Correlation Between the Conformation Space and the Sequence Space of Peptide Chain

T.N. SASAKI and M. SASAI

Graduate School of Human Informatics, Nagoya University, Nagoya 464-8601, Japan

Abstract. Effects of point mutations are investigated in a computer model of sequence selection. In the model, sequences are selected according to their functional ability defined by the configuration at the active site. Through such functional selection on the local configuration, sequences which fold globally into a unique conformation can be selected from random sequences. The selected sequence accepts many point mutations as neutral mutations. More structurally fluctuating sites in the chain accept a larger variety of point mutations, which shows correlation between the conformational fluctuation and the sequence fluctuation.

Key words: functional selection, knowledge-based potential, mutational robustness, neutral evolution, off-lattice model, sequence space, super-funnel

1. Introduction

Proteins are robust against mutations. Effects of mutations on structure [1], stability [2–4], activity [5, 6] and folding kinetics [7] were examined, which confirmed that proteins are relatively insensitive to mutations unless some crucial residues at the active site are altered [1]. Folding kinetics is largely determined by topology of the native conformation [8, 9] and peptides synthesized with the reduced alphabet from 20 amino acids show similar stability and activity as wild type sequences [10, 11], both of which support the picture that proteins are insensitive to the detailed difference of sequences. N. Go noted such ubiquitously found insensitivity of proteins with a phrase, ‘principle of mutational robustness’ [12].

In the last decade much effort has been devoted to elucidate the folding mechanism of proteins, which revealed that sequence of proteins is designed in a specific way to facilitate the rapid folding into a unique conformation in a biological time scale [13]. This design principle of foldable sequence was referred to as ‘consistency principle’ [14] or ‘principle of minimal frustration’ [15]. Since the consistency principle stresses the restriction on sequences and the principle of mutational robustness stresses the allowance of sequences, two principles, at a first glance, seem to contradict each other.

On the relation between two principles, there are two possible ways of thinking, one of which could be called ‘scenario of opposite selective pressures’. In this

scenario biologically selected sequences belong into a limited category in which opposite requirements of two principles are satisfied at the same time. Sequences might have evolved to be in such category due to the presence of two oppositely working selection pressures [4]. The other could be ‘the parallel scenario’ in which mutational robustness is a consequence or a byproduct of selection of sequences to satisfy the consistency principle. Results of lattice models of folding seem to support the idea of the parallel scenario [16, 17]. In the present paper, computer simulation of sequence selection is carried out and point mutations on the selected sequence are examined to discuss which of two scenarios holds in the model.

In order to simulate the selection process we employ a knowledge-based potential model of folding [18]. In the model, peptide is represented by a connected off-lattice chain of beads, each of which corresponds to β carbon (for glycine α carbon is used). 20×20 types of two-body potentials between beads are inferred from the frequency of spatial distances found in conformations in the protein data bank (PDB). This knowledge-based potential model has ability to fold short helical proteins such as calcium-binding protein or DNA-binding protein into the native conformation.

Since selection pressure in Darwinian process should act on the biological function of the sequence, it is also natural in simulation to select sequences according to their function. Simulations with a simple spin-glass model [19], a lattice model [20], and the knowledge-based potential model [21] showed that random sequences can evolve to be protein-like through such functional selection. In Reference 21, Yomo *et al.* showed that protein-like properties including compactness, secondary structure, and ability to fold into the unique native conformation develop from random sequences through selection on the functional ability. Thus, theoretical study suggested a scenario that the selection on the functional ability is sufficient to find sequences satisfying the consistency principle from a vast number of sequences. This scenario was indeed verified by recent experimental selection from library of random sequences [22, 23]. In this paper, we analyze effects of point mutations of the sequence selected by the method of Reference 21.

2. Functional Selection and Point Mutation

Procedure of the functional selection in computer is illustrated in Figure 1. Starting from a random sequence of 47 residues at the 0th generation, the sequence at the i th generation, $\text{seq}(i)$ and its mutant, $\text{mut}(i)$, are compared, where $\text{mut}(i)$ is generated by substituting a randomly chosen site of $\text{seq}(i)$ by a randomly chosen kind of residue. Then, folding simulation is conducted by using the knowledge-based potential model for $\text{mut}(i)$. Starting from the stretched linear conformation, folding of the chain is simulated by numerically integrating the Langevin dynamics under the knowledge-based potentials among residues. This folding simulation is repeated 25 times for the sequence, and hence 25 independent trajectories are obtained to assess the activity.

Figure 1. Algorithm of computer simulation of functional selection. Starting from a random sequence at $i = 0$ th generation, sequence of the i th generation is subjected to a point mutation by replacing the residue at a randomly chosen site with a randomly chosen type of residue. Then, folding simulation of the mutant is carried out and the measure of the functional ability, $DCAC$, defined by Equation 1 is calculated. The mutant is judged whether it is handed over to the $i + 1$ st generation or not with the Metropolis algorithm by regarding $1/DCAC$ as the fictitious energy. When the mutant is rejected, sequence at the i th generation is passed on to the $i + 1$ st generation.

Here, we assume that activity of the sequence is measured by similarity of configuration at the active site to the one expected in a highly functional protein: As a reference active site, we use the four residues near the C terminal of engrailed homeodomain (PDB code, 1enh). These four residues are relevant to DNA binding of 1enh and are highly conservative [24]. The activity of the sequence is estimated in the model by measuring the degree of convergence of active-site configuration (DCAC),

$$DCAC = 1 / \left\langle \frac{1}{6} \sum_{\substack{a,b \in \\ \text{active site}}} |r_{ab} - r_{ab}^{enh}| \right\rangle_{\text{over } 20000 \times 25 \text{ structures}}, \quad (1)$$

Figure 2. A globally ordered unique conformation develops through the functional selection. (a) An example of random conformations of the sequence at the 10th generation. (b) The native-like conformation of the sequence at the 80th generation. Snapshots at 60 000th step in the Langevin dynamics simulation are shown. Dark parts of the chain represent the active site.

where r_{ab} is the pair distance between residues at the active site of the simulated chain, r_{ab}^{enh} is that of the referenced active-site configuration in 1enh, and $\langle \dots \rangle$ is the average over the structures of 20 000 snapshots from 40 000 to 60 000 steps in 25 trajectories of the Langevin dynamics. Summation in Equation (1) is the sum over $4!/(2!2!) = 6$ pair-distances among 4 residues. Then, $\text{mut}(i)$ is selected as the sequence at the $i + 1$ st generation, $\text{seq}(i + 1)$, with the probability,

$$P = \begin{cases} 1 & \text{if } DCAC_{\text{mut}(i)} > DCAC_{\text{seq}(i)} \\ \exp \left[\left((DCAC_{\text{seq}(i)})^{-1} - (DCAC_{\text{mut}(i)})^{-1} \right) / \tilde{T} \right] & \text{if } DCAC_{\text{mut}(i)} \leq DCAC_{\text{seq}(i)} \end{cases}, \quad (2)$$

otherwise, $\text{seq}(i)$ is passed on as $\text{seq}(i + 1)$. This is the Metropolis Monte Carlo walk in the sequence space with the fictitious temperature, $\tilde{T}$, and the fictitious energy, $1/DCAC$.

In the present calculation $\tilde{T}$ was set to be 0.1 from 0th to 72nd generation and 0.01 after 73rd generation. At early generations in the selection process, different folding trajectories of the same sequence end up with differently frozen random conformations, one of which is shown in Figure 2a. At later generations different folding trajectories of the same sequence converge into a unique conformation as shown in Figure 2b: Through the selection process, structural constraint imposed on the active site was transmitted to other sites through interactions and the whole

Figure 3. Distributions of the number of mutants as functions of $DCAC_{\text{mut}}/DCAC_{\text{seq}(80)}$, where $DCAC_{\text{seq}(80)} = 2.90 \text{ \AA}^{-1}$ and the binning width is 0.0345. Black bars represents the number of mutants accepted with $\tilde{T} = 0.01$, gray bars are the number of mutants accepted with $\tilde{T} = 0.1$, and white bars are the total number of mutants.

chain acquired consistency to suppress the excess structural fluctuation at the active site by taking a globally ordered compact conformation. Similar development of the protein-like features were found for 5 runs among 15 trial selection runs [21]. Thus, a necessary condition for the activity (suitable configuration at the active site) is a sufficient condition for selecting sequences which satisfy the consistency principle in the model.

In this paper, thus selected sequence at the 80th generation, $\text{seq}(80)$, is subjected to point mutations. All the $19 \times 47 = 893$ mutants are generated and each mutant is folded 25 times from the stretched chain conformation with the knowledge-based potentials. Then, $DCAC$ and foldability of mutants are assessed from folding trajectories and probability of acceptance of mutation is estimated.

3. Results and Discussion

893 mutants can be categorized into three classes: (1) superior mutants whose $DCAC$ is larger than $DCAC_{\text{seq}(80)}$, (2) neutral mutants, and (3) inferior mutants whose $DCAC$ drops significantly from $DCAC_{\text{seq}(80)}$. Distribution of mutants is shown in Figure 3 as a function of the relative $DCAC$. Neutral mutants are populated as a large peak centered at $DCAC < DCAC_{\text{seq}(80)}$. When the selection

Figure 4. Entropy, S_n , of residue variability defined at each site, n , is connected by the real line and amplitude, A_n , of structural fluctuation of each residue around the configuration at the 80th generation is connected by the dashed line.

parameter $\tilde{T} = 0.1$ is used, 67% of 893 mutants are accepted as a sum of neutral and superior mutants with the criterion of Equation (2) and when $\tilde{T} = 0.01$, 25% of mutants are accepted, which shows the abundance of neutral mutations: The selected sequence in the simulation is not tuned perfectly but allows a large neutral fluctuation and leaves some room of further improvement. These features are similar to those in real proteins. From Figure 1 of Reference 5 on catalase activity, we can see that about 30% of random mutants severely lose the activity but the rest of mutants are populated around the wild type sequence without a drastic change in activity. The simulation results show that mutational robustness similar to that found in natural proteins arises solely from the selection on function as consistency among interactions develops through the functional selection, showing that the parallel scenario holds in the model.

Acceptable sequences of neutral and superior mutants are analyzed by comparing them with the structural fluctuation. Entropy of sequences, S_n , is defined at each site by $S_n = -\sum_{\mu} P_n(\mu) \ln P_n(\mu)$, where $P_n(\mu)$ is the probability that the amino acid residue of type μ is allowed at the n th site in an ensemble of sequences of seq(80) and its mutants, estimated from the relation,

$$P_n(\mu) = \exp(-\Delta(DCAC)^{-1}/\tilde{T})P_n(\alpha), \quad (3)$$

where $\Delta(DCAC)^{-1}$ is the change in $1/DCAC$ arisen from the mutation of the residue type α to the type μ at the n th site. In Figure 4, S_n is compared with the

Figure 5. $DCAC_{mut}/DCAC_{seq(80)}$ and $foldability_{mut}/foldability_{seq(80)}$ are plotted for 893 mutants generated by point mutations of $seq(80)$.

amplitude of structural fluctuation at the n th site $A_n = [\langle (\vec{r}_n - \langle \vec{r}_n \rangle)^2 \rangle]$ where $\vec{r}_n$ is a position vector of the n th site, $\langle \dots \rangle$ is the average along each trajectory of the Langevin dynamics around the folded conformation of the 80th generation, and $[\dots]$ is the average over 25 trajectories. S_n and A_n show different patterns for $n > 30$, which should be due to the relatively strong interactions between these residues and the active site at $n = 39, 40, 42$, and 44 . For $n < 30$, however, patterns of S_n and A_n have some similarity. Correlation coefficient is not very large, $r = 0.48$ ($p = 0.008$) when S_n and A_n are compared for $1 \leq n \leq 30$, and $r = 0.57$ ($p = 0.002$) for $1 \leq n \leq 27$, but from Figure 4 we can find existence of positive correlation: The more structurally fluctuating sites tend to accept a more variety of point mutations except in the region near the active site.

Because $DCAC$ is measured after the folding simulation, $DCAC$ should have correlation to foldability of the sequence. Here, foldability is defined as

$$foldability = 1 / \left\langle \frac{1}{47} \sum_{\text{all sites}} |\vec{r}_n^k - \vec{r}_n^l|^2 \right\rangle_{\text{over all pairs of 25 structures}}, \quad (4)$$

where $\vec{r}_n^k$ and $\vec{r}_n^l$ are position vectors of the n th site in the final structure of k th and l th folding trajectories and the bar represents average over 100 instances along the folding trajectories taken every 200 steps from 40 000 to 60 000 steps in the Langevin dynamics. $DCAC$ and $foldability$ of mutants are compared in Figure 5.

These two quantities have a weak correlation such that the inferior mutants have small *foldability*. Correlation, however, is not evident for acceptable mutants. Thus, for acceptable mutations, small change in sequence such as point mutation is more relevant to the small conformational fluctuation than to the large conformational movement in folding.

These results may be interpreted in a transparent way by introducing a concept of 'super-funnel', which is the funnel-like energy surface defined both in the conformation space and in the sequence space at the same time: Interactions in the foldable sequence are designed to be consistent to each other to yield a bias in energy surface such that energy tends to decrease as conformation becomes closer to the native conformation. Such biased energy surface of the foldable sequence was called the 'funnel-like' surface [13]. The funnel-like surface is smooth without a large abrupt change in energy upon small conformational change. Here, the small random conformational change results in a small change in interaction energy, and the similar change in interaction should also be yielded by the small random change in sequence. Thus, we can expect that the energy surface does not show an abrupt change by point mutation around the foldable sequence, which allows a large room of neutral fluctuation. A globally biased manifold in energy can be defined in the 'product space' of the conformational space and the sequence space, which should be called 'super-funnel'. The observed Mt. Fuji-type landscape in the sequence space [3] is consistent with this super-funnel picture. Existence of the super-funnel around the foldable sequence-conformation was suggested by using the lattice model [17]. If the super-funnel should govern the sequence-conformation fluctuation, then we can expect that the site-dependent curvature of the conformational funnel correlates to the site-dependent mutability along the sequence-funnel, either of which is a projection of a single manifold onto the conformation or the sequence space. Structural fluctuation, therefore should correlate to sequence variability within the super-funnel picture.

In summary, the selection simulation with a coarse-grained model of folding showed that mutational robustness is a natural consequence of the functional selection. Many point mutations of the selected sequence are allowed as neutral mutations, which are correlated to the structural fluctuation around the native conformation. In the coarse-grained model, however, atomistic details of real proteins are neglected, so that a different explanation of mutational robustness still remains: Design of sequence which enables the packing of side-chain atoms at the native state is not easy, and thus the principle of mutational robustness is difficult to be satisfied once the sequence is well designed in general. With this picture, two opposite selective pressures are needed to find sequences which satisfy the consistency of side-chain packing and the mutational robustness at the same time. Further examination of the super-funnel hypothesis with atomistic models and in experiments will help to resolve the problem on which of two scenarios, the opposite pressure scenario or the parallel scenario, holds in real proteins.

Acknowledgement

We gratefully acknowledge fruitful discussions with T. Yomo. This work was supported by Japan Science and Technology Corporation, Japan Society for the Promotion of Science, and by grants from Ministry of Education, Culture, Sports, Science and Technology, Japan.

References

1. Sinha, N. and Nussinov, R.: Point mutations and sequence variability in proteins, *Proc. Natl. Acad. Sci. USA* **98** (2001), 3139–3144.
2. Serrano, L., Day, A.G. and Fersht, A.R.: Step-wise mutation of barnase to binase, *J. Mol. Biol.* **233** (1993), 305–312.
3. Aita, T., Uchiyama, H., Inaoka, T., Nakajima, H., Kokubo, T. and Husimi, Y.: Analysis of a local fitness landscape with a model of the rough Mt. Fuji-type landscape, *Biopolymers* **54** (2000), 64–79.
4. Wintrobe, P.L. and Arnold, F.H.: Temperature adaptation of enzymes, in: F.H. Arnold (ed.), *Advances in Protein Chemistry* **55**, Academic Press, San Diego, 2001, pp. 161–225.
5. Matsuura, T., Yomo, T., Trakulnaleamsai, S., Ohashi, V., Yamamoto, K. and Urabe, I.: Nonadditivity of mutant effects on the properties of catalase I and its application to efficient directed evolution, *Protein Engineering* **11** (1998), 789–795.
6. Voigt, C.A., Mayo, S.L., Arnold, F.H. and Wang, Z.-G.: Computational method to reduce the search space for directed protein evolution, *Proc. Natl. Acad. Sci. USA* **98** (2001), 3778–3783.
7. Martinez, J.C. and Serrano, L.: The folding transition state between SH3 domain is conformationally restricted and evolutionary conserved, *Nat. Struct. Biol.* **6** (1999), 1010–1016.
8. Riddle, D.S., Grantcharova, V.P., Santiago, J.V. and Baker, D.: Experiment and theory highlight role of native state topology in SH3 folding, *Nat. Struct. Biol.* **6** (1999), 1016–1024.
9. Tsai, C.J., Maizel, J.V. and Nussinov, R.: Anatomy of protein structures: Visualizing how a one-dimensional protein chain folds into three-dimensional shape, *Proc. Natl. Acad. Sci. USA* **97** (2000), 12038–12043.
10. Riddle, D.S., *et al.*: Functional rapidly folding proteins from simplified amino acid sequences, *Nat. Struct. Biol.* **4** (1997), 805–809.
11. Brown, B.M. and Sauer, R.T.: Tolerance of Arc repressor to multiple-alanine substitutions, *Proc. Natl. Acad. Sci. USA* **96** (1999), 1983–1988.
12. Go, N.: The consistency principle revisited, in: K. Kuwajima and M. Arai (eds.), *Old and New Views of Protein Folding*, Elsevier, Amsterdam, 1999, pp. 97–105.
13. Onuchic, J.N., Luthey-Schulten, Z. and Wolynes, P.G.: Theory of protein folding: The energy landscape perspective, *Ann. Rev. Phys. Chem.* **48** (1997), 545–600.
14. Go, N.: Theoretical studies of protein folding, *Ann. Rev. Biophys. Bioeng.* **12** (1983), 183–210.
15. Bryngelson, J.D. and Wolynes, P.G.: Spin glasses and the statistical mechanics of protein folding, *Proc. Natl. Acad. Sci. USA* **84** (1987), 7524–7528.
16. Li, H., Helling, R., Tang, C. and Wingreen, N.: Emergence of preferred structures in a simple model of protein folding, *Science* **273** (1996), 666–669.
17. Bornberg-Bauer, E. and Chan, H.S.: Modeling evolutionary landscapes, *Proc. Natl. Acad. Sci. USA* **96** (1999), 10689–10694.
18. Sasai, M.: Conformation, energy, and folding ability of selected amino acid sequences, *Proc. Natl. Acad. Sci. USA* **92** (1995), 8438–8442.
19. Saito, S., Sasai, M. and Yomo, T.: Evolution of the folding ability of proteins through functional selection, *Proc. Natl. Acad. Sci. USA* **94** (1997), 12976–12979.

20. Hirst, J.D.: The evolutionary landscape of functional model proteins, *Protein Engineering* **12** (1999), 721–726.
21. Yomo, T., Saito, S. and Sasai, M.: Gradual development of protein-like global structures through functional selection, *Nat. Struct. Biol.* **6** (1999), 743–746.
22. Keefe, A.D. and Szostak, J.W.: Functional proteins from a random-sequence library, *Nature* **410** (2001), 715–718.
23. Yomo, T.: *Molecular evolution in a static and dynamical landscape*, an invited talk at the 4th ICBP, Kyoto, 2001.
24. Kissinger, C.R., Liu, B., Martin-Blanco, E., Kornberg, T.B. and Pabo, C.O.: Crystal structure of an engrailed homeodomain-DNA complex at 2.8Å resolution: A framework for understanding homeodomain-DNA interactions, *Cell* **63** (1990), 579–590.