Correlated Flexible Molecular Coding and Molecular Evolvability

Y. HUSIMI^{1,*}, T. AITA^{1,2} and I. TABUCHI¹

¹Department of Functional Materials Science, Saitama University, Saitama 338-8570, Japan

²Tsukuba Research Institute, Novartis Pharma K.K., Ohkubo 8, Tsukuba, 300-X611, Japan

(*Author for correspondence, e-mail: husimi@fms.saitama-u.ac.jp)

Abstract. Evolvability of biopolymers is based on molecular coding. The molecular coding is represented by biopolymer function vs monomeric sequence relationship, that is, a proper fitness landscape on the sequence space. On the other hand, molecular coding is mostly realized by monomeric sequence vs biopolymer structure relationship. We suggest the evolution of evolvability based on flexible or multiplex coding originating from flexible or polymorphic conformation of evolving biopolymers. We report a finding supporting that the amino acid landscape of the standard genetic code for an amino acid property which is more important to the protein function gives higher value of an evolvability measure. We developed a promising molecular construct which realized genotype-phenotype linking in order to study the *in vitro* protein evolution to clarify above mentioned protein evolvability.

Key words: fitness landscape, genetic code, *in vitro* virus, remote homolog proteins, sequence space

1. Introduction

Life is characterized by its high evolvability. Biological evolvability is demonstrated even at the molecular level as revealed by recent advancement in evolutionary molecular engineering. Evolvability of biopolymers is based on molecular coding. Evolution of the evolvability means evolution of molecular coding mechanism. The molecular coding is represented by biopolymer function vs monomeric sequence relationship, that is, a proper ‘fitness’ landscape on the sequence space. On the other hand, molecular coding is mostly realized by monomeric sequence vs biopolymer structure relationship, because its function is realized mostly by specific interaction originating from the complementarity between molecular surfaces.

If we accept the RNA world hypothesis, the first molecular coding was realized when some RNA function was encoded on its base sequence. Contrary to nucleic acids, evolution of proteins required the second coding mechanism, which required both the proper genetic code table and the genotype-phenotype linking strategy.

Evolvability of a protein is based on the structure of the protein ‘fitness landscape’ on a base sequence space (= DNA space), when random mutagenesis takes place at the base sequence level. Here, in evolutionary molecular engineering, the

concept of 'fitness' in biology is expanded to a quantitative measure of a certain physicochemical property of a protein (i.e. enzymatic activity, affinity to a ligand or structural stability). The structure of a fitness landscape on the DNA space follows the two elemental coding mechanisms. The first coding mechanism is the coding of a protein's function into an amino acid sequence, and is represented by the fitness landscape on the amino-acid sequence space (= protein space). The structure of a fitness landscape on the protein space is physicochemically determined by the primary structure and the environmental condition. Many experiments demonstrated statistical additivity in mutational effects in protein. These observations suggest that fitness landscapes on a local protein space around current protein are most likely the Mt. Fuji-type fitness landscape, of which we have theoretically studied (Aita and Husimi, 1996; Aita et al., 2000). The second coding mechanism is the coding of an amino acid by a triplet codon according to a genetic code table. Amino-acid allocations in the genetic code table are critical in linking the fitness landscape on the protein space to that on the DNA space. Therefore, we studied the organization of the standard genetic code (SGC) in terms of the amino-acid property 'landscape' on the 4-valued 3-dimensional sequence space (codon space) (Figure 1). A type of the landscapes satisfying the contradictory requirements (changeability and mutational robustness) is the Mt. Fuji-type landscape based on the mutational additivity. The height of the mountain represents the changeability, while the smoothness on the slope represents the mutational robustness. On the condition that a set of 20 property-values corresponding to 20 amino acids and a termination signal is given, the problem is what the beneficial allocation of these values to 64 codons is. We showed the mutational robustness for SGC based on apparent additivity of the contribution of individual letters in a codon to an amino-acid property (Aita, Urata, Husimi, 2000).

We also showed the correlation between the amino-acid property and the 'site-fitnesses' in order to clarify the meaning of the mutational robustness of SGC. Here, the term 'site-fitness' is a free energy contribution from a certain amino acid residue at a site in an amino acid sequence, on the assumption of additivity of residue contribution.

We next examined the effectiveness through the climbability and stability for adaptive walkers on fitness landscapes in DNA space for a model peptide where the fitness is defined as the sum of the site-fitnesses (Aita, Urata, Husimi, 2000). Another study was done by Arden and Sella (Ardell and Sella, 1999), in terms of the quasispecies theory (Eigen et al., 1989). They suggested that individuals with codes that are more error-correcting with respect to mutation will have higher fitness than those with less correcting codes.

In the first half of this article, we describe another finding supporting that the amino acid landscape of the SGC for a more important amino acid property for the protein structure/function gives higher value of an evolvability measure. We also describe the concept of flexible and/or multiplex coding originating from flexible or polymorphic conformation of evolving biopolymers.

Figure 1. Hydropathy landscape on the codon space (4-valuated 3-dimensional sequence space). Diameter of each solid circle is proportional to the hydropathy index (Kyte) of the amino acid encoded by the base triplet. Main axis: 2nd letter of codons (UCGA). Second axis: 1st letter of codons (GUAC). Third axis: 3rd letter of codons (UCAG). This landscape is nearly Mt. Fuji-type.

In the last half, we review our recent work on the genotype-phenotype linking strategy. We developed a very promising molecular construct to study the *in vitro* protein evolution to clarify above mentioned protein evolvability.

2. Correlated and flexible molecular coding

2.1. CORRELATION BETWEEN AMINO-ACID PROPERTY AND CONSERVATIVENESS IN REMOTE HOMOLOG PROTEINS

Amino acid allocation in the genetic code table maps the fitness landscape on the amino-acid-sequence space to that on the base-sequence space, where random mutation actually takes place. Therefore we studied the mapping correlatedness of

the standard genetic code (SGC) in terms of the amino acid property 'landscape' on the four-valued three-dimensional sequence space (codon space). Landscapes of 11 properties were analyzed. Correlatedness for each landscape was evaluated by θ_{SGC} -value, which represents the ratio of the mean slope to the degree of roughness. The θ_{SGC} -values for hydrophathy indices, polarity, specific heat, and β sheet propensity were considerably large for SGC judging from Z-score against randomly shuffled codes. We showed the simultaneous satisfaction of mutational robustness in various properties (correlated coding). This was beyond the expectation from the structural regularity due to synonymous codon blocks in the SGC table.

The large θ -values for SGC and their large Z-scores suggest that several amino acid properties such as hydrophathy have significantly affected the functions and physicochemical properties of proteins, in their evolution. As this point is not self-evident, we examined correlations between the θ_{SGC} -value and the actual 'site-fitnesses' of a protein (Aita, Urata, Husimi, 2000). The site-fitnesses were derived from a set of binding preference scores of amino acid residues at every site in MHC class I molecule binding properties. We found as expected that the SGC's θ_{SGC} -value for an amino acid property is correlated with the significance of the property in the protein function.

As another finding in this context, we found the correlations between the SGC's θ_{SGC} -value for an amino acid property and conservativeness of the property in remote homologue proteins, as follows.

Remote homolog proteins are those proteins, of which sequence homologies are very poor (less than 25%) but conformations are nearly identical. They are compiled in the database FSSP.

Statistics of the variety of amino acids at each corresponding site in the main chain structure among remote homologs leads to conservativeness of the amino acid properties. Eight kinds of the amino acid properties are analyzed as shown in Figure 2. We chose 22 representative structures from FSSP database and analyzed remote homologs for each of them. Calculation procedure for the conservativeness is as follows. You calculate the standard deviation σ_j of the distribution of the value of amino acid property k for each site j of the structure 'A'. Next you shuffle randomly the assignment between amino acid and the value of the property and calculate the standard deviation σ'_j . You determine the distribution of σ'_j as the background distribution. You calculate Z-score z_j for σ_j against the background distribution. Now the conservativeness $\bar{z}$ of amino acid property k for the structure A is defined as the mean of z_j over all the sites.

In Figure 2, the correlation between $\bar{z}$ and θ_{SGC} -value is shown for 8 properties of amino acids. It shows positive correlation just as site-fitness. It was the globin fold (2FAL) that has the maximum correlation coefficient and its value was $r = 0.86$ ($p = 0.006$). Average value of correlation coefficient over 22 folds was $r = 0.64$ ($p = 0.09$). Base sequence homology is about 25% among genes for 5 remote homologs of the globin fold. As this value is equivalent to homology among random sequences, the positive correlation as shown in Figure 2 is not a direct

Figure 2. Correlation between conservativeness of amino acid properties in remote homolog proteins ($\bar{z}$) and θ_{SGC} -value of standard genetic code. 'o': Transfer free energy from water to octanol ΔG_{oct} , 'x': Transfer free energy from water to cyclohexane ΔG_{chx} , 'v': Molecular volume, 's': Specific heat, 'a': α -helix propensity, 'b': β -sheet propensity, 'h': Sweet's hydropathy as positive control, 'n': Alphabetical order as negative control. The deepness of the letters is proportional to the correlation coefficient for each of 22 folds. Bold letters correspond to the globin fold (2FAL).

consequence from the robustness of SGC against mutation. The conservativeness of the property k indicates the general importance of the amino acid property in the physicochemical property of the protein. We confirmed the positive correlation between the importance and the measure of evolvability θ_{SGC} .

Random mutation in the base sequence space is mapped into the correlated mutation in amino acid sequence space through SGC. Therefore, Mt. Fuji-type landscape of SGC plays the role of a guardrail set near the sharp ridge of the protein landscape. Moreover, it is speculated that the correlated mapping of SGC reduces the ruggedness of a protein landscape, that is, the mutational change in non-additive terms of a protein fitness. Thus proteins make steadier adaptive walk

than in the case without second correlated coding. This is an example of evolution of evolvability of biopolymers.

2.2. 'KAMIKOCHI'-SEQUENCE AS AN EXAMPLE OF FLEXIBLE/MULTIPLEX CODING

Evolution of single biopolymers in the laboratory is described as an adaptive walk on the fitness landscape on the sequence space. Mutational analysis showed that contemporary proteins are located not at the peak, but, say, at the eight tenth of ascending path of the landscape. The consistency principle or the principle of minimum frustration stated that contemporary proteins are characterized as fast folding to the unique conformation specified by its amino acid sequence. We consider a point climbing down somewhat on the landscape. It is reported that a germ-line immunoglobulin is somewhat flexible and becomes tight during maturation process, judging from X-ray structural analysis (Wedmayer et al., 1997; Furukawa et al., 1999). Many data of *in vitro* evolution of RNA or ssDNAs from the random sequence pool showed that they are structurally polymorphic (Schmitt and Lehman, 1999; Schultes and Bartel, 2001; Eli et al., in preparation) in contrast with contemporary tRNAs. Thus it is speculated that fast folding to a flexible and/or multiplex conformation is characteristic to evolving biopolymers. We call a monomeric sequence of such a biopolymer as 'Kamikochi'-sequence (Kamikochi in the Japan Alps is the name of a highland from which you can climb up many peaks.) To reach such a sequence means the evolution of evolvability. Fast genetic adaptability may be attained through evolution of Kamikochi-sequence rather than evolution of individual biopolymers of a specific function. Searching for efficient Kamikochi-sequences is important because they serve as a starting sequence in evolutionary molecular engineering for enzyme design.

3. *In Vitro* DNA Virus for Genotype-Phenotype Linking

Darwinian selection process for proteins or peptides requires some linking strategy between phenotype and genotype in order to give the genetic feedback of the fitness. In other word, the second molecular coding required the establishment of such a linking mechanism. There are at least three such strategies both in laboratory and in Nature (Nemoto and Husimi, 1995): virus-type, cell-type, and external intelligence-type. In virus-type linking strategy, the genotype molecule (DNA or RNA) is bound to the phenotype molecule (protein) directly just as in a simple bacteriophage particle (virion). A phage display (Smith, 1985) and a cellstat (Husimi et al., 1982) were developed for the study of molecular evolution in laboratory using this strategy. *In vitro* version of these methods was developed under the name of '*in vitro* virus' (Nemoto et al., 1997) or 'mRNA-peptide fusion' (Roberts and Szostak, 1997). We called it *in vitro* virus because it can be regarded as the simplest virus with a single gene and a single coat protein, and its 'life cycle' is

turned over *in vitro*, that is, its 'host' is a test tube. Main reason is, of course, its genotype-phenotype bonding strategy other than the compartmentalization strategy of a cellular organism. By attaching puromycin to 3' end of an mRNA, a covalent linkage between a nascent protein and its encoding mRNA was made on a ribosome during *in vitro* translation. It increased dramatically the library diversity and the system flexibility. In fact, functional proteins have been evolved starting from a random library in this approach (Keefe and Szostak, 2001).

Classical *in vitro* virus had some problems; for example, it was required to modify the template mRNA to attach puromycin linker; the template could not tolerate harsh selective conditions because the genotype RNA molecule is fragile. Recently, some *in vitro* selection systems consisting of cDNA-peptide fusion have been developed to overcome the latter drawback (Kurz et al., 2001). But because these methods also required enzymatic or chemical modifications of evolving template, it was difficult to make an autonomously evolving molecular system.

We developed a new type of *in vitro* DNA virus (Tabuchi et al., 2001). Virion formation of our *in vitro* DNA virus consisted of following processes as shown in Figure 3: hybridization of a reverse transcription (RT)-primer with a puromycin-linker to mRNA, *in vitro* translation in a wheat germ extract and reverse transcription, which lead to cDNA-protein fusion. We developed a RT-primer with a puromycin-linker named T-linker-primer.

The T-linker-primer was synthesized by the use of a standard DNA synthetic method. It was synthesized in the first step as two separate parts by a DNA synthesizer and each primary amine on the two parts were linked together via bisfunctional N-hydroxysuccinimid ester. 5' protection of the puromycin-linker moiety of the T-linker-primer (5'-NH₂-(spacer18)₄-(CC)OMe-puromycin-3'-CPG) and base protections of the RT-primer moiety of T-linker-primer (5'-(GCCCCCCG)_{OMe}(T-NH₂)CCC-3') were deprotected by standard protocols. The puromycin-linker moiety of the T-linker-primer was activated by FluoroLink Cy3 reactive dye in DMSO and washed with DMSO. DNA-primer moiety was added and incubated in DMSO and washed with DMSO and TE buffer. Concentrated NH₄OH was added to remove the linker-primers from CPG and deprotection of bases was performed. The linker-primers were purified with a reverse phase cartridge.

In vitro DNA virus consists of stable cDNA-protein fusion without any modification of the evolving template pool. We demonstrated efficient virion formation of the *in vitro* DNA virus and selective enrichment in a test system consisting of a protein with two types of affinity tag.

In vitro DNA virus virion consists of a cDNA/mRNA hybrid double helix and a protein bound covalently to the cDNA. This construct must be more stable than mRNA-peptide fusion, and tolerate harsh and stringent selection conditions, such as temperature stress-guided selection. If higher stability will be required, the virion should be converted to a double-stranded DNA.

As it is free from necessity of mRNA modifications, *in vitro* virus will take the next step toward autonomous evolution, such as *in vivo* viral evolution in a cellstat.

Figure 3. 'Life cycle' of 'in vitro DNA virus'. *In vitro* virus 'genome' (dsDNA) is transcribed to mRNA. The mRNA is hybridized to a RT-primer with a puromycin-linker (T-linker-primer) and translated *in vitro* (in the wheat germ extract). On the ribosome, puromycin is incorporated to C-terminus of the nascent protein. Reverse transcription starting from the RT-primer gives a stable cDNA-protein fusion (*in vitro* virus 'virion'). The DNA moiety of screened virions judging from the 'fitness' of the protein moiety is amplified.

Combining the *in vitro* DNA virus with an isothermal amplification method for RNA/DNA such as 3SR (self-sustained sequence replication), a continuous 'culture' of the *in vitro* virus will be realized. A flow reactor for this viral life-cycle reaction will be a 'natural selection'-type evolution reactor (Husimi et al., 1982) for a protein evolution.

Acknowledgements

A part of this work was performed as a part of the R&D Project of the Industrial Science and Technology Frontier Program supported by NEDO (New Energy and Industrial Technology Development Organization). Authors thank Dr N. Nemoto (GenCom Co.) for his discussions and support.

References

1. Aita, T. and Husimi, Y.: Fitness Spectrum among Random Mutants on Mt. Fuji-type Fitness Landscape. *J. Theor. Biol.* **182** (1996), 469–485.
2. Aita, T., Uchiyama, H., Inaoka, T., Nakajima, M., Kokubo, T. and Husimi, Y.: Analysis of a Local Fitness Landscape with a Model of the Rough Mt. Fuji-type Landscape: Application to Prolyl Endopeptidase and Thermolysin, *Biopolymers* **54** (2000), 64–79.
3. Aita, T., Urata, S. and Husimi, Y.: From Amino Acids Landscape to Protein Landscape: Analysis of Genetic Code in terms of Fitness Landscape, *J. Mol. Evol.* **50** (2000), 313–323.
4. Arden, D.H. and Sella, G.: The Impact of Message Mutation on the Fitness of a Genetic Code, *Preliminary Proceedings DIMACS Workshop on Evolution as Computation*, 1999, pp. 175–180.
5. Eigen, M., McCaskill, J.S. and Schuster, P.: The Molecular Quasispecies, *Adv. Chem. Phys.* **75** (1989), 149–263.
6. Eli, P., Tabuchi, I., Iizuka, F. and Husimi, Y.: unpublished data.
7. Furukawa, K. et al.: Junction Amino Acids Determine the Maturation Pathway of an Antibody, *Immunity* **11** (1999), 329–338.
8. Husimi, Y.: Selection and Evolution of Bacteriophages in Cellstat, *Adv. Biophys.* **25** (1989), 1–43.
9. Keefe, A.D. and Szostak J.W.: Functional Proteins from a Random-Sequence Library, *Nature* **410** (2001), 715–718.
10. Kurz, K., Gu, K., Al-Gawara, A. and Lohse, P.: cDNA-Protein Fusion, *Chem. Bio. Chem.* **2** (2001), 666–672.
11. Nemoto, N. and Husimi, Y.: A Model of Virus-Type Strategy in the Early Stage of Encoded Protein Evolution, *J. Theor. Biol.* **176** (1995), 67–77.
12. Nemoto, N., Miyamoto-Sato, E., Husimi, Y. and Yanagawa, H.: *In vitro* Virus, *FEBS Lett.* **414** (1997), 405–408.
13. Roberts, R.W. and Szostak, J.W.: mRNA-Protein Fusion, *Proc. Natl. Acad. Sci. USA* **94** (1997), 12297–12300.
14. Schmitt, T. and Lehman, N.: Non-Unity Molecular Heritability Demonstrated by Continuous Evolution *in vitro*, *Chem. Biol.* **6** (1999), 857–869.
15. Schultes, E.A. and Bartel, D.P.: One Sequence, Two Ribozymes: Implications for the Emergence of New Ribozyme Folds, *Science* **289** (2001), 448–452.
16. Smith, G.P.: Filamentous Fusion Phage: Novel Expression Vectors that Display, Cloned Antigens on the Virion Surface, *Science* **228** (1985), 1315–1317.
17. Tabuchi, I., Soramoto, S., Nemoto, N. and Husimi, Y.: An *in vitro* DNA Virus for *in vitro* Protein Evolution, *FEBS Lett.* **508** (2001), 309–312.
18. Wedmayer, G.J. et al.: Structural Insights into the Evolution of an Antibody Combining Site, *Science* **276** (1997), 1665–1669.