J Mol Evol (2000) 50:313–323
DOI: 10.1007/s002399910035

JOURNAL OF MOLECULAR EVOLUTION

From Amino Acid Landscape to Protein Landscape: Analysis of Genetic Codes in Terms of Fitness Landscape

Takuyo Aita,[1,2] Satoshi Urata,[1] Yuzuru Husimi[1]

[1] Department of Functional Materials Science Saitama University Urawa 338, Japan
[2] Takarazuka Research Institute, Novartis Pharma K. K., 10-66, Miyuki-cho, Takarazuka, 665 Japan

Received: 7 June 1999 / Accepted: 15 December 1999

Abstract. Assigning the values of a certain physicochemical property for individual amino acids to the corresponding codons, we can make an amino acid property "landscape" on a four valued three dimensional sequence space from a genetic code table. Eleven property landscapes made from the standard genetic code (SGC) were analyzed. The evaluation of correlation for each landscape is done by θ value, which represents the ratio of the mean slope (as an additive term) to the degree of roughness (as a nonadditive term). The θ-values for hydropathy indices, polarity, specific heat, and β-sheet propensity were considerably large with respect to SGC. This implies that the additivity of the contribution from each letter holds for these properties. To clarify the meaning of the so-called mutational robustness of SGC, we next examined correlations between the amino acid property and the actual "site fitnesses" of a protein. 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 peptides (Udaka et al. in press). We found that the SGC's θ value for an amino acid property is correlated with the significance of the property in the protein function. Adaptive walk simulation on fitness (= affinity) landscapes in a base sequence space for these model peptides confirmed better evolvability due to the introduction of SGC.

Key words: Standard genetic code — Mutational robustness — Evolvability — Adaptive walk — Codon — Evolutionary molecular engineering

Correspondence to: Yuzuru Husimi; *e-mail:* husimi@fms.saitama-u.ac.jp

Introduction

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. 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 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 proteins (e.g., Wells 1990). These observations suggest that fitness landscapes on a local protein space around current proteins are most likely the Mt. Fuji–type fitness landscape, 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 four-valued three-dimensional sequence space (codon space).

A vast number of studies have been made on the natural genetic code, including the SGC. Particularly, the mutational robustness against base substitutions in SGC has been studied in various ways, where the "mutational robustness" means that amino acids that have similar physicochemical properties also have similar codons (e.g., Haig and Hurst 1991; Di Giulio 1997; Freeland and Hurst 1998; Trinquier and Sanejouand 1998). Some research focused on what the most optimized code is (Di Giulio et al. 1994) and evaluated the degree of the optimization for SGC (Wong 1980; Di Giulio 1989). The ideal genetic code seems to have a wide dynamic range for each of the important properties and a mutational robustness against base substitution. 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, and the smoothness on the slope represents the mutational robustness. On the condition that a set of 20 property values and a termination signal is given, the problem is what the beneficial allocation of these values to 64 codons is. Our approach to the mutational robustness for SGC is based on apparent additivity of the contribution of individual letters in a codon to an amino acid property. Thus, in the first half, we analyzed amino acid property landscapes for 11 typical properties, including 3 hydropathy indices using a model of the Mt. Fuji–type landscape (Aita et al. 2000).

In the latter half, we studied the correlation between the amino acid property and the "site fitness" to clarify the meaning of the mutational robustness of SGC. Here, the term *site fitness* is a free energy contribution (divided by *RT*) from a certain amino acid residue at a site in an amino acid sequence on the assumption of additivity of residue contribution. Using a set of "binding preference scores" of amino acid residues at every site of MHC class I molecule binding peptides (Udaka et al. 1999) as a set of model site fitnesses, we examined the correlation between the amino acid property and the site fitness and linked the correlation with the θ values for SGC. The usefulness of these scores was verified by Udaka et al. (1999), who have succeeded in predicting the affinity of an arbitrary peptide to a given MHC class I molecule using these scores.

Our interest focuses on the degree to which the SGC provides the benefit on protein's evolution, rather than on what the most optimized code possible is (Di Giulio et al. 1994) or on the origin of the genetic code (e.g., Di Giulio 1997). Then we examined the effectiveness through the climbability and stability for adaptive walkers on fitness landscapes in DNA space for the MHC class I molecule binding peptides, where the fitness is defined as the sum of the site fitnesses. Another study was done by 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. We also confirmed the effectiveness of SGC on evolvability for the realistic model peptides.

Analysis Method of an Amino Acid Property Landscape in the Codon Space

The codon space is the four-valued three-dimensional sequence space comprising of all codons. A base and an amino acid are denoted by β (β ∈ {a,u,g,c}) and α (α ∈ {A,C,D, . . . ,Y}), respectively. An amino acid, which is assigned to a certain codon C, is denoted by α(C). A value of a particular physicochemical property, such as hydropathy or volume, for an amino acid α is denoted by $\mathcal{V}(\alpha)$. We used the centered and autoscaled data of the amino acid properties. Centering was done by subtracting the averages from all data, and autoscaling was conducted by the division of each variable by its standard deviation. The amino acid property for a codon C is defined by $\mathcal{V}_C = \mathcal{V}(\alpha(C))$. An amino acid property landscape in the codon space is defined as a set of $\mathcal{V}_C$.

Let $\{C_1, C_2, \ldots, C_n\}$ be a set of all codons except stop codons, where $n = 61$ for the standard genetic code (SGC). We carried out the linear regression analysis of an amino acid property landscape for a genetic code, using the following model. By introducing a quantity $\upsilon_l(\beta)$, that is an apparent contribution to the amino acid property from a base β at the *l*th ($l = 1, 2, 3$) letter, an additive part V_C of a $\mathcal{V}_C$ for a codon C is defined as

$$V_C = V_O + \sum_{l=1}^{3} \upsilon_l(\beta_{Cl}) \qquad \text{(Eq. 1)}$$

O is a particular codon that corresponds to the peak on this model landscape (where we call this landscape "Mt. Fuji–type"). β_{Cl} represents the base at the *l*th letter in a codon C.

The base sequence of the codon O and unknown parameters V_O and $\upsilon_l(\beta)$ are determined by the least squares method to minimize the value of $\Sigma_{i=1}^{n}(\mathcal{V}_{C_i} - V_{C_i})^2$ (we have 10 unknowns). The values of $\upsilon_l(\beta)$ is assigned to satisfy

$$\upsilon_l(\beta) \begin{cases} = 0, & \text{if } \beta = \beta_{Ol} \\ < 0, & \text{if } \beta \neq \beta_{Ol} \end{cases} \qquad \text{(Eq. 2)}$$

where β_{Ol} denotes the base at the *l*th letter in the codon O. For the *l*th letter, the mean value of $\upsilon_l(\beta)$'s over three bases except β_{Ol} is denoted by ε_l. We define the "mean

Table 1. Characteristics of amino acid property landscapes for the standard genetic code

Index k	Property	r	O	V_O	$\langle\|\varepsilon_l\|\rangle$	σ	θ	Z_1	Z_2	Z_3
1	hydropathy[a]	0.90	uuu (Phe)	1.89	0.85	0.41	2.10	6.95	6.00	14.3
2	hydropathy[b]	0.92	guu (Val)	1.88	0.81	0.41	1.97	5.97	5.74	13.3
3	hydropathy[c]	0.80	guu (Val)	1.72	0.79	0.67	1.17	3.74	2.88	7.20
4	polarity	0.84	gaa (Glu)	1.38	0.63	0.48	1.31	3.52	2.73	7.22
5	volume	0.78	cug (Leu)	1.15	0.58	0.61	0.94	1.11	1.47	4.69
6	ASA	0.71	cag (Gln)	1.10	0.56	0.71	0.79	0.03	0.80	3.36
7	mass	0.74	cag (Gln)	1.04	0.54	0.66	0.83	−0.01	0.94	3.58
8	specific heat	0.78	cua (Leu)	1.68	0.74	0.63	1.18	2.49	2.30	6.87
9	isoelectric point	0.69	cga (Arg)	1.80	0.73	0.77	0.94	1.12	1.33	4.52
a	α helix	0.61	gug (Val)	1.27	0.61	0.77	0.80	0.83	0.76	3.50
b	β sheet	0.77	uuc (Phe)	1.53	0.69	0.63	1.10	2.25	2.18	6.24
D_m								8.58	8.23	18.3

The 11 amino acid properties were chosen: three hydropathy indices (a: Sweet and Eisenberg 1983; b: Kyte and Doolittle 1982; c: Eisenberg et al. 1982); polarity (Woese et al. 1966); volume (Zamyatnin 1975); accessible surface area (ASA; Chothia 1976); mass; specific heat (Privalov et al. 1989); isoelectric point (Alff-Steinberger 1969); α helix propensity; and β sheet propensity (Chou and Fasman 1978). The particular codon O corresponds to the peak on the model landscape. Z_m represents the Z score of the SGC's θ value, based on a frequency distribution of the θ values obtained from 1,000 randomly generated variant codes ((m, ∞)-VC) according to shuffling manner m (see text). r is a correlation coefficient between the original value ($\mathcal{V}_C$) and additive one (V_C). D_m represents the Mahalanobis's generalized distance (see Appendix A). Other notations are defined in the text.

slope" of the amino acid property landscape as follows. Consider the mean change in the property when, in a certain codon, a certain base except β_{Ol} is replaced by β_{Ol}. The mean change in the property is averaged over all possible codons except stop codons in the codon space. The mean slope is defined as this doubly averaged value and is approximately given by $\langle|\varepsilon_l|\rangle = 1/3 \sum_{l=1}^{3} |\varepsilon_l|$.

If the residual $\mathcal{V}_{C_i} - V_{C_i}$ ($i = 1, 2, \ldots, n$) after the fitting procedure is small and distributes randomly in the codon space, according to a near Gaussian distribution with mean 0 and standard deviation σ, then we regard this landscape as being a rough Mt. Fuji–type that has the mean slope of $\langle|\varepsilon_l|\rangle$, and the roughness of σ. We introduce an index $\theta \equiv \langle|\varepsilon_l|\rangle/\sigma$, that is the ratio of the mean slope to the roughness. Completing the above procedures, we can identify "landscape properties," such as $\langle|\varepsilon_l|\rangle$, σ, and θ.

It is desirable that the mean slope $\langle|\varepsilon_l|\rangle$ is large to give a wide dynamic range to various protein characters, while it is also desirable that the roughness σ is small to hold conservativeness or robustness to base substitution. Therefore, ideal genetic codes should take large θ values. We use the θ value as a measure of evaluation for a genetic code.

Analysis of the Standard Genetic Code

θ *Value for the Standard Genetic Code*

We selected the following 11 amino acid properties that seem important for protein properties: three hydropathy indices (Eisenberg et al. 1982; Kyte and Doolittle 1982; Sweet and Eisenberg 1983), polarity (Woese et al. 1966), volume (Zamyatnin 1975), accessible surface area (Chothia 1976), mass, specific heat (Privalov et al. 1989), isoelectric point (Alff-Steinberger 1969), α helix propensity, and β sheet propensity (Chou and Fasman 1978). Based on the method described above, we analyzed the 11 amino acid property landscapes for the SGC. Sweet's and Kyte's hydropathy indices were derived, respectively, from Dayhoff's mutation matrix of amino acid replacement and an average of physicochemical properties of amino acids and spatial environmental data of proteins residues. Eisenberg's hydropathy index is roughly proportional to the free energy required to transfer an amino acid residue from the interior to the surface of a water soluble protein. Polarity (polar requirement) was derived from the R_F values in amino acid chromatography. The three hydropathy indices and polarity are mutually correlated by $|r| > 0.7$, where r is a correlation coefficient. The scales of volume, accessible surface area, and molecular weight are mutually correlated by $r > 0.91$. The scale of specific heat is correlated with that of accessible surface area ($r = 0.64$) and volume ($r = 0.79$). The scale of β sheet propensity is correlated with that of Kyte's hydropathy ($r = 0.67$), Sweet's hydropathy ($r = 0.82$), and polarity ($r = -0.77$). Other combinations do not show high correlation.

The result from analysis of the 11 property landscapes is shown in Table 1. It is obvious that the three hydropathy landscapes and polarity landscape have large θ values ($\theta = 1.1 \sim 2.1$); that is, their surfaces are considerably correlated not only locally (within each codon block composed of synonymous codons) but also globally (over all codons). This result is consistent with the well-known observations as mutational robustness in SGC (e.g., Haig and Hurst 1991). Taking the Kyte's hydropathy landscape as a representative case, we show the overview of the landscape in Fig. 1A and the distribution of

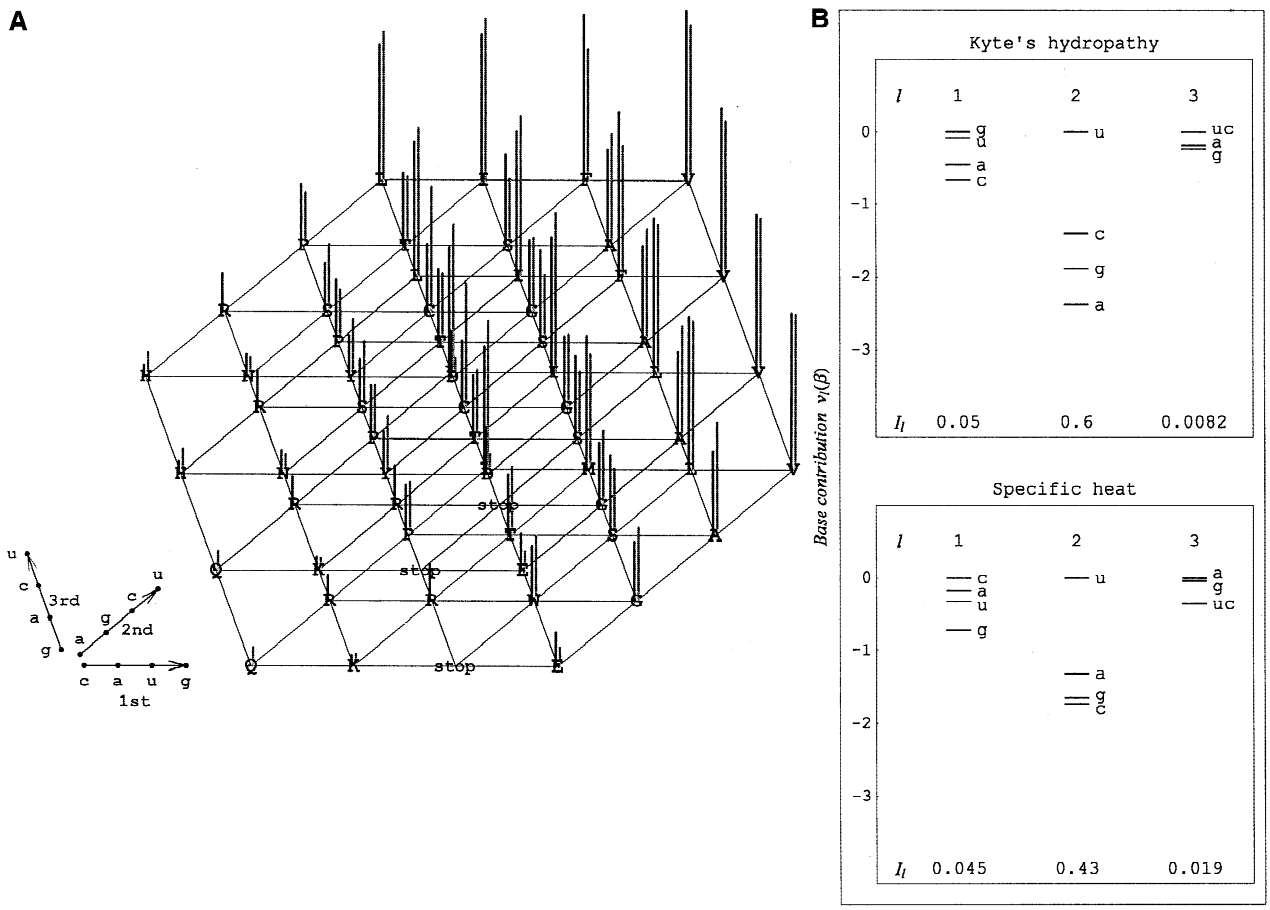


Fig. 1. Results of analysis of the Kyte's hydropathy landscape for the standard genetic code (SGC). **A** Overview of the Kyte's hydropathy landscape on the codon space. The codon space is projected on the plane of the page. Each codon is located at the corresponding node, and each of the amino acids allocated to the codons is represented by the corresponding one-letter abbreviation. "stop" represents the stop codon. Single point mutations are represented by the vectors shown as arrows. The amino acid property of a codon is represented by a bar standing at the corresponding node (arbitrary unit). The left side bars and right side bars are original values ($\mathcal{V}_C$) and additive values (V_C) obtained from the analysis, respectively. The correlation coefficient between them is 0.92. Note that the ordering of letters for each axis is intentional to show the mountainous structure, however, the ordering does not affect the results in this study at all. **B** Distribution of the contributions ($v_l(\beta)$) from each of the three letters for Kyte's hydropathy (top) and specific heat (bottom). I_l is the site information for the lth letter (Aita and Husimi 1996).

$v_l(\beta)$ for each letter in Fig. 1B. The most sensitive letter is the second one ($l = 2$), whereas the insensitive letter is the third one ($l = 3$). These results are consistent with the well-known observation. We calculated the virtual V_{C_i} values for the three stop codons from Eq. 1 with the obtained parameters for Kyte's hydropathy: uaa $\rightarrow$ −0.78; uag $\rightarrow$ −0.83; and uga $\rightarrow$ −0.28. These values are hydrophilic.

The characteristics for other landscapes are qualitatively similar to those shown in Fig. 1. We found no significant correlation between the residuals ($\mathcal{V}_{C_i} - V_{C_i}$) and codon sequences.

Z *Score for the Standard Genetic Code*

To compare the θ value for the SGC with those of other genetic codes, we generated a set of 1,000 randomly sampled variant codes by partial or complete shuffling of the amino acid allocations in the SGC and subsequently calculated the θ value for each of the variant codes by using the analysis method mentioned previously and determined the frequency distribution for the θ values. A variant code was generated by repetition of interchanging two amino acid allocations in the SGC, where the interchanging process is performed in either of the following three manners.

Manner 1 (m_1): Block Interchange with Restriction. In the SGC, each amino acid $\alpha \in \{A, C, D, \ldots, Y\}$ is allocated in a codon block composed of synonymous codons. The mutual interchange of two amino acid allocations is performed by these blocks under the condition that the shuffling process conserves the degeneracy of each amino acid. The codon space is partitioned by the blocks defined in the SGC, where the six synonymous codons for serin are dealt with as a single block and the translation termination signal is regarded as the 21st phenotype "Z" participating in the shuffling. Let G_k be a set of amino acids having the same degeneracy of k: G_1 =

{M, W}; G_2 = {C, D, E, F, H, K, N, Q, Y}; G_3 = {I, Z}; G_4 = {A, G, P, T, V}; and G_6 = {L, R, S}. $|G_k|$ is the number of elements in G_k. First, we pick out G_k among {G_1, G_2, G_3, G_4, G_6} with a probability of P_k, where P_k is given by

$$P_k = \binom{|G_k|}{2} \Big/ \sum_k \binom{|G_k|}{2}$$

Next, we pick out two arbitrary amino acids from G_k and interchange their allocations.

Manner 2 (m_2): Block Interchange Without Restriction. The interchange of two amino acid allocations is performed by the block, as in Manner 1, except that the interchange of two amino acid allocations is unrestricted and "Z" does not participate in the shuffling. Thus, the degeneracy of each amino acid can drastically alter through the shuffling process.

Manner 3 (m_3): Interchange by a Codon Unit. The interchange of two amino acid allocations is performed by a codon unit. Thus, the degeneracy of each amino acid is conserved through the shuffling process.

A variant code made by repetition of the interchanging operation according to manner m ($m = m_1, m_2, m_3$) by d times is denoted by (m, d)-VC (d represents something like "distance" between the SGC and each of the variant codes). We generated a set of 1,000 (m_1, d)-VCs for each of d = 1, 2, 3, 4, 5, and 200 and obtained distributions of θ values calculated from these variant codes. Figure 2 shows these distributions for Kyte's hydropathy, specific heat, α helix propensity, and β sheet propensity. The θ values for (m_1, d)-VCs tend to decrease as d increases. This suggests that the SGC is considerably optimized with respect to the four amino acid properties. Furthermore, we obtained θ values for several natural deviant codes (Fig. 2). In general, θ values for these deviant codes are close to that for the SGC, while the code for yeast mitochondria has very low values.

The θ value distributions for variant genetic codes ((m, d)-VC) tend to a steady distribution as d increases. This steady distribution is almost attainable when $d > 20$. A completely shuffled code when $d \to \infty$ is denoted by (m, ∞)-VC. Based on a frequency distribution of the θ values obtained from 1,000 randomly generated (m, ∞)-VCs, we define Z score for the θ value of SGC, with regard to each shuffling manner (m):

$$Z_m \equiv \frac{\theta_{SGC} - E[\theta_{(m,\infty)\text{-VC}}]}{SD[\theta_{(m,\infty)\text{-VC}}]} \qquad \text{(Eq. 3)}$$

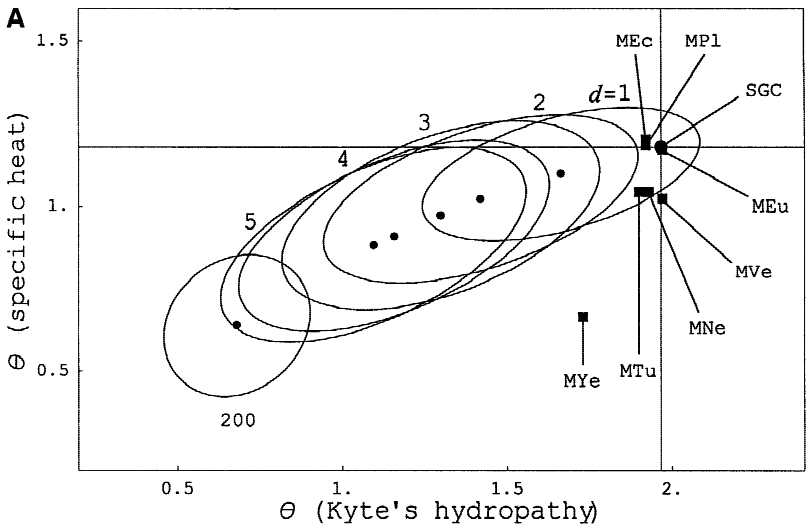
where θ_{SGC} and $\theta_{(m,\infty)\text{-VC}}$ are the θ value of the SGC and that of a (m, ∞)-VC, respectively. E[] and SD[] are the mean and standard deviation for a set of 1,000 (m, ∞)-VCs. Three types of Z scores (Z_1, Z_2, and Z_3) are shown in Table 1 for the 11 amino acid properties. The three hydropathy indices—polarity, specific heat, and β sheet propensity—show $Z_1 > 2.0$ and $Z_2 > 2.0$. This means that the smoothness of the landscape for the SGC is beyond the expectation from the structural regularity due to codon blocks composed of synonymous codons in SGC table (Maeshiro and Kimura 1998) and that the apparent rough additivity holds in mutational effects on these properties at the base level. The Z_1 seems slightly larger than the Z_2, with respect to seven properties. The reason is that the $SD[\theta_{(m_2,\infty)\text{-VC}}]$ for each property is much larger than its $SD[\theta_{(m_1,\infty)\text{-VC}}]$.

From Amino Acid Landscape to Protein Landscape: Case of MHC Class I Binding

Correlation Between Amino Acid Property and Site Fitness

The large θ values for the SGC and their large Z scores suggest that several amino acid properties, such as hydropathy (or polarity), have dramatically affected the functions and physicochemical properties of proteins in their evolution. In our previous study on Mt. Fuji–type fitness landscapes (e.g., Aita and Husimi 1996), we introduced the "site fitness," $w_j(\alpha)$, that is, a free energy contribution (divided by RT) from a certain amino acid residue α at the jth site in an amino acid sequence, on the assumption of additivity of residue contribution. For a protein or peptide in which the mutational additivity holds, the site fitnesses can be experimentally measured by using the positional scanning method (e.g., Houghten et al. 1991; Udaka et al. 1995). We examined correlations between site fitnesses ($w_j(\alpha)$) and amino acid properties ($\mathcal{V}(\alpha)$), by using a set of "binding preference scores" of amino acid residues at every site of MHC class I binding peptides (Udaka et al. 1995, 1999) as a set of model site fitnesses. The score is the experimental value obtained by the positional scanning method and defined as the logarithm of the molar concentration of the mixture of peptides that have bound half of the MHC class I molecules (defined as log SD_{50}). A set of the binding preference scores of 19 amino acids (except cysteine) at individual sites in the peptides is available for MHC class I molecule D^b, K^b, and L^d. The chain length of D^b-, K^b-, and L^d-binding peptides are 9-mer, 8-mer, and 9-mer, respectively. We gathered all the sites and numbered them as j = 1, 2, . . . , 26. The site fitnesses, $w_j(\alpha)$ (j = 1, 2, . . . , 26; α = A, D, . . . , Y), were derived from the binding preference scores with a slight modification. Details of the derivation are described in the Appendix B.

Subsequently, we calculated the "site information" I_j as a measure of tolerance to residue substitutions for



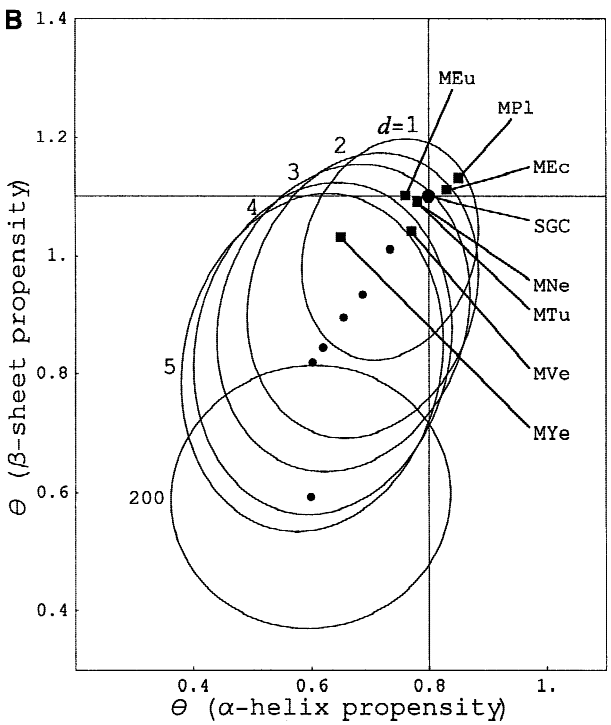
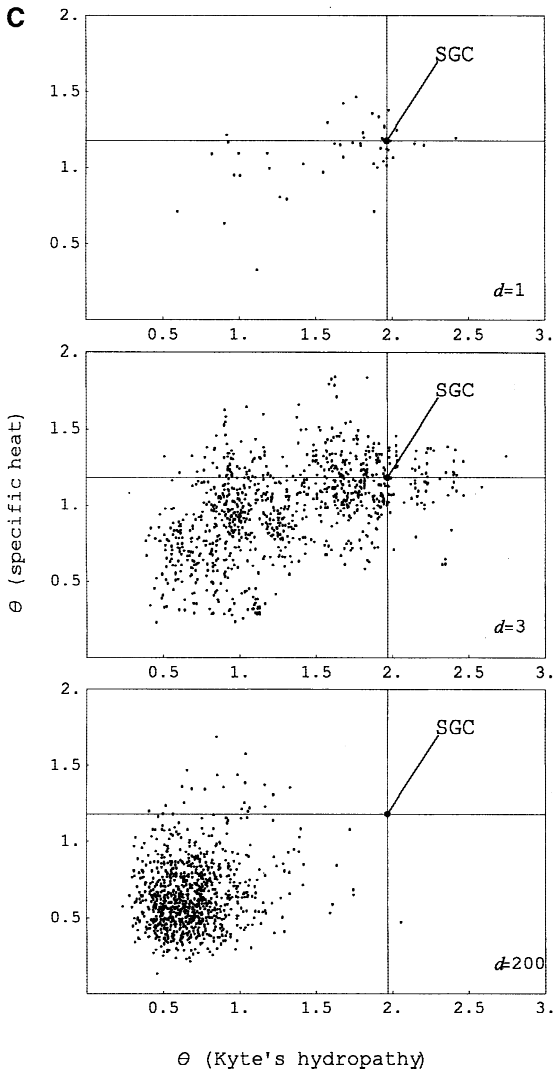


Fig. 2. Distribution for the θ values obtained from 1,000 (m_1, d)-VCs for each d value of d = 1, 2, 3, 4, 5, and 200. **A** Kyte's hydropathy versus specific heat, and **B** α helix propensity versus β sheet propensity. The dot represents the mean point of the θ distribution for each d value, and the ellipse surrounding the dot represents deviations of the first and second principal components. Because the number of all interchanges among amino acid allocations according to shuffling manner m_1 is only 51, the distribution for $d = 200$ is close to that for $d \to \infty$. The location for the standard genetic code is indicated by "SGC," and those for several mitochondrial codes are indicated by "MYe" (yeasts), "MPl" (platyhelminths), "MNe" (nematoda), "MEc" (echinodermata), "MTu" (tunicata) and "MVe" (vertebrata). **C** Details of the θ distribution for **A**. The number of samples of variant codes is 51 for $d = 1$ and 1,000 for $d = 3$ and $d = 200$.

each site (Aita and Husimi 1996). I_j is derived from a set of site fitnesses at the jth site and takes a value between 0 and $\log_2\lambda$, where λ is the number of residue types available at each site; $\lambda = 19$ in this case. Several "anchor" sites showing low tolerance to residue substitutions take the large values of site information ($I_j = 1.4 \sim 3.6$), while other tolerant sites take low values ($I_j = 0.1 \sim 0.8$). Udaka et al. (1995) used another measure of the tolerance and made the similar evaluation.

We calculated a correlation coefficient, r_j, for the data set $\{(\mathcal{V}(\alpha), w_j(\alpha))|\alpha = \text{A, D, . . . , Y}\}$ at each site, with respect to each of the 11 amino acid properties. It was obvious that the three hydropathy indices and polarity are outstandingly well correlated with the site fitnesses at half of the sites, including the anchor sites (Fig. 3). Furthermore, the mean and weighted mean of $|r_j|$ over 26 sites, $\langle|r_j|\rangle$, were calculated for each property. At a weighting operation, we used $I_j/\sum_{j=1}^{26} I_j$ as a weight for the jth site. The reason is that the correlation between site fitnesses and amino acid properties is more meaningful for critical sites taking large site information. Interestingly, irrespective of the mean or weighted mean, there is a positive correlation between $\langle|r_j|\rangle$ and θ_{SGC} for the nine properties, except Kyte's and Sweet's hydropathy indices (Fig. 3). The correlation coefficient obtained from the nine properties is 0.92 ($p < 0.001$). We also observed a similar correlation (correlation coefficient is 0.73–0.82) when using another set of site fitnesses, which were derived from Houghten et al. (1992; Dooley et al. 1993). It is likely that the conservativeness of an amino



acid property in SGC is correlated with the significance of the property in a protein function.

We note that our aim is not to study MHC class I binding peptides. Therefore, we omitted the details on the substantial characteristics for them. The details were discussed by Udaka et al. (1995, 1999).

Adaptive Walks on a Model Fitness Landscape in a Base Sequence Space

To examine the effect of amino acid allocations in the SGC or in other conceivable genetic codes on the evolution of proteins, we introduced a model fitness function, defined as the sum of the site fitnesses of the 26 sites, and carried out numerical simulations of adaptive walks on the model fitness landscape in the four-valued 78 ($= 26 \times 3$)-dimensional base sequence space. We define the fitness of a base sequence P as a concatenation of 26 codons by

$$W_{\mathrm{P}} = \sum_{j=1}^{26} w_j(\alpha(\mathrm{C}_{\mathrm{P}j})) \leq 0 \qquad \text{(Eq. 4)}$$

where $\mathrm{C}_{\mathrm{P}j}$ is the codon at the jth position in the concatenation P, and $\alpha(\mathrm{C}_{\mathrm{P}j})$ is the amino acid encoded by $\mathrm{C}_{\mathrm{P}j}$. The fitness W_{P} defined in Eq. 4 is considered on the binding free energy scale of a ligand peptide having a chain length of 26 to a virtual MHC-like receptor molecule: the dissociation constant between them is proportional to $\exp(-W_{\mathrm{P}})$. The rough additivity of the contribution of each of the amino acid residues in several MHC class 1 binding peptides was verified by Udaka et al. (1999). Therefore, we considered that the validity of using Eq. 4 as a model of the fitness function has been demonstrated. We note that the fitness landscape in the 19-valued, 26-dimensional amino acid sequence space is of the Mt. Fuji–type uniquely defined, while the fitness landscape in the 4 valued, 78-dimensional base sequence space is affected by the genetic code used. We adopted the following five cases with regard to genetic codes:

Code 0: SGC
Code 1: 1,000 examples of (m_1, 1)-VC
Code 2: 1,000 examples of (m_1, 2)-VC
Code 3: 1,000 examples of (m_1, 3)-VC
Code 4: 1,000 examples of (m_1, 5)-VC
Code 5: 1,000 examples of (m_1, ∞)-VC

In this article, we call the fitness landscape defined by SGC the "original landscape," and call a fitness landscape defined by a (m_1, d)-VC the "variant landscape." The question is whether the climbability and stability of adaptive walkers on the original landscape are larger than those on each of the variant landscapes.

Modeling of adaptive walks is critical in evaluating the climbability for adaptive walkers. In evolutionary molecular engineering, the environmental condition of an evolving biopolymer can be controlled (Husimi 1989) and any adaptive walk strategy can be taken. To simplify the phenomena, we used the (1, N)-ES as the adaptive walk strategies (Aita and Husimi 1998, 2000). This strategy follows a simple rule: a parent on a fitness landscape produces N descendants with single point mutations, and subsequently the fittest among the N descendants will become a new parent in the next generation (Rechendberg 1984). We set $N = 50$, $N = 10$, or $N = 3$ through a single adaptive walk process. Each walk was carried out starting from a randomly chosen sequence through 1,000 generations. When nonsense mutations and mutations to cysteine's codons occurred, we resolved these affairs by regenerating other mutants.

Let y_t be the scaled fitness of a single adaptive walker in the tth generation, where scaling is conducted by division of walker's fitness (W_{P} value) by |mean fitness of randomly chosen sequences| $= 19.4$. Figure 4 shows examples of time ($=$ generation) course of y_t on the original landscape, for each case of $N = 50$, $N = 10$, and $N = 3$. These time courses show alternation of plateau and epochal changes due to the existence of terraces of

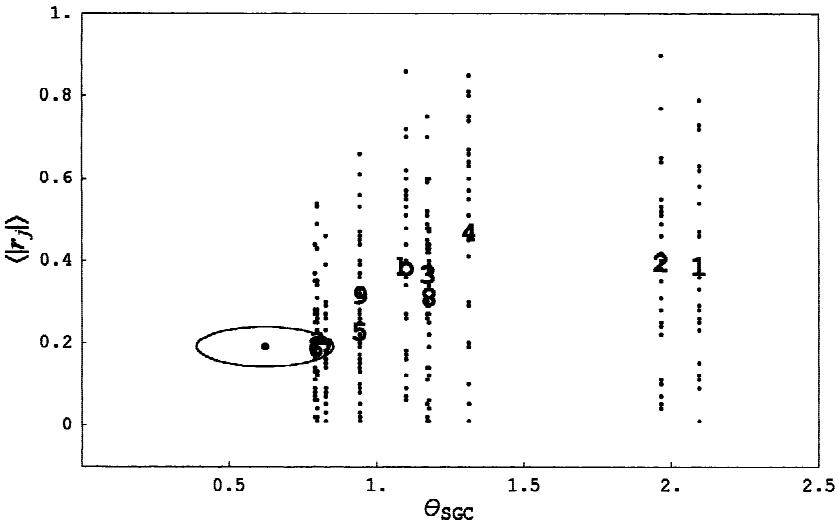


Fig. 3. Correlation between $\langle|r_j|\rangle$ and θ_{SGC}. r_j is a correlation coefficient between an amino acid property $\mathcal{V}(\alpha)$ and site fitness $w_j(\alpha)$ for the jth site in MHC class I binding peptides. $\langle|r_j|\rangle$ represents the mean of $|r_j|$ over 26 sites. The 26 values of $|r_j|$ ($j = 1, 2, \cdots, 26$) and $\langle|r_j|\rangle$ are plotted with small dots and the index characters defined in Table 1, respectively, against the θ_{SGC} value for each of the 11 amino acid properties. A single dot and the ellipse surrounding the dot represent the mean and standard deviation of distributions for randomly assigned property values, respectively. The correlation coefficients obtained from the nine properties except Kyte's and Sweet's hydropathy indices are 0.92 ($p < 0.001$).

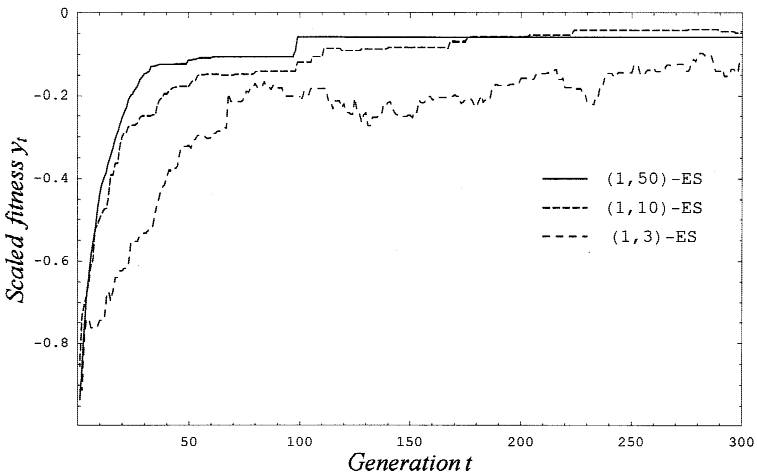


Fig. 4. Time courses of three adaptive walks on the original landscape (= fitness landscape defined by the standard genetic code) in the base sequence space. The abscissa is the generation t. The ordinate is the scaled fitness y_t.

synonymous codon blocks on the original landscape. Similar characteristics can be observed with respect to variant landscapes. Walkers taking the (1, N)-ES reached a steady state after 100–300 generations at the most because a mutation-selection balance sets in.

One thousand walks were carried out in each of *Code 0–Code 5* (where 1,000 walks were carried out for the original landscape [*Code 0*] and a single walk was carried out for each of the 1,000 variant landscapes [*Code 1–Code 5*]). We evaluated the climbability and stability for the stationary state of a single adaptive walk by

$$\overline{y_{ss}} \equiv \frac{1}{500}\sum_{t=501}^{1,000} y_t$$

$$\delta y_{ss} \equiv \sqrt{\frac{1}{500}\sum_{t=501}^{1,000} y_t^2 - \overline{y_{ss}}^2}$$

respectively. We averaged these values over 1,000 walks for each of *Code 0–Code 5*.

In Fig. 5A, the trial average of the climbability for each *Code* is plotted against the average of θ value for Kyte's hydropathy for the corresponding *Code*. Interestingly, there is a positive correlation between them, where the correlation coefficient is larger than 0.91 ($p < 0.01$) for each N value. The climbability on the original landscape seems better than that on most of the variant landscapes. The difference in y_t value by 0.01 corresponds to the change in association constant by 1.2-fold. The difference between the trial average of $\overline{y_{ss}}$ value for *Code 0* and that for *Code 5* is about 0.04–0.05, which corresponds to the change in association constant by 2.2–2.6-fold. If we assume the deterministic selection, this affinity gap seems large enough for the SGC to be selected among other competitors in the origin of genetic codes. Meanwhile, Fig. 5A shows that the climbability in the case when $N = 50$ is less than that in the case when $N = 10$. The reason for the intuitively unexpected event may be that the walker is likely to get trapped in a terrace corresponding to a local optimum, as the search strategy is close to the exhaustive search (see Fig. 4).

In Fig. 5B, the trial average of the stability for each *Code* is plotted against the average of θ value for Kyte's

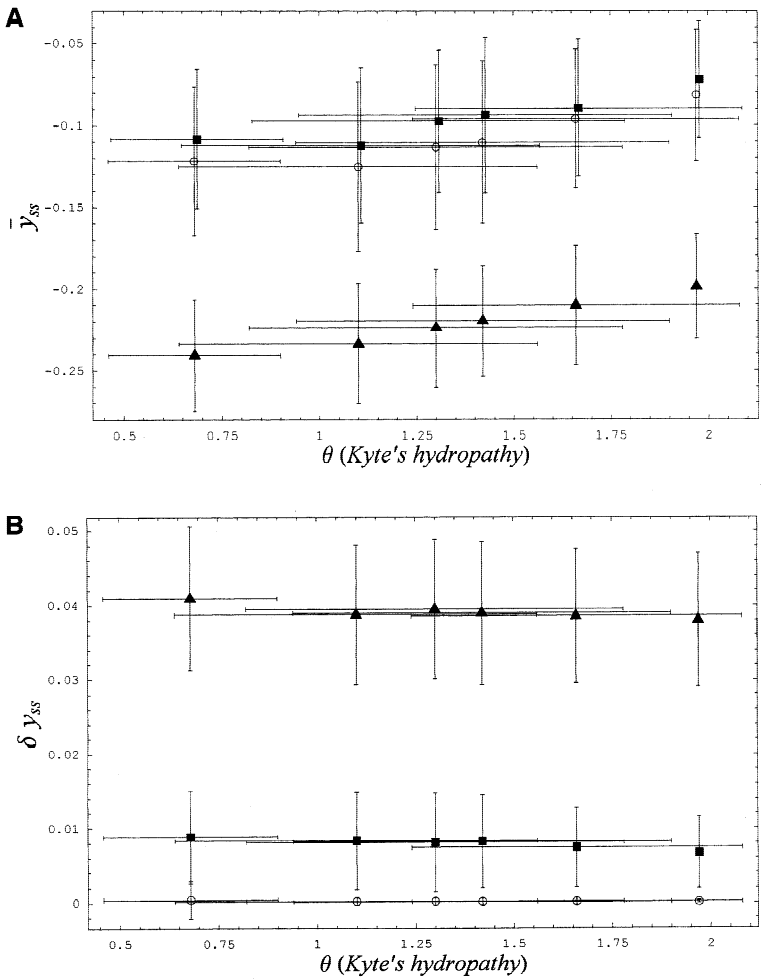


Fig. 5. **A** Correlation between the average of the climbability ($\overline{y_{ss}}$) for each type of genetic code (*Code 0–Code 5*) and the average of θ value for Kyte's hydropathy for the corresponding *Code*. Each error bar represents the standard deviation. The manner of calculating climbability is described in the text. The statistics of θ value for Kyte's hydropathy for each *Code* are taken from Fig. 2A. ○, ■, and ▲ show the case of $N = 50$, $N = 10$, and $N = 3$, respectively. The correlation coefficient is as follows: 0.91 ($p < 0.01$) for ○; 0.91 ($p < 0.01$) for ■; 0.99 ($p < 0.0001$) for ▲. **B** Correlation between the average of the stability (δy_{ss}) for each *Code* and the average of θ value for Kyte's hydropathy for the corresponding *Code*. The manner of calculating the stability is described in the text. The correlation coefficient is as follows: −0.90 ($p < 0.01$) for ○; −0.96 ($p < 0.001$) for ■; −0.89 ($p < 0.01$) for ▲.

hydropathy for the corresponding *Code*. There is a negative correlation between them where the correlation coefficient is less than −0.88 ($p < 0.01$) for each N value. The tendency that the steady state for walkers is more stable with the θ value increasing is conspicuous in the case when $N = 3$, which corresponds to the random sampling search.

Our results demonstrates the θ value of a genetic code affects the climbability and stability in an adaptive walk, and supports the predominance of SGC over almost other conceivable genetic codes.

Discussion

Our approach to the evaluation of genetic codes is based on the structure of amino acid property landscapes in the codon space. Selecting 11 different physicochemical properties for 20 amino acids, we analyzed their amino acid landscapes for the SGC. As a result, we showed these landscapes for several properties, such as hydropathy or polarity, have a globally correlated structure. This is beyond the expectation from the structural regularity due to synonymous codon blocks in the SGC table. The contribution from each letter in a codon is apparently roughly additive to these properties. The results shown in this paper are consistent with the conclusion of several previous studies on mutational robustness in SGC (Haig and Hurst 1991; Freeland and Hurst 1998; Trinquier and Sanejouand 1998). Trinquier and Sanejouand observed that hydrophobicity scales based on spatial environment data of protein residues or on mutation matrices of amino acid replacement generally show stronger conservativeness by the genetic code than those based on pure physicochemical properties of isolated residues. This observation is compatible with ours that Sweet's and Kyte's hydropathy indices showed outstandingly large θ_{SGC} and Z_m values (they ranked first and second, respectively, according to these criteria). It seems that this observation is partly trivial because their frequency data of natural protein residues reflects on amino acid allocations in the SGC.

Our results indicate the simultaneous satisfaction of the mutational robustness in various properties. This fact is not fully understood with another hypothesis that the mechanism of emergence of SGC inevitably brought the mutational robustness. The hypothesis states that the biosynthetically related amino acids that share similar phys-

icochemical properties might be assigned to similar codons by historical constraints. A principle for minimizing the frustration among various properties had to be worked in the evolution of genetic code to improve the mutational robustness. This principle might be kept easier by evolution through introducing a new member (or through improving the specificity of primitive fuzzy allocation) than by evolution through shuffling the existing code table (Maynard-Smith and Szathmary 1995). Evolution of genetic code from a primitive code table of small number of amino acids has been discussed extensively. Eigen created a plausible scenario for the evolution of the genetic code, that is, GNC $\rightarrow$ RNY $\rightarrow$ NNN (Eigen and Schuster 1978). We confirmed that this scenario satisfies the condition of mutational robustness judged from Z score for θ (data not shown).

To link the 11 amino acid property landscapes with a protein fitness landscape on the base sequence space, we used a set of binding preference scores of amino acids as a model set of site fitnesses. Our examination showed that the SGC's mutational robustness for several amino acid properties is actually meaningful for the binding of peptides to MHC class I molecules and that the SGC works well on the high climbability and stability for adaptive walkers on the fitness (= affinity) landscape formed with the set of site fitnesses. The number of sites where site-fitness data were available is too small to generalize our findings. Using data of site fitnesses derived from Houghten et al. (1992; Dooley et al. 1993), we calculated the correlations between amino acid property and site fitness and confirmed the similar positive correlations. We therefore believe our findings may have some generality.

Thus, we gave a clearer meaning to the concept of mutational robustness of the SGC than before. Mutational robustness of the SGC alters the interpretation of the site fitness distributions in individual sites for the protein landscape. If the random mutation takes place in the amino acid sequence space, the wide dynamic range of the site fitness leads to a drastic change of the property of the protein, usually very deleterious for an evolved protein. If the random mutation takes place in the base sequence space, this wide dynamic range of site fitness becomes apparently narrow in point mutagenesis with a small mutation rate due to the correlated mapping of the genetic code (SGC). Note that the situation is quite different in evolutionary molecular engineering using the saturation mutagenesis.

Therefore, even if a protein would have the replication ability, the RNP (= RNA + Protein) world would not have been taken over by the protein world. The RNP world must have more stable evolvability than this imaginary protein world.

Random mutation in the base sequence space is mapped into the correlated mutation in amino acid sequence space through the SGC. Therefore, Mt. Fuji–type landscape of the SGC plays the role of a guardrail set near the sharp ridge of the protein landscape. Moreover, it is speculated on the analogy of the correlation between amino acid property and site fitness that the correlated mapping of the SGC reduces the ruggedness of a protein landscape, that is, the mutational change in nonadditive terms of a protein fitness. Thus, proteins make steadier adaptive walk than in the case without second correlated coding (the first coding was realized by the monomeric sequence versus the polymer function relationship). This is an example of evolution of evolvability of biopolymers.

Acknowledgments. This work was supported by Research Fellowships of the Japan Society for the Promotion of Science for Young Scientists and a Grant-in-Aid from the Ministry of Education, Science, Sports and Culture of Japan. 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). The binding preference scores for MHC class I binding were kindly provided from Dr. Keiko Udaka (Kyoto University) before publication.

Appendix A: Mahalanobis's Generalized Distance

We calculated Mahalanobis's generalized distance D_m to make a simultaneous evaluation of the θ_{SGC} values for 11 properties, for each shuffling manner (m). D_m is defined as

$$D_m = \sqrt{{}^t\mathbf{Z_m}\, R^{-1}\mathbf{Z_m}}$$

where R^{-1} is the inverse matrix of a correlation matrix R, and ${}^t\mathbf{Z_m}$ and $\mathbf{Z_m}$ are the row vector and column vector, respectively:

$$R = \begin{pmatrix} 1 & \rho_{12} & \cdots & \rho_{1K} \\ \rho_{12} & 1 & \cdots & \rho_{2K} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{1K} & \rho_{2K} & \cdots & 1 \end{pmatrix} \text{ and } \mathbf{Z_m} = \begin{pmatrix} Z_m^{(1)} \\ Z_m^{(2)} \\ \vdots \\ Z_m^{(K)} \end{pmatrix}$$

$\rho_{kk'}$ is the correlation coefficient between $\theta_{(m,\infty)\text{-VC}}$ values for the kth property and those for the k'th one (k, $k' = 1, 2, \cdots, K(= 11)$). $Z_m^{(k)}$ is the Z_m for the kth property.

Appendix B: Derivation of Site Fitnesses from Binding Preference Scores

First, we multiplied each of the binding preference scores by ln 10 to deal with them on the free energy scale. Let $score_j(\alpha)$ be the score of an amino acid residue α located at the jth site in a peptide sequence. We here take an assumption that the binding free energy ΔG_P between a particular MHC class I molecule and a peptide with a sequence P is approximately described by

$$\Delta G_{\mathrm{P}} = a \times \sum_{j=1}^{\nu} score_j(\alpha_{\mathrm{P}j}) + b \qquad \text{(Eq. A.1)}$$

where a and b are constants, ν represents the chain length of the peptide, and $\alpha_{\mathrm{P}j}$ represents the particular amino acid residue at the jth site in the peptide sequence P. With $\max(score_j)$, that represents the maximum of $score_j(\alpha)$ over all αs for the jth site, Eq. A.1 is rewritten as

$$\Delta G_{\mathrm{P}} = \sum_{j=1}^{\nu} w_j(\alpha_{\mathrm{P}j}) + a \sum_{j=1}^{\nu} \max(score_j) + b \qquad \text{(Eq. A.2)}$$

where

$$w_j(\alpha) = a(score_j(\alpha) - \max(score_j)) \qquad \text{(Eq. A.3)}$$

Only the first term in Eq. A.2 depends on the sequence. Therefore, we pick out this term and call $W_{\mathrm{P}} = \sum_{j=1}^{\nu} w_j(\alpha_{\mathrm{P}j})$ the "fitness" of the peptide sequence P and call $w_j(\alpha)$ the "site fitness" of an amino acid residue α at the jth site, respectively, in this study. The values of a and original scores are available in Udaka et al. (1999).

References

Aita T, Husimi Y (1996) Fitness spectrum among random mutants on Mt. Fuji-type fitness landscape. J Theor Biol 182:469–485

Aita T, Husimi Y (1998) Adaptive walks by the fittest among finite random mutants on a Mt. Fuji-type fitness landscape. J Theor Biol 193:383–405

Aita T, Husimi Y (2000) Adaptive walks by the fittest among finite random mutants on a Mt. Fuji-type fitness landscape. II: Effect of small non-additivity. J Math Biol (submitted)

Aita T, Uchiyama H, Inaoka T, Nakajima M, Kokubo T, Husimi Y (2000) Analysis of a local fitness landscape with a model of the rough Mt. Fuji-type landscape: application to prolyl endopeptidase and thermolysin. Biopolymers (in press)

Alff-Steinberger C (1969) The genetic code and error transmission. Proc Natl Acad Sci USA 64:584–591

Ardell DH, Sella G (1999) The impact of message mutation on the fitness of a genetic code. *In:* Preliminary Proceedings DIMACS Workshop on Evolution as Computation, pp 175–180

Chothia C (1976) The nature of the accessible and buried surfaces in proteins. J Mol Biol 105:1–14

Chou PY, Fasman GD (1978) Prediction of the secondary structure of proteins from their amino acid sequence. Adv Enzymol 47:45–148

Di Giulio M (1989) The extension reached by the minimization of the polarity distances during the evolution of the genetic code. J Mol Evol 29:288–293

Di Giulio M, Capobianco M, Medugno M (1994) On the optimization of the physicochemical distances between amino acids in the evolution of the genetic code. J Theor Biol 168:43–51

Di Giulio M (1997) On the origin of the genetic code. J Theor Biol 187:573–581

Dooley CT, Chung NN, Schiller PW, Houghten RA (1993) Acetalins: opioid receptor antagonists determined through the use of synthetic peptide combinatorial libraries. Proc Natl Acad Sci USA 90:10811–10815

Eigen M, Schuster P (1978) A hypercycle: Part C: a realistic hypercycle. Naturwiss 65:341–369

Eigen M, McCaskill JS, Schuster P (1989) The molecular quasispecies. Adv Chem Phys 75:149–263

Eisenberg D, Weiss R, Terwilliger T, Wilcox W (1982) Hydrophobic moments in protein structure. Faraday Symp Chem Soc 17:109–120

Freeland S, Hurst L (1998) The genetic code is one in a million. J Mol Evol 47:238–248

Haig D, Hurst L (1991) A quantitative measure of error minimization of the genetic code. J Mol Evol 33:412–417

Houghten RA, Pinilla C, Blondelle SE, Appel JR, Dooley CT, Cuervo JH (1991) Generation and use of synthetic peptide combinatorial libraries for basic research and drug discovery. Nature 354:84–86

Houghten RA, Appel JR, Blondelle SE, Cuervo JH, Dooley CT, Pinilla C (1992) The use of synthetic peptide combinatorial libraries for the identification of bioactive peptides. Bio Techniques 13:412–421

Husimi Y (1989) Selection and evolution of bacteriophage in cellstat. Adv Biophys 25:1–43

Kyte J, Doolittle RF (1982) A simple method for displaying the hydropathic character of a protein. J Mol Biol 157:105–132

Maeshiro T, Kimura M (1998) The role of robustness and changeability on the origin and evolution of genetic codes. Proc Natl Acad Sci USA 95:5088–5093

Maynard-Smith J, Szathmary E (1995) The major transitions in evolution. W. H. Freeman/Spektrum Akad. Verlag

Privalov PL, Tiktopulo EI, Venyaminov SY, Griko YV, Makhatadze GI, Khechinashvili NN (1989) Heat capacity and conformation of proteins in the denatured state. J Mol Biol 205:737–750

Rechenberg I (1984) The evolution strategy: a mathematical model of Darwinian evolution. In: Frehland E (ed.) Synergetics: from microscopic to macroscopic order. Springer Series in Synergetics, vol. 22, pp 122–132

Sweet RM, Eisenberg D (1983) Correlation of sequence hydrophobicities measures similarity in three-dimensional protein structure. J Mol Biol 171:479–488

Trinquier G, Sanejouand Y (1998) Which effective property of amino acids is best preserved by the genetic code? Protein Eng 11:153–169

Udaka K, Wiesmüller KH, Kienle S, Jung G, Walden P (1995) Tolerance to amino acid variations in peptides binding to the major histocompatibility complex class I protein H-2K^{b}. J Biol Chem 270:24130–24134

Udaka K, Wiesmüller KH, Jung G (1999) Repertoire forecast of MHC class I binding peptides with peptide libraries. *In:* Masanori Kasahara (ed) Proceedings of the Sixth International Workshop on MHC Evolution, 1999. Springer-Verlag (in press)

Wells JA (1990) Additivity of mutational effects in proteins. Biochem 29:8509–8517

Woese CR, Dugre DH, Dugre SA, Kondo M, Saxinger WC (1966) On the fundamental nature and evolution of the genetic code. Cold Spring Harbor Symp. Quant Biol 31:723–736

Wong JT (1980) Role of minimization of chemical distances between amino acids in the evolution of the genetic code. Proc Natl Acad Sci USA 77:1083–1086

Zamyatnin AA (1975) Protein volume in solution. Prog Biophys Mol Biol 24:109–123