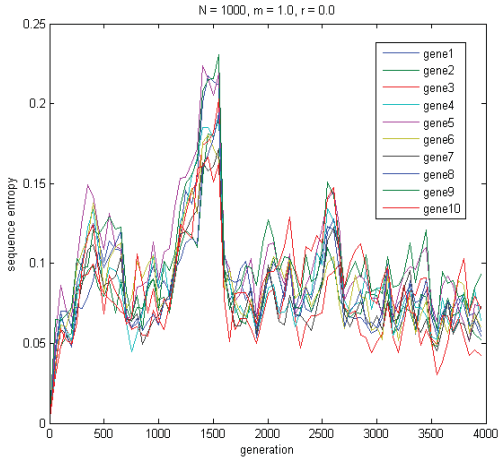
A



B



C



D

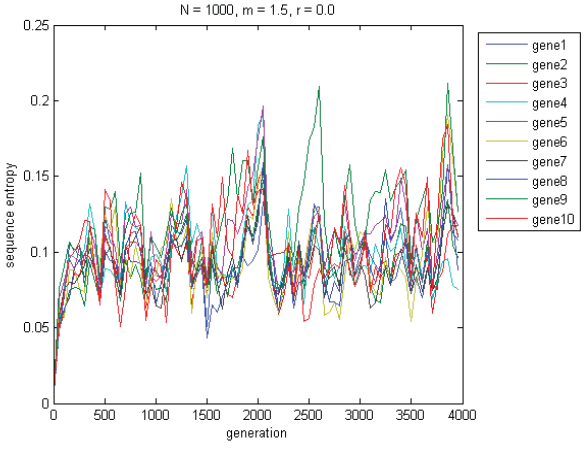
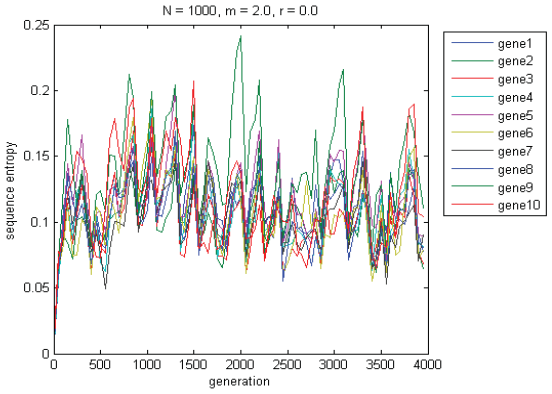
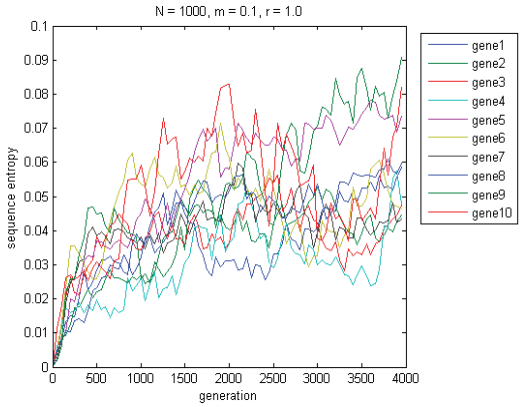


Figure 3.5 Evolution of Sequence Entropy S(p) using Eris Model for proteins with different population size N, mutation rate m and recombination strength r.

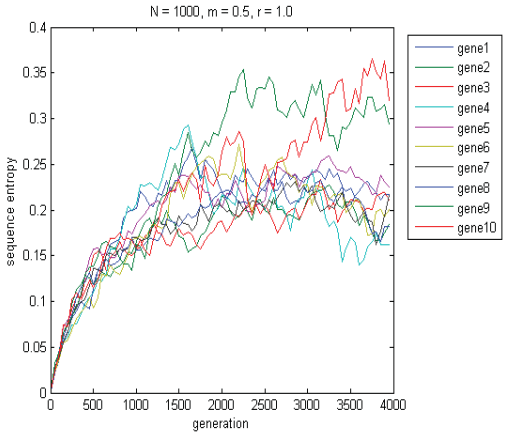
E



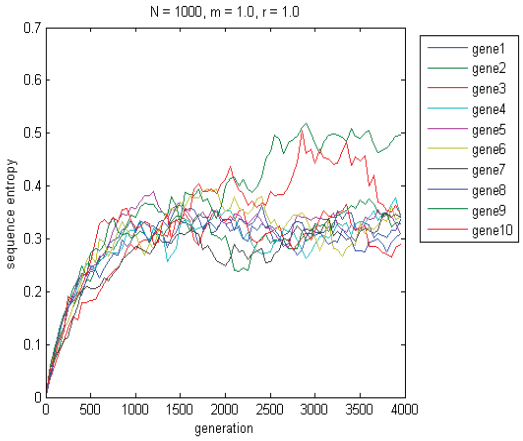
F



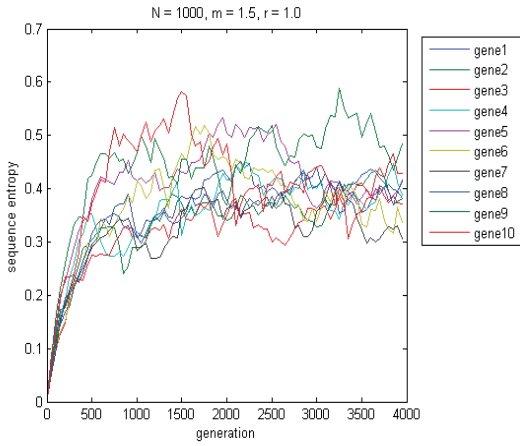
G



H



I



J

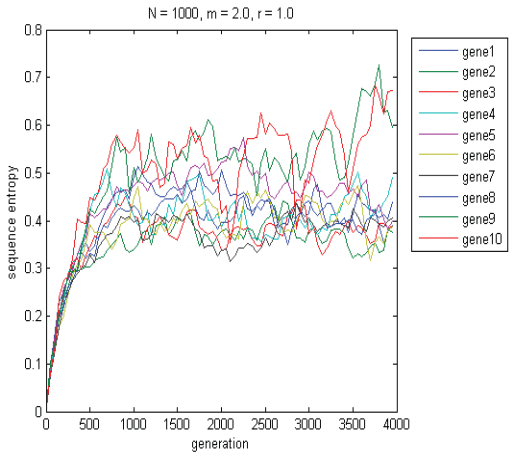


Figure 3.5 (Continued)

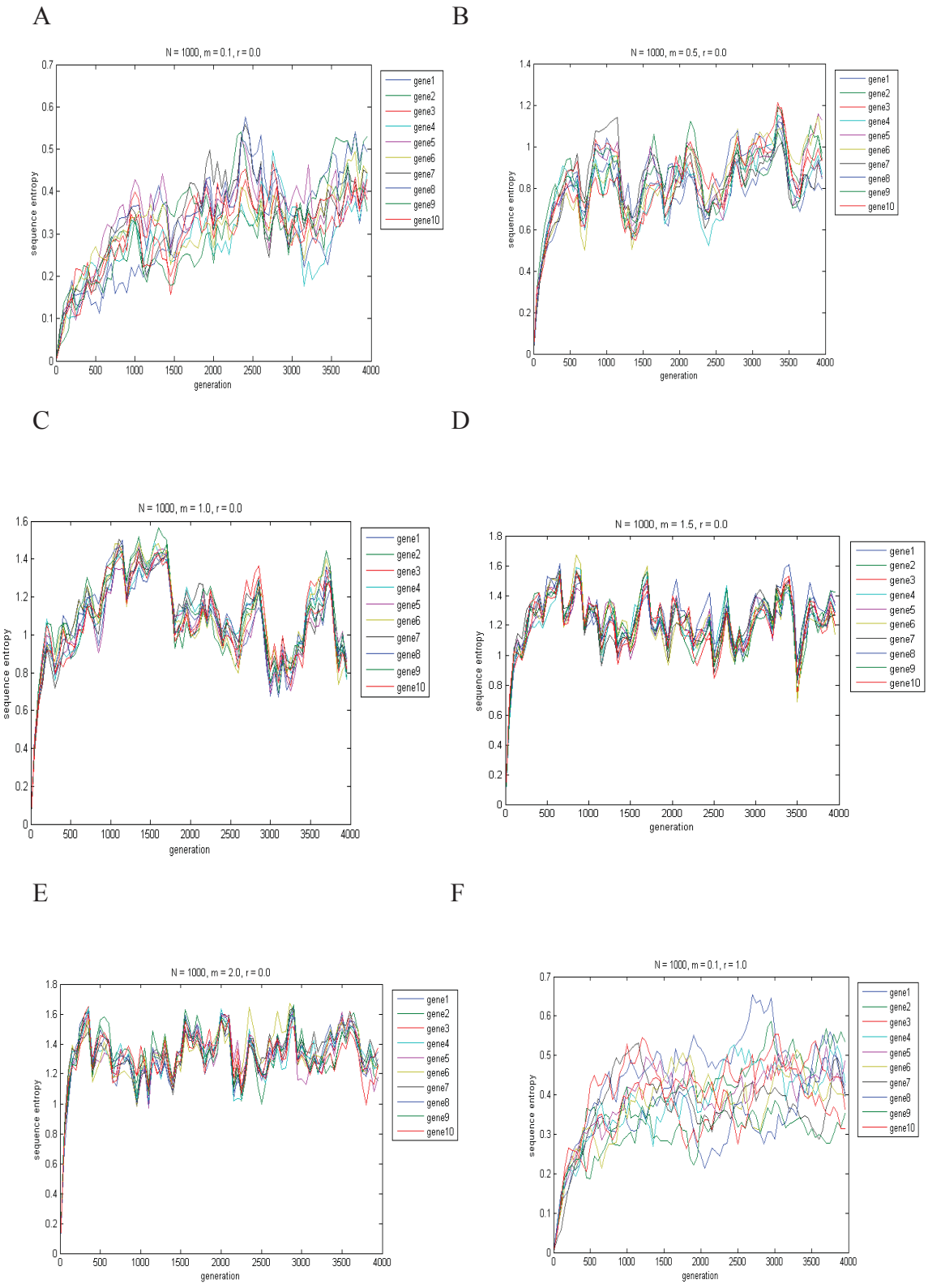
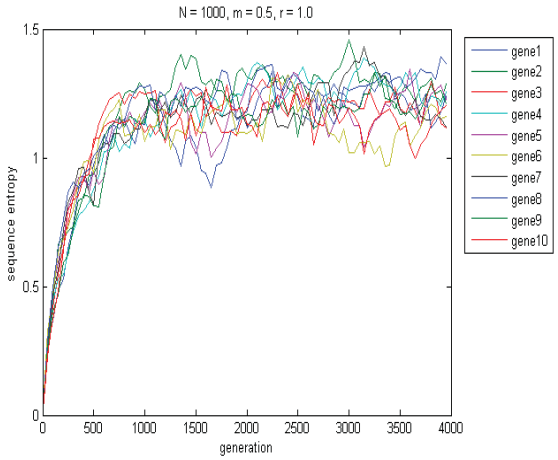
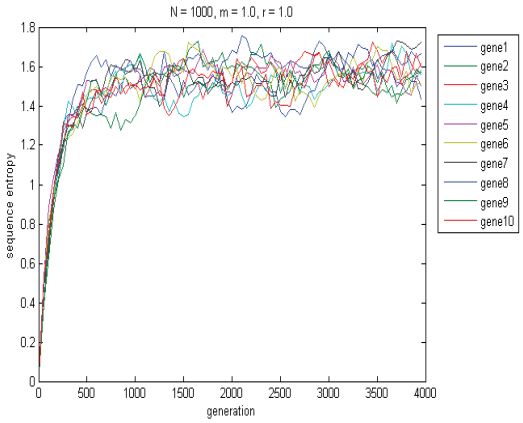


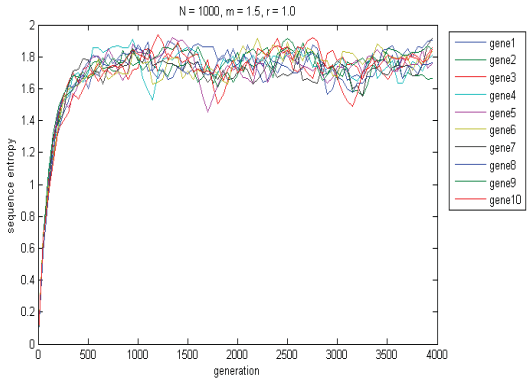
Figure 3.6 Evolution of Sequence Entropy S(p) using Lattice Model for proteins with different population size N, mutation rate m and recombination strength r.

G H





I J



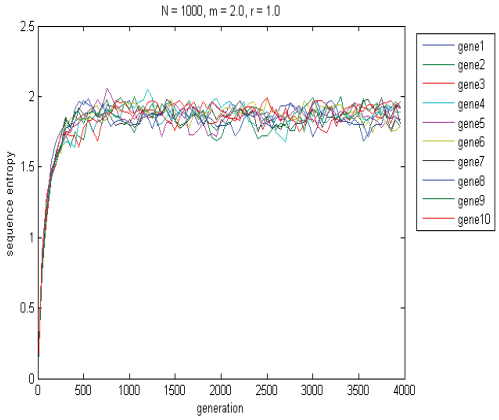


Figure 3.6 (Continued)

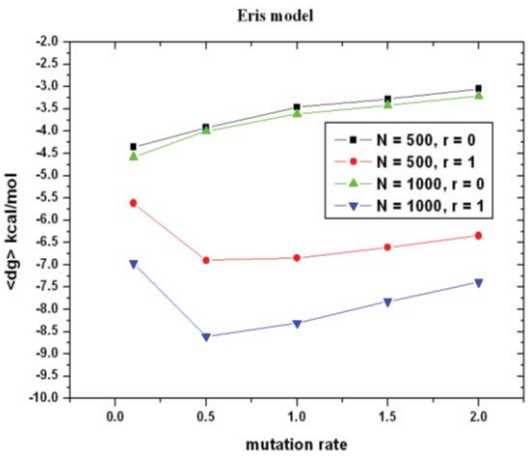
3.4.2 Protein thermodynamics effect

We now turn to the effect of recombination on the equilibrium protein free energies. We perform a series of simulations using different population sizes and mutation rates to investigate the equilibrium protein free energy for pure-mutation and mutation-recombination processes. The result for the Eris model is shown in Figure 3.7(A). There are several interesting features of the above figure. First, for a

pure-mutation process, the protein free energy decreases with an increase in mutation rate. Second, for the mutation-recombination process, protein stability increases with the increase of mutation rate at low mutation rates but decreases with the increase of mutation rate at high mutation rates. This phenomenon may be attributed to several factors. First, we should understand that recombination acts as a filter to deleterious mutations under the mutation inflows. Additionally, this filter can only be effective if mutation inflow is small. If the mutation inflow is large, the recombination filter will fail because it cannot remove all deleterious mutations. In the low mutation rate regime, an increase in the mutation rate will increase both the beneficial mutation supply and deleterious mutation supply. The "recombination filter" will remove the deleterious mutations and retain the beneficial mutations. Therefore, an increase in the mutation rate increases protein stability. However, in the high mutation rate regime, the "recombination filter" will not be able to remove deleterious mutations because the mutation inflow is too large. Therefore, a higher mutation rate will lead to a decrease in protein stability. Third, the mutation-recombination process is associated with higher protein stability than is the pure-mutation process at the same population size and mutation rate. This is because the mutation-recombination process explores the sequence space more efficiently and therefore can achieve greater protein stability. Fourth, protein stability increases with population size at the same mutation rate for both the pure-mutation and mutation-recombination processes. This is consistent with previous findings for the pure-mutation process [14] and occurs because the selection effect is more pronounced in a large population, thus enabling the protein to become

more stable at equilibrium. Fifth, the gap in protein stabilities between different population sizes is larger for mutation-recombination process than for the pure-mutation process. This is because the recombination-mutation process searches the sequence space more efficiently and therefore produces more stable proteins and more unstable proteins compared to the pure-mutation process. Under the more stringent selection of a larger population size, the unstable proteins are purged, and the recombination-mutation process retains more stable proteins than does the pure-mutation process. Therefore, the gap in protein stabilities between different population sizes is larger for the mutation-recombination process than for the pure-mutation process. Figure 3.7 (B) shows the results of a control experiment using the lattice model. The features that appeared in the Eris model also appear in the lattice model, showing that there is no major discrepancy between the two models. Therefore, the epistasis does not significantly affect the linear additive assumption of protein free energy.

A



B

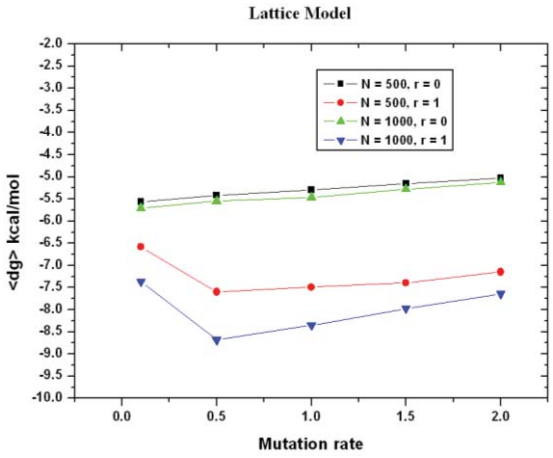
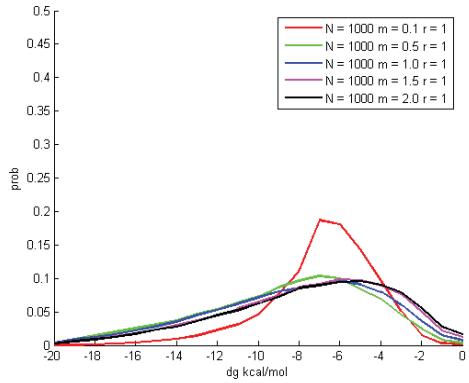


Figure 3.7 (A) The dependence of stability of all proteins averaged over all organisms in a population on mutation rate with different population size N and recombination strength r using Eris Model. Figure 3.7 (B) The dependence of stability of all proteins averaged over all organisms in a population on mutation rate with different population size N and recombination strength r using Lattice Model.

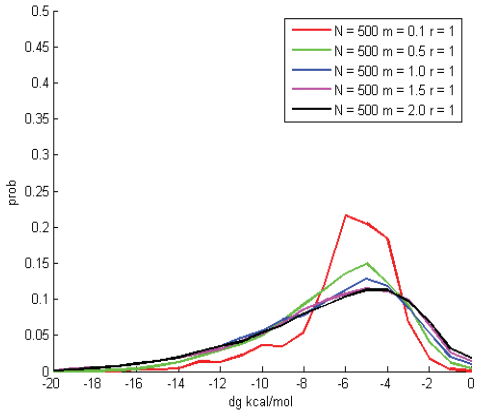
We also investigated the distribution of protein stability in the high recombination and pure-mutation regimes. Figure 3.8 (A-D) shows the protein stability distribution for different population sizes and mutation rate under both pure-mutation and mutation-recombination regimes under the Eris model. Figure 3.9 (A-D) shows the protein stability distributions for different population sizes and mutation rates for both pure regimes under the lattice protein model. Both models make the same prediction: the distribution of protein stability is much broader in the mutation-recombination regime than that in the pure-mutation regime. The reason for the broader distribution in the recombination regime is that recombination explores

the sequence space more thoroughly, and it is thus more likely to produce both very stable proteins and marginally stable proteins.

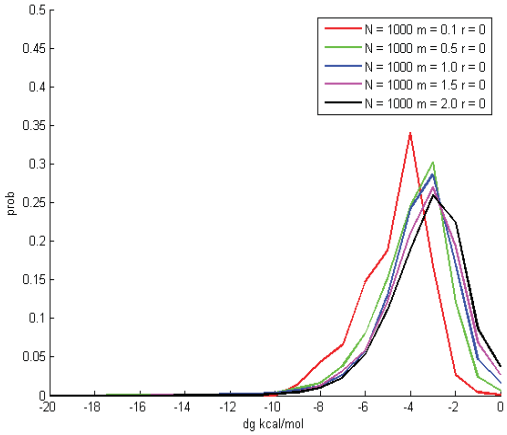
A



B



C



D

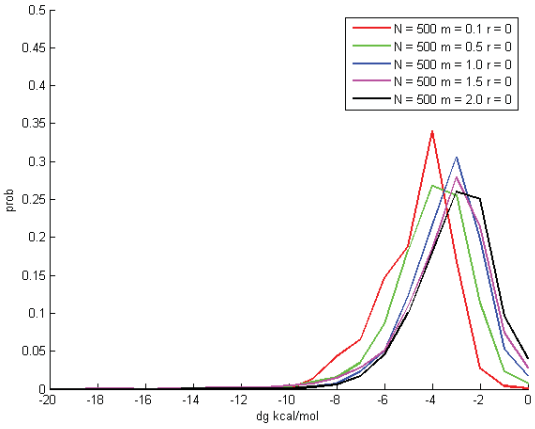


Figure 3.8 (A-D) Distribution of protein stabilities with different population size N, mutation rate m and recombination strength r under Eris Model.

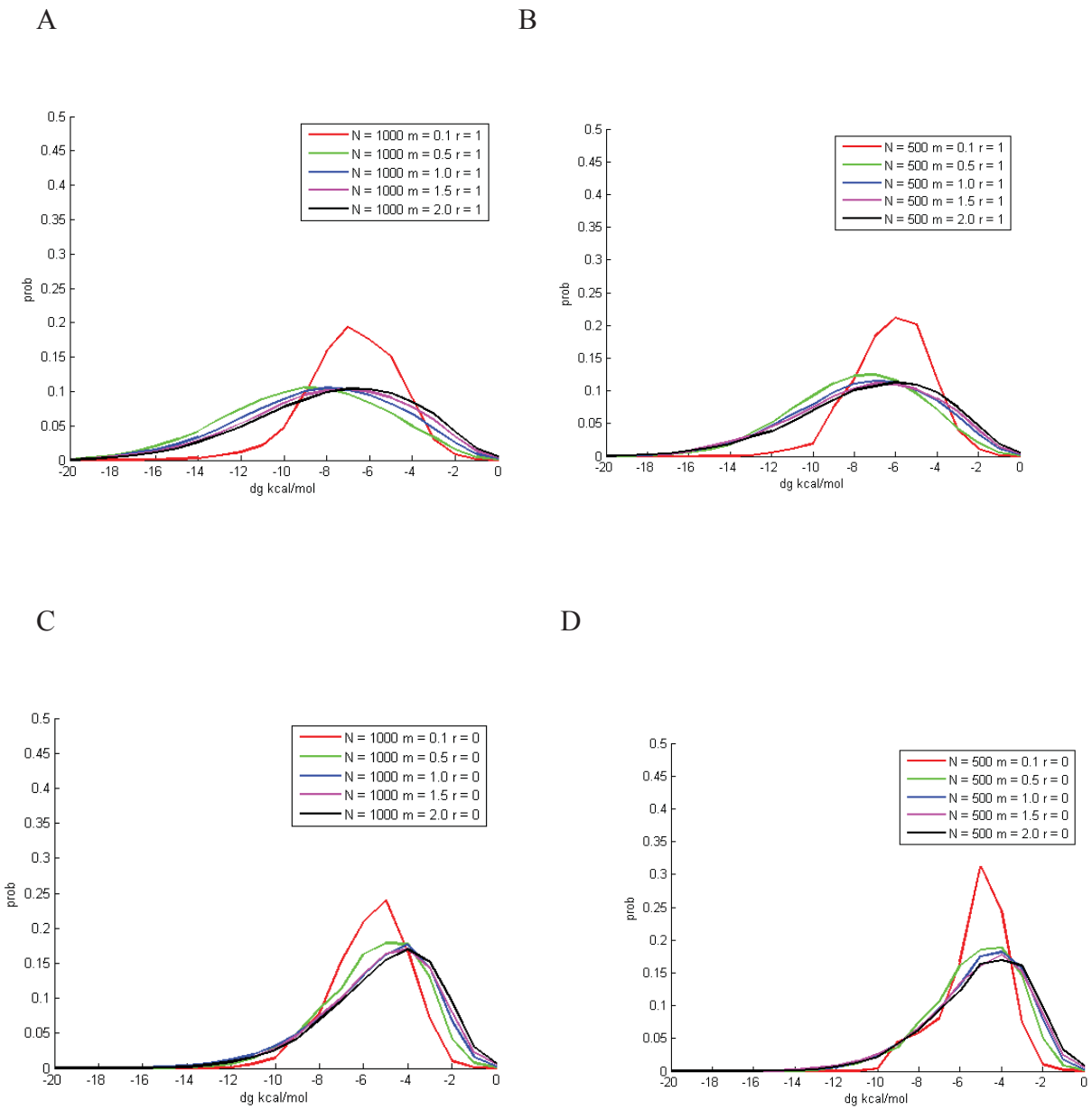


Figure 3.9 (A-D) Distribution of protein stabilities with different population size N, mutation rate m and recombination strength r under Lattice Model.

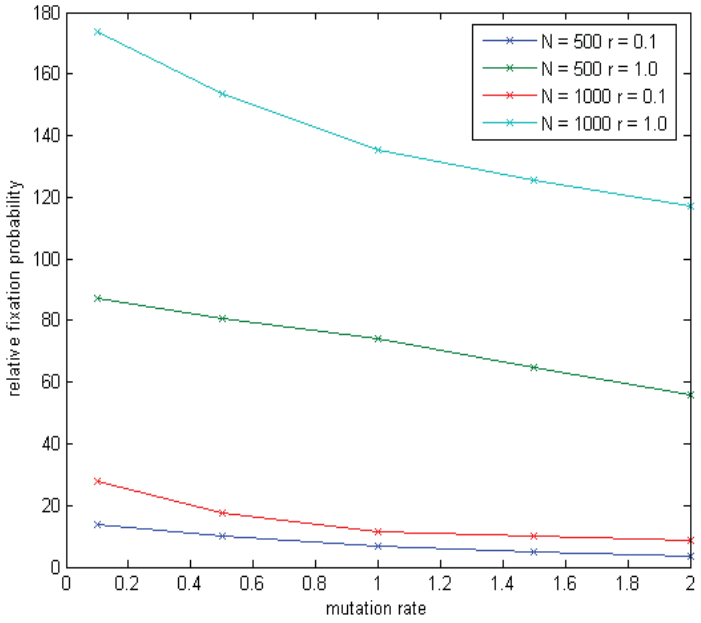
3.4.3 Fixation probability of recombination alleles

Our final objective is to investigate the fixation probability of recombinant alleles in a well-adapted pure-mutation population. It is important to note the

difference between this issue and the previously discussed issue of adaptation dynamics. Adaptation dynamics investigates the features of pure-mutation and mutation-recombination processes in the context of a novel environment to which the initial population is not well adapted. Here, in contrast, we study the origin of recombination in a pure-mutation population in same environment. Our approach is as follows: at the start of each simulation, each organism in the population has an inactive recombination modifier allele. We let the population evolve for N generations to reach an approximate mutation-selection equilibrium. Then, we randomly choose an individual at which a recombination modifier allele becomes activated. The population then evolves until the fixation or loss of the recombination modifier allele in the population. We run five thousand simulations for each case with different population sizes, mutation rates and recombination strengths to estimate the probability that the recombination modifier allele becomes fixed. The probability is then compared with the neutral expectation so that the relative probability of fixation is quantified. In the neutral case, the probability of the fixation of the modifier allele is 1/N, where N is the population size. Figure 3.10 (A) and (B) show the relative probability of fixation of the modifier allele under the Eris and lattice models. First, we study how the strength of the modifier allele affects its evolutionary fate. We observe that modifier alleles with higher recombination strength have higher relative fixation probabilities. Second, Figure 3.10 (A) and (B) show that, under the same recombination strength and mutation rate conditions, the recombinant allele has a greater relative probability of becoming fixed in a larger population. This result was

previously observed in a traditional population genetics simulation [13, 29]. The larger population maintains more polymorphisms; therefore, the variation in the stability of each gene is larger in the larger population, and there are more very stable and very unstable proteins in large populations than in small populations. Therefore, recombination can bring those very stable proteins together and increase their fixation probabilities. Third, our results show that, with the same recombination strength and population size, higher mutation rates lead to lower fixation probabilities of recombinant alleles. This is because high mutation rate will generate more deleterious mutations to the organism that carries the recombinant allele, therefore reducing the chance that the lineage will survive in the long run.

A



B

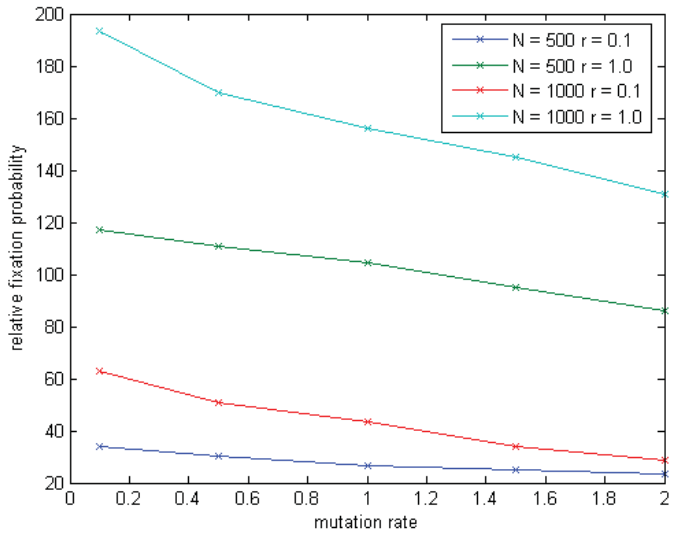


Figure 3.10 (A) The dependence of relative fixation probability of the recombination allele on different mutation rate m with different population size N and recombination strength r under Eris model. Figure 3.10 (B) The dependence of relative fixation probability of the recombination allele on different mutation rate m with different population size N and recombination strength r under lattice model.

3.5 Discussion

In this work, we studied the effects of mutation and recombination on various aspects of protein evolution, including the adaptation process, the equilibrium thermodynamic properties and the evolutionary origin of recombination under different mutation rates, recombination strengths, and population sizes. Our realistic biophysical protein model allows us to observe a rich set of behaviors. Our model is based on a simple but biologically reasonable assumption that the organismal birth rate is determined by the fraction of correctly folded proteins. The realistic biophysical protein model naturally includes epistatic effects and the sequence depletion effect. Furthermore, the mutational and recombinational effects in our model are realistic and do not assume prespecified fitness effects of mutation and recombination. Our model also permits recurrent and back mutations. Another important feature of our model is that it admits intragenic recombination, which cannot be captured by the traditional population genetics approach. In fact, intragenic recombination is a very important factor in the protein evolution process. For example, almost all recombination events in bacteria are intragenic.[18]

We observed that recombination can increase the adaptation speed of a population and the final fitness level of a population. Previous lattice model simulations also showed that the inclusion of recombination results in a faster increase in fitness and greater overall fitness level during evolution.[30, 31] Adaption is slower in pure-mutation populations because beneficial mutations must compete with each other, while recombination brings together beneficial mutations, breaking the linkage of

different genes. This phenomenon is a manifestation of the Hill-Robertson effect, which states that linkage can decrease the efficacy of selection.[32] Our results are also consistent with those of a previous study which showed that, at a high mutation rate, recurrent mutations can unite beneficial mutations in the same background and that the relative difference in adaptation time between sexual and asexual populations is small when there is a large mutation supply, making asexual populations behave progressively more like sexual populations.[33] Indeed, we observed that when the mutation rate increases, the speed of adaptation in a pure-mutation population increases, and the difference in adaptation speed between pure-mutation and recombination populations decreases. Therefore, mutation rate is a double-edged sword. On one hand, it increases the adaptation speed, but on the other hand, it lowers the final fitness level of the whole population.

The examination of sequence entropy allows us to study the change in the sequence space during the adaptation process. We observed that recombination tends to increase sequence entropy by breaking the linkage between different residues and alleviating the hitchhiking effect. This observation is consistent with previous studies showing that recombination rate is positively correlated with sequence diversity.[34, 35] We observed that the sequence entropies of different proteins move up and down synchronously in the pure-mutation populations, while there is no such synchronous movement in mutation-recombination populations. The synchronous movement of sequence entropy in a pure-mutation population was observed in a previous simulation using lattice protein models.[12] Our results are also consistent with the

results of a previous population dynamics simulation, which showed that there is a transition from genotype selection to allele selection when the recombination rate increases from zero to a very high number.[36] When there is no recombination, every gene is linked and the genotype serves as the unit of selection. When recombination is very high, however, the linkage between each gene is broken, and each gene is selected independently. In that study, the author used a traditional population genetics model that modeled each gene as an allele and thus could not incorporate the intragenic homologous recombination effect. Our realistic protein model generalizes their results to the amino acid residue level by taking into account of the intragenic homologous recombination. Additionally, our model revealed behavior different from that observed in a previous study using a two-letter HP lattice protein model. [18] Their result showed that when mutation dominates, the average sequence distance increases monotonically, but while recombination dominates, the sequence distance increases and decreases several times before reaching an equilibrium state. Our result shows the contrary: in the pure-mutation case, sequence entropy increases and decreases several times, while in the mutation-recombination case, sequence entropy increases monotonically. This discrepancy likely arises because in our model, proteins evolve under selection, which takes the form of a Fermi function. This means that sequences folding into the same structure may have different fitness levels. Therefore, a beneficial mutation can become fixed so that the sequence entropy increases and decreases several times during the pure-mutation process. In contrast, in the mutation-recombination process, sequence entropy increases monotonically because

the linkage between different residues is disrupted by recombination. In the two-letter HP lattice protein model, however, there is no fitness difference among protein sequences folding into the correct target structures. Essentially, this model treats fitness as a step function: the fitness is one if the protein sequence folds into the target structure and zero if the protein sequence does not fold into the target structure. Therefore, sequence diversity increases monotonically when mutation dominates because mutations typically result in moderate sequence changes that do not greatly decrease the probability of correctly folding into the target structure. Therefore, the mutant sequence can be preserved and the sequence diversity increases. In contrast, when recombination dominates, sequence diversity increases and decreases several times. This is because recombination typically causes drastic sequence changes, with a higher probability of disrupting the ability to fold into the target structure. Therefore, the sequence diversity can decrease during this process. This demonstrates that a correct fitness function plays an important role in shaping the protein evolution process in the sequence space. The different effects of step-function fitness and Fermi-function fitness on protein stability distribution have been studied previously.[27] In that study, the result shows that Fermi-function fitness can produce better agreement with experimental protein stability distributions by taking into account the gradual loss of fitness as proteins become marginally stable. Our result shows that the gradual loss of fitness as proteins become marginally stable also has important effects on determining the sequence evolutionary trajectories in the pure-mutation and mutation-recombination processes.

We also examined the different effects of pure-mutation and mutation-recombination models on protein thermodynamics. Our structure-based protein model reproduces the experimental protein stability distribution. In addition, our model predicts that protein stability will decrease when the mutation rate increases in a pure-mutation process. This result is consistent with the results of previous simulations, which showed that a high mutation rate is detrimental to protein folding stability.[14, 27] Our result generalizes this finding; the previous studies used phenomenological protein models without explicit structures and therefore without the sequence depletion effect. One experimental study showed that some features of protein stability, such as contact density, are significantly different between RNA virus proteins and proteins from DNA-based organisms, which suggests that different mutation rates might result in different protein stabilities.[37] Moreover, our model extends the previous simulation results by explicitly considering the effects of recombination and predicts that protein stability should increase with mutation rate when the mutation rate is low and decrease with mutation rate when the mutation rate is high, although a full verification of this prediction will require the measurement of the stabilities of homologous proteins from both RNA and DNA viruses. We also observe that the mutation-recombination process leads to increased protein stability compared to that in the pure-mutation process at the same mutation rate. This finding is consistent with the fact that many retroviruses use frequent recombination as a strategy to increase their survival probability.[38, 39] Retroviruses generally require high mutation rates to avoid immune system attack. If they did not use recombination, their

proteins would become extremely unstable and they would not be able to survive. To increase the protein stability, they use frequent recombination to purge deleterious mutations.

Finally, we studied the origin of recombination in a pure-mutation population. We observed that modifier alleles with higher recombination strength have a higher relative fixation probability. This finding is consistent with the results of a previous simulation study using a traditional population genetics approach that did not consider protein folding biophysics.[13] This means that stronger recombination modifiers that increase the recombination rate substantially will have a much greater advantage over weak recombination modifiers that only increase recombination by a small amount. This result also reflects the fact that nature not only selects for the recombinational mode of reproduction but also selects for the right amount of recombination. If the recombination modifier is too weak to reduce the rate of mutation accumulation, the recombination mode of reproduction will have no advantage over the pure-mutation mode of reproduction. We also observed that the fixation of recombinant alleles occurs more rapidly in a large population given the same recombination strength and mutation rate. This result is consistent with previous findings demonstrating that the Hill-Robertson effect is more pronounced, and selection for recombination modifiers is stronger, in large populations with many sweeping loci. [40] Therefore, recombinant alleles have a greater chance of becoming fixed in a larger population.

3.6 Conclusion

In this study, we developed a realistic biophysical model with which to investigate the roles of mutation and recombination in protein evolution. Our biophysical protein folding model and fitness function allowed us to incorporate the epistasis effect and sequence depletion effect naturally in our study. Furthermore, our model considers intragenic recombination, which is a significant improvement upon previous models. The conclusions drawn from our study are as follows: First, during the adaptation process, mutation-recombination processes increase the adaptation speed compared to the pure-mutation process. Additionally, the mutation-recombination process has a higher equilibrium fitness level than does the pure-mutation process because the mutation-recombination process is able to explore the sequence space more thoroughly. Additionally, a high mutation rate is a double-edged sword because although it can increase the speed of the adaptation process, it also decreases the final fitness level. In addition, we found that a high mutation rate increases the sequence entropies in both pure-mutation and mutation-recombination processes. Recombination increases the sequence entropies compared to the pure-mutation process. The sequence entropies of different genes move synchronously in a pure-mutation process, and the synchronous rise and fall of sequence entropies reflect the occurrence and fixation of beneficial mutations in the population. These synchronous movements lead to the hitchhiking effect of different genes in an organism. The time interval between each synchronous move becomes smaller when the mutation rate increases because this increases the rate of the generation of beneficial mutations. No synchronous movement is observed in

mutation-recombination processes, however, because the linkage between genes is disrupted. Second, the effects of mutation and recombination on protein stability can be summarized as follows: Protein stability decreases with the increase of mutation rate; furthermore, for mutation-recombination processes, protein stability increases with the increase of mutation rate at low mutation rates and decreases with the increase of mutation rate at high mutation rates. The mutation-recombination process is associated with higher protein stability than is the pure-mutation process at the same population size and mutation rate. Protein stability increases with population size at the same mutation rate for both processes. Additionally, the distribution of protein stability is much broader in the recombination regime than in the pure-mutation regime. Finally, for the fixation of recombinant alleles in a pure-mutation population, modifier alleles with higher recombination strength have a higher relative fixation probability. The recombinant allele has a higher relative probability of becoming fixed in a large population. Finally, higher mutation rates lead to lower fixation probabilities for recombinant alleles.

3.7 Reference

1. Wagner GP & Altenberg L (1996) Complex adaptations and the evolution of evolvability. Evolution 50(3):967-976.

2. Kirschner M & Gerhart J (1998) Evolvability. Proc. Natl. Acad. Sci. 95:8420-8427.

3. Earl DJ & Deem MW (2004) Evolvability is a selectable trait. Proc. Natl. Acad.

Sci. 101:11531-11536.

4. Serohijos AWR, Rimas Z, & Shakhnovich EI (2012) Protein Biophysics Explains Why Highly Abundant Proteins Evolve Slowly. Cell Reports 2:249-256.

5. Zeldovich KB, Chen P, Shakhnovich BE, & Shakhnovich EI (2007) A First-Principles Model of Early Evolution: Emergence of Gene Families, Species, and Preferred Protein Folds. PLoS Comput. Biol. 3(7):1224-1238.

6. Zeldovich KB, Chen P, & Shakhnovich EI (2007) Protein stability imposes limits on organism complexity and speed of molecular evolution. Proc. Natl. Acad. Sci. 104:16152-16157.

7. Bloom JD, Raval A, & Wilke CO (2007) Thermodynamics of neutral protein evolution. Genetics 175:255-266.

8. Desai MM, Fisher DS, & Murray AW (2007) The Speed of Evolution and Maintenance of Variation in Asexual Populations. Curr. Biol. 17:385.

9. Boerlijst MC, Bonhoeffer S, & Nowak MA (1996) Viral quasi-species and recombination. Proc. R. Soc. Lond. B 263:1577-1584.

10. Feldman MW, Christiansen FB, & Brooks LD (1980) Evolution of recombination in a constant environment. Proc Natl Acad Sci USA-Biol Sci 77:4838-4841.

11. Eigen M, J M, & Schuster P (1989) The molecular quasi-species. Adv. Chem. Phys. 75:149-263.

12. Heo M, Kang L, & Shakhnovich EI (2008) Emergence of species in evolutionary "simulated annealing". Proc Natl Acad Sci USA 106:1869-1874.

13. Gordo I & Campos PRA (2008) Sex and deleterious mutations. Genetics

179:621-626.

14. Wylie CS & Shakhnovich EI (2011) A biophysical protein folding model accounts for most mutational fitness effects in viruses. Proc. Natl. Acad. Sci. 108(24):9916-9921.

15. Gouveia JF, Oliveria VMd, Satiro C, & Campos PRA (2009) Rate of fixation of beneficial mutations in sexual populations. Phys. Rev. E 79.

16. Sambrook J (1977) Adenovirus amazes at Cold Spring Harbor. Nature 268:101-104.

17. Walter G (1978) Why genes in pieces. Nature 271(5645):501-501.

18. Xia Y & Levitt M (2002) Roles of mutation and recombination in the evolution of protein thermodynamics. Proc. Natl. Acad. Sci. 99:10382-10387.

19. Cui Y, Wong WH, Bornberg-Bauer E, & Chan HS (2002) Recombinatoric exploration of novel folded structures: A heteropolymer-based model of protein evolutionary landscapes. Proc. Natl. Acad. Sci. 99:809-814.

20. Yin S, Ding F, & Dokholyan NV (2007) Eris: an automated estimator of protein stability. Nature Methods 4:466-467.

21. Yin S, Ding F, & Dokholyan NV (2007) Modeling backbone flexibility improves protein stability estimation. Structure 15:1567-1576.

22. Miyazawa S & Jernigan RL (1996) Residue-residue potentials with a favorable contact pair term and an unfavorable high packing density term for simulation and threading. J Mol Biol 256(3):623-644.

23. D KM, et al. (2006) ProTherm and ProNIT: thermodynamic databases for

proteins and protein-nucleic acid interactions. Nucleic Acids Res. 34:D204-D206.

24. Gerrish PJ & Lenski RE (1998) The fate of competing beneficial mutations in an asexual population. Genetica 102/103:127-144.

25. Barton NH & Charlesworth B (1998) Why Sex and Recombination? Science 281:1986-1990.

26. Wagner A (2011) The low cost of recombination in creating novel phenotypes. Bioessays 33:636-646.

27. Chen P & Shakhnovich EI (2009) Lethal Mutagenesis in Viruses and Bacteria. Genetics 183:639-650.

28. Yano T & Hasegawa M (1974) Entropy increase of amino acid sequence in protein. J Mol Evol 4:179-187.

29. Keightley PD & Otto SP (2006) Interference among deleterious mutations favours sex and recombination in finite populations. Nature 443:89-92.

30. Williams PD, Pollock DD, & Goldstein RA (2006) Selective Advantage of Recombination in Evolving Protein Populations:. a Lattice Model Study. International Journal of Modern Physics C 17(01):75-90.

31. Watson RA, Weinreich DM, & Wakeley J (2010) Genome structure and the benefit of sex. Evolution 65-2:523-536.

32. Hill WG & Robertson A (1966) The effect of linkage on limits to artificial selection. Genet Res 8:269-294.

33. Bollback JP & Huelsenbeck PJ (2007) Clonal interference is alleviated by high mutation rates in large populations. Mol Biol Evol 24(6):1397-1406.

34. Andolfatto P & Przeworski M (2001) Regions of Lower Crossing Over Harbor More Rare Variants in African Populations of Drosophila melanogaster. Genetics 158(2):657-665.

35. Begun DJ & Aquadro CF (1992) Levels of naturally occurring DNA polymorphism correlate with recombination rates in D. melanogaster. Nature 356:519-520.

36. Neher RA & Shraiman BI (2009) Competition between recombination and epistasis can cause a transition from allele to genotype selection. Proc. Natl. Acad. Sci. 106(16):6866-6871.

37. Tokuriki N, Oldfield CJ, Uversky VN, Berezovsky IN, & Tawfik DS (2009) Do viral proteins possess unique biophysical features? Trends Biochem. Sci. 34:53-59.

38. Hu WS & Temin HM (1990) Retroviral recombination and reverse transcription. Science 250:1227-1233.

39. Jung A, et al. (2002) Recombination: Multiply infected spleen cells in HIV patients. Nature 418:144-144.

40. Neher RA, Shraiman BI, & Fisher DS (2010) Rate of Adaptation in Large Sexual Populations Genetics 184:467-481.