Available online at www.sciencedirect.com
ScienceDirect
Journal of Theoretical Biology 245 (2007) 715–725
Journal of Theoretical Biology
www.elsevier.com/locate/yjtbi

Solvent accessibility, residue charge and residue volume, the three ingredients of a robust amino acid substitution matrix

Hani Goodarzi[a,*], Ali Katanforoush[b,c], Noorossadat Torabi[d], Hamed Shateri Najafabadi[d]

[a]*Molecular Biology Department, Princeton University, Princeton, NJ, USA*
[b]*Department of Bioinformatics, Institute of Biochemistry and Biophysics, University of Tehran, Tehran, Iran*
[c]*Institute for Studies in Theoretical Physics and Mathematics, Niavaran Square, Tehran, Iran*
[d]*Department of Biotechnology, Faculty of Science, University of Tehran, Tehran, Iran*

Received 27 July 2006; received in revised form 31 October 2006; accepted 8 December 2006
Available online 19 December 2006

Abstract

Cost measure matrices or different amino acid indices have been widely used for studies in many fields of biology. One major criticism of these studies might be based on the unavailability of an unbiased and yet effective amino acid substitution matrix. Throughout this study we have devised a cost measure matrix based on the solvent accessibility, residue charge, and residue volume indices. Performed analyses on this novel substitution matrix (i.e. solvent accessibility charge volume (SCV) matrix) support the uncontaminated nature of this matrix regarding the genetic code. Although highly similar to a number of previously available cost measure matrices, the SCV matrix results in a more significant optimality in the error-buffering capacity of the genetic code when compared to many other amino acid substitution matrices. Besides, a method to compare an SCV-based scoring matrix with a number of widely used matrices has been devised, the results of which highlights the robustness of this matrix in protein family discrimination.

Keywords: Load minimization; Cost measure matrix; Optimality; Amino acid substitution matrix; Genetic code; Scoring matrix

1. Introduction

1.1. The canonical genetic code: theories and concepts

The codon assignments in the canonical genetic code play an undeniable role in connecting the genomic world where the mutations take place to the proteomic world where natural selection is capable of affecting survival of the corresponding organisms (Fitch and Upper, 1987; Zhu and Freeland, 2006). Yet, the nature of this role and the notion of the very evolution of the code itself are highly debated among the researchers (reviewed in Freeland et al., 2003; Di Giulio, 2005). As Di Giulio (2005) puts it, two deterministic forces might be speculated in the evolution of the genetic code:

1. *Physicochemical determinism*: Natural selection had produced a redundant code that buffered phenotype from genetic errors by ensuring a high proportion of silent point mutations (Sonneborn, 1965; Zuckerkandl and Pauling, 1965) and a less drastic effect even when the mutation does occur (Alff-Steinberger, 1969; Goldberg and Wittes, 1966). Quantitative support for this hypothesis is developed by the use of computer simulations in many studies (Haig and Hurst, 1991; Freeland and Hurst, 1998; Gilis et al., 2001; Goodarzi et al., 2004; Goodarzi et al., 2005a–c).
2. *Historical determinism*: The structure of the genetic code reflects the biosynthetic pathway of amino acid formation. This idea presented by Wong (1975) is studied intensely by Di Giulio (1997a, b, 1999, 2000).

Abbreviations: SCV, solvent accessibility charge volume; CRM, code-reflecting matrix

*Corresponding author. Tel.: +1 609 258 3658.
E-mail address: goodarzi@princeton.edu (H. Goodarzi).

0022-5193/$ - see front matter © 2007 Elsevier Ltd. All rights reserved.
doi:10.1016/j.jtbi.2006.12.014

The search for the major selection pressure for the evolution of the genetic code seems far from over and the debates seem unresolved; yet, many clues are added to our knowledge of the code since the introduction of these theories (Ellington et al., 2000; Di Giulio, 2001, 2005; Goodarzi et al., 2005b, c; Zhu and Freeland, 2006).

1.2. The case for an error minimizing code

In spite of a sophisticated growth in analytical evidence to support the case for an "error minimization" hypothesis, many facets of this theory remain ill-explored, including the mechanism and pathway by which an adaptive pattern of codon assignments emerged, the extent to which natural selection created synonym redundancy, its role in shaping the amino acid and nucleotide languages, and even the correct interpretation of the adaptive codon assignment pattern (Freeland et al., 2003).

Many recent studies on this subject include computer simulations in quantifying the extent to which the canonical genetic code is capable of minimizing the effects of mistranslations, point mutations, and more recently shown, ins/del mutations (Goodarzi et al., 2005c). These analyses are based on two major constituents:

1. *Fitness functions*: Created to assign a score to every given code based on its load minimizing ability, these functions are used extensively in many analytical studies (Haig and Hurst, 1991; Freeland and Hurst, 1998; Gilis et al., 2001; Freeland, 2002; Archetti, 2004; Goodarzi et al., 2004, 2005a–c). In this study, one of the recently proposed fitness functions is used:

$$\varphi^{faa} = \sum_{c=1}^{64} \frac{p[a(c)]}{n[a(c)]} \sum_{c'=1}^{64} p(c'|c) \cdot g[a(c), a(c')] \quad \text{(Giles, 2001)}, \tag{1}$$

   where $p[a(c)]$ returns the relative frequency of the amino acid $a(c)$ coded by codon c, $n[a(c)]$ is an integer standing for the number of synonymous codons of amino acid $a(c)$, and $g[a(c),a(c')]$ is a cost measure function which illustrates the deleterious effect of the amino acid substitution resulted from the misinterpretation of codon c as c'. $p(c'|c)$ is the probability of codon c being misinterpreted as codon c' (the values chosen by Freeland and Hurst, 1998).
2. *Random code generation*: Fitness scores are calculated for randomly generated codes; by comparing the results to that of the canonical one, a quantitative measure of the optimality of the code is obtained. Haig and Hurst (1991) chose the following rules for generating random codes: the codon space (i.e., the possible 64 codons) is divided into the same 21 non-overlapping sets of codons observed in the standard code, each set comprising all codons specifying a particular amino acid in the standard code; the three stop codons remain in the same position of the standard code for all alternative codes, while each of the 20 amino acids is assigned randomly to one of these sets to form an alternative code.

Each of these basic ideas has faced its own critiques. For example, Di Giulio (2000) claimed that although the frequency of the codes that perform better than the canonical one is roughly 10^6 (or 10^9 in other studies), considering the vast number of possible codes (20! in case of the classic method of random code generation) results in a very high number of better alternative codes. Besides, many of the random codes differ drastically from the canonical genetic code or its presumed ancestors and are unlikely to be obtained by small mutations. These arguments have resulted in an alternative method of generating random codes (Archetti, 2004).

Another debated parameter in the abovementioned studies is the function $g[a(c),a(c')]$. Many amino acid substitution matrices such as PAM_{74-100} are used to serve as cost measure functions (Table 1); yet, as Di Giulio (2001) has shown the amino acid substitution matrices that are based on the observations on contemporary and real proteins are flawed by the contamination of the genetic code itself and are useless in such analyses. It is difficult to untangle contamination with the genetic code from significant associations between measures of mutation cost and the genetic code, in part because almost all matrices include subtle biases (for example, the set of proteins chosen for crystallographic work is decidedly non-random and biased towards members of large multigene families involved in core metabolic and regulatory processes). Thus, the main source of information for measuring the cost of mutations should be obtained from the amino acid indices that include only the physicochemical characteristic of each of the amino acids. Many amino acid indices such as Polar requirement (Woese et al., 1966), Hydrophobicity scales (Engelman et al., 1986; Nozaki and Tanford, 1971), and Hydropathic character (Kyte and Doolittle, 1982) are used in different studies to measure optimality of the genetic code. Yet, each of these indices covers a limited portion of the physicochemical properties of each amino acid as a whole. In this study, we have tried to devise an amino acid substitution matrix, which includes the polarity, the charge, and the volume of each residue to measure the cost of each mutation. This matrix, designated the solvent accessibility charge volume (SCV) matrix, was shown to be unbiased towards the genetic code (i.e. not contaminated by the genetic code). Moreover, the optimality measurement analyses using this matrix further reveal the robustness of the genetic code in minimizing the effects of mistranslations. Applying this novel matrix to untangle protein classification problems reveals that this matrix is as efficient as database-driven matrices such as PAM50, which further highlights the credibility of this matrix.

Table 1
A number of common cost measure matrices and indices previously used in prior studies

Cost measure	Introduced in	Used in	Corresponding cost function
PAM_{74-100}	Benner et al. (1994)	Freeland et al. (2000), Gilis et al. (2001), Goodarzi et al. (2004)	$g(a_1,a_2) = -h(a_1,a_2)$
Mutation	Gilis et al. (2001)	Gilis et al. (2001), Goodarzi et al. (2004)	$g(a_1,a_2) = -h(a_1,a_2)$
Polar requirement	Woese et al. (1966)	Haig and Hurst (1991), Freeland and Hurst (1998), Gilis et al. (2001)	$g(a_1,a_2) = \lvert h(a_1)-h(a_2)\rvert$
Hydrophobicity scale #1	Engelman et al. (1986)	Zhu et al. (2003)	$g(a_1,a_2) = \lvert h(a_1)-h(a_2)\rvert$
Hydrophobicity scale #2	Nozaki and Tanford (1971)	Zhu et al. (2003)	$g(a_1,a_2) = \lvert h(a_1)-h(a_2)\rvert$
Hydropathic character	Kyte and Doolittle (1982)	Zhu et al. (2003)	$g(a_1,a_2) = \lvert h(a_1)-h(a_2)\rvert$

2. Materials and methods

2.1. The amino acid properties used to construct the SCV matrix

Initially, three amino acid properties were used to construct the SCV matrix: the percentage of exposure for each amino acid (Chen et al., 2004), residue volume (Zamyatin, 1972), and residue charge. The SCV matrix is devised based on a linear relationship among these properties that would define each amino acid:

$$g(a,a') = \left|S(a) - S(a')\right| + \alpha \cdot \left|V(a) - V(a')\right| + \beta \cdot \left|C(a) - C(a')\right|, \quad (2)$$

where S is the solvent accessibility, C is the residue charge, and V is the residue volume of each amino acid (all these parameters are normalized to a range of 0–1). α and β, the values of which are chosen later, are linear constants combining these amino acid properties to form a proper cost measure matrix. This equation is similar to the one introduced by Grantham (1974) based on amino acid composition, molecular volume, and polarity; yet, our analyses showed that SCV is a more reliable matrix in terms of protein family discrimination through pair-wise alignments (data not shown).

2.2. Rules for generating random codes

In this study, two methods have been used for generating random codes: (i) the degenerate method that theoretically is capable of producing any possible code and (ii) the constrained classic method, which is a limited permutation of amino acids in their fixed positions.

2.2.1. The degenerate method

When using this method, we are not interested in studying the canonical genetic code and its optimality; yet, we are trying to scan all the possible alternative codes that might have been proved to be more efficient than the canonical one in load minimization. Generating random codes is not simply assigning each codon randomly to a chosen amino acid due to the fact that the degeneracy of the code would not be preserved. To this end, the degenerate method of random code generation introduced by Goodarzi et al. (2005c) has been used (for details see the original paper). This method of generating random codes assigns a random number of synonymous codons as well as random codons to each amino acid while preserving the degeneracy of the generated codes (for details see Goodarzi et al., 2005c).

2.2.2. The classic method

The following method (as in Archetti, 2004) has been used to create random codes:

1. The "codon space" is divided into 21 non-overlapping sets of codons observed in the canonical code, each set specifying an amino acid in the natural genetic code (one set consists of stop codons).
2. Each alternative code is obtained by randomly assigning each of the 20 amino acids to one of these sets. All three stop codons remain invariant, in position for all alternative codes.
3. The first or the second bases are always kept invariant in order to limit the space of possible variant codes.

This method creates random codes that are positioned relatively near the canonical genetic code in the space of all possible codes and might have been accessed by evolution in terms of small gradual mutations. In the classic method, the codon boxes are fixed and the amino acids are permutated in these fixed positions; yet, in the degenerate method the codon boxes are variable as well and the number of synonymous codons in each box is subject to change Goodarzi et al., 2005c).

In this study, based on two reasons we have used z-scores instead of the empirical p-values: (i) the distribution of fitness scores are significantly normal (data not shown), and (ii) some of the searches returned 0 better codes that made such studies incomparable if p-values were used. The definition of z-score is as follows:

$$z = \frac{\varphi_{gc} - \text{mean}}{\text{sde}}, \quad (3)$$

where φ_{gc} is the fitness score of the canonical genetic code and mean and standard deviation values are determined by generating a defined number of random codes (10^5–10^9 in this study).

2.3. Code-reflecting matrices

After choosing proper values as α and β in Eq. (2), the SCV matrix would be complete; yet, the possibility of the contamination with the properties of the canonical genetic code should be addressed for this matrix. In order to study the proposition that this matrix is or is not biased towards the genetic code, a regression analysis has been used to compare this matrix to the matrices that reflect the structure of the canonical genetic code (as in Di Giulio, 2001).

In this part of our study, two substitution matrices were used that are deemed to be obtained from the structure of the canonical genetic code itself:

1. *Angle measurement*: Introduced in 1989 by Di Giulio, this matrix measures the angle between the 20 amino acid vectors in the 21-dimension space of the canonical genetic code. As reported by Di Giulio (1989) himself, this matrix highlights the role of polar character and size of amino acids in the evolutionary history of the genetic code.
2. *CRM (code-reflecting matrix)*: In addition to angle measurement, we have also devised CRM based on the structure of the canonical genetic code. This matrix is obtained from the probabilities of mistranslations between the amino acids (applying the weightings and biases theoretically chosen by Freeland and Hurst (1998) for mistranslations):

$$CRM(a_i, a_j) = \sum_{c(a_i)} \sum_{c(a_j)} p(c'|c). \tag{4}$$

In case $CRM(a_i, a_j) = 0$:

$$CRM(a_i, a_j) = -\mathrm{Max}_{value}(a_i, a_k), \tag{5}$$

where $\mathrm{Max}_{value}(a, b)$ returns the greatest of a and b.

The results of the Mantel tests between these two matrices and the SCV matrix are compared with those of other cost measure matrices, both contaminated and uncontaminated, in order to comment on the genuineness of the SCV matrix. All these regressions were performed between matrices of 190 points, i.e. excluding the diagonal elements of these matrices. In the matrices of the polarity distances of amino acids, every component represents the absolute value of the difference between the value of the polarity of the ith amino acid and that of the jth one.

2.4. SCV matrix vs. other amino acid indices

We also made a comparison between the SCV matrix and other amino acid cost measure indices available. To this end, 516 amino acid indices were downloaded (available online at www.genome.ad.jp/dbget/aaindex.html) and were compared to the SCV matrix. Regression analysis between matrices of 190 points of these amino acids indices and the SCV matrix was performed and the eight indices that showed an R-value of greater than 0.75 were identified.

In the next step, z-scores were calculated for each of these 516 amino acid indices as cost measure functions by generating 10^7 randomly generated codes (the classic method):

$$g(a, a') = |h(a) - h(a')|, \tag{6}$$

where $h(a)$ returns the corresponding index of amino acid a.

The computed z-scores were then compared to the z-score of the SCV matrix to study the load minimization capacity of the canonical genetic code and the robustness of the SCV matrix to highlight such optimality.

2.5. The ability of the SCV matrix in protein alignment

As described above, the SCV matrix is derived as a linear function of three amino acid properties with chosen parameters; yet, it should be proven that this matrix actually reflects the distances between amino acids. To this end, we tried to compare the SCV matrix to other amino acid substitution matrices based on scoring matrices comparison methods which test the capability of scoring matrices to discriminate protein families through alignment (Henikoff and Henikoff, 1993). When using a scoring matrix based on the distance values (e.g. SCV), one should consider some necessities on the scoring matrices used in sequence alignment algorithms. For measuring the similarity between two sequences using alignment algorithms, the scoring matrix has to consist of a log-odd discrimination measure of amino acids; otherwise, the score of the "best" alignment between two sequences could not bear any meaningful concept about their "similarity". Another essential property of a scoring matrix used in local alignment algorithms is that the expected value of the scores obtained from aligning every two "random" subsequences should be less than zero (see Altschul, 1991). By applying a negative exponential function on values of SCV matrix, they are transformed to a form of two-variable distribution function:

$$f(a, a') = \mathrm{e}^{-\lambda . g(a,a')}, \tag{7}$$

where $g(a, a')$ denotes the SCV measure between amino acids a and a' Eq. (2). λ is a constant, set to 12 for SCV and 3 for Polar Requirement, chosen to satisfy the constraints of a scoring matrix suitable for local alignments.

Log-odds are calculated from this distribution as "expected", and the relative frequency of individual amino acids in Swissprot database (http://tw.expasy.ch/sprot/relnotes/relstat.html) as "random" models, respectively:

$$sc(a, a') = \log_2 \left(\frac{f(a, a')}{\rho . q_a . q_{a'}} \right), \tag{8}$$

where q_a is the relative frequency of amino acid a and ρ is calculated from

$$\rho = \frac{\sum_{a,a'} f(a,a')}{20\sum_a q_a}. \quad (9)$$

Briefly, our goal has been to apply ρ as a *shift value* in order to make a negative expected score in alignments and subsequently choose λ such that $f(a,a')$ satisfies the probability density function as much as possible. Note that, in the last step all the elements are rounded to the nearest integer.

In order to compare the efficiency of SCV as a substitution-scoring matrix against other conventional scoring matrices regarding protein alignment, we have studied the ability of discrimination between protein families by utilizing local sequence alignment algorithm on some pairs of "representative" sequences of families. The idea of this simple analysis is the same as Henikoff and Henikoff (1993) with the exception of some slight differences in choosing the protein family database.

It implies the three following steps:

1. Choosing the representative members for each family.
2. Selecting the "hard-to-discriminate" families.
3. Measuring the power of the scoring matrix to discriminate between these families.

For the first step, we imagine that members of each family share a phylogenetic tree. By the assumption that the distance between each pair on the tree is proportional to the inverse of the alignment score on them, one can find out the furthest two members of a family by a conventional algorithm in graph theory used to find the diameter of a tree. The algorithm would compute the greatest distance in the family provided that the distance function obeys a line distance condition, i.e. if x is between y and z then $dist(x,y)+dist(x,z)$ equals to $dist(y,z)$. Although this property is not necessarily true on the score of local sequence alignment as a distance, we accept it as a heuristic approach. Besides these two members, we chose another protein as the "center" of the family to point to the sequence that has the most likely alignment with both "furthest" members. In other words, the alignment score of the "center" and one of "furthest" members is as near as possible to the alignment score of the "center" and the other one. By adding the longest sequence of the family up to the sequences described above, we introduce four representatives of each family. At this step, we used the 6/−1 identity matrix for alignments.

Among these protein families, those that their representatives appear to be more similar to other families (determined by the corresponding alignments) are called "hard-to-discriminate". As discussed in Bastien et al. (2004) considering z-score rather than conventional E-value for measuring protein relationships could lead to a more reliable conclusion when statistical requirements to an E-value analysis are not clearly given. So to determine such families, all pairwise alignments between families' representatives were computed. Then 120 most similar sequences with this representative were screened for each family. z-Score of the family's representative out of these 120 samples of alignment scores approximately makes a sense on "hard-to-discrimination" wherever the z-score is small (for instance less than 10).

We applied all the pairwise alignments between each family's representative and all other proteins in the same family. Consider a representative of a family, for instance "center" protein and do its pairwise alignment with all other sequences of the family. Now the difference of the highest and lowest alignment z-score could estimate a measure for the *firmness* of the family. Formerly by computing all pairwise alignments scores between the family's representative and representatives of other families, a set of random samples that is necessary for calculating z-score, was provided. Scoring matrices to be comprised are BLOSUM65, BLOSUM50, PAM250, PAM50, SCV and Polar requirement derived scoring matrices. The 6/−1 identity matrix also was applied as an inefficient-assumed testifier.

For this part of the study, Pfam-A (ver. 18 release 23 July 2005) (Finn et al., 2006), was used for extracting proteins the sequences of which are later provided by Swissprot (release 48.5 September 2005) (Bairoch et al., 2004). Consequently, 33 813 proteins were obtained; each of which belongs to one or more of 7973 families.

3. Results

3.1. Choosing appropriate values for α and β

Generating 10^5 random codes (the degenerate method), z-scores were calculated for φ^{faa} applying SCV matrix (constructed by incremental values of α and β) as the cost measure function. The results are tabulated in Table 2, where a z-score is reported for every α and β. The value −7.111 was apparently the minimal z-score and the corresponding $\alpha = 0.5$ and $\beta = 0.188$ values of this z-score were chosen to construct the final SCV matrix.

Applying the abovementioned α and β values in Eq. (2), the SCV matrix was constructed which is basically a

Table 2
z-Values calculated for incremental values of α and β by generating 10^7 randomly generated codes (degenerate method)

α	β				
	0	0.25	0.5	0.75	1
0	−6.72744	−6.98423	−7.08071	−7.06041	−6.966
0.063	−6.74819	−7.00255	−7.10084	−7.08434	−6.99402
0.125	−6.75709	−7.00963	−7.11062	−7.09898	−7.01383
0.188	−6.75561	−7.00675	−7.11124	−7.10545	−7.02654
0.25	−6.74496	−6.99508	−7.10355	−7.10423	−7.03226
0.313	−6.72597	−6.97528	−7.08818	−7.09597	−7.03169

Table 3
The SCV amino acid substitution matrix

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
A	0.00																			
R	0.72	0.00																		
N	0.53	0.19	0.00																	
D	0.74	0.21	0.21	0.00																
C	0.28	0.87	0.68	0.87	0.00															
Q	0.84	0.34	0.31	0.35	0.82	0.00														
E	0.95	0.42	0.43	0.21	0.94	0.32	0.00													
G	0.42	0.60	0.41	0.60	0.41	0.41	0.57	0.00												
H	0.67	0.52	0.51	0.70	0.66	0.36	0.68	0.24	0.00											
I	0.27	0.98	0.79	1.01	0.28	1.10	1.22	0.69	0.93	0.00										
L	0.32	0.90	0.71	0.90	0.11	0.94	1.06	0.53	0.77	0.16	0.00									
K	0.98	0.26	0.45	0.42	0.99	0.47	0.34	0.72	0.64	1.24	1.08	0.00								
M	0.32	0.90	0.71	0.90	0.10	0.73	0.87	0.32	0.56	0.37	0.21	1.02	0.00							
F	0.27	0.86	0.67	0.89	0.16	0.98	1.10	0.57	0.81	0.12	0.05	1.12	0.25	0.00						
P	0.61	0.44	0.25	0.44	0.60	0.23	0.41	0.19	0.27	0.88	0.71	0.56	0.50	0.76	0.00					
S	0.45	0.44	0.25	0.44	0.44	0.39	0.50	0.16	0.27	0.72	0.55	0.56	0.46	0.60	0.16	0.00				
T	0.50	0.56	0.37	0.56	0.49	0.33	0.53	0.08	0.16	0.77	0.61	0.68	0.40	0.65	0.12	0.12	0.00			
W	0.25	0.97	0.78	1.00	0.27	1.09	1.21	0.68	0.92	0.22	0.29	1.23	0.36	0.25	0.87	0.70	0.76	0.00		
Y	0.06	0.68	0.49	0.71	0.22	0.80	0.92	0.39	0.63	0.30	0.26	0.94	0.26	0.21	0.58	0.41	0.47	0.29	0.00	
V	0.59	1.17	0.98	1.17	0.30	0.82	1.14	0.57	0.65	0.45	0.29	1.29	0.26	0.33	0.73	0.73	0.60	0.56	0.52	0.00

distance matrix. As presented in Table 3, the SCV matrix is a symmetrical one where the cost of substitution for each amino acid pair has a positive value.

3.2. Analyzing the correlation between the SCV matrix and the CRMs

As previously mentioned, two code-reflecting matrices, namely the angle measurement and the CRM, were used to study the possible contamination of the SCV matrix with the structure of the canonical genetic code. The correlation coefficients for each of the pairwise regressions between each of the cost measure matrices and the two code-reflecting matrices are tabulated in Table 4. The *p*-values obtained from Mantel test are also provided. The Mantel test was performed by zt software (Bonnet and Van de Peer, 2002). The R^2-values are calculated for the SCV matrix, a number of polarity measurements (Polar requirement [Woese et al., 1966], Hydrophobicity scale #1 [Nozaki and Tanford, 1971], Hydrophobicity scale #2 [Engelman et al., 1986], and Hydropathic character [Kyte and Doolittle, 1982]), and PAM_{74-100} that is previously discussed to be biased towards the canonical genetic code (Di Giulio, 2001). With the exception of Hydrophobicity scale #1, the reported *p*-values (<0.05) corroborate the reliability of the obtained R^2-values.

Table 4
Pairwise regression analyses driven to study the possible bias of the SCV matrix towards the canonical genetic code

Cost measure matrix	Angle measurement	Code-reflecting matrix
SCV matrix	0.0573 (5×10^{-4})	0.0215 (2×10^{-2})
Polar requirement	0.0884 (1×10^{-4})	0.0442 (3×10^{-3})
Hydropathic char.	0.0475 (1×10^{-3})	0.0263 (2×10^{-2})
PAM_{74-100}	0.2265 (1×10^{-4})	0.1217 (1×10^{-4})
Mutation matrix	0.0313 (8×10^{-3})	0.0529 (1×10^{-3})
Hydrophobicity scale 1	0.0164 (4×10^{-2})	0.0006 (3×10^{-1})
Hydrophobicity scale 2	0.0770 (3×10^{-4})	0.0702 (4×10^{-4})

The *p*-values computed from Mantel test are provided in parentheses. The corresponding *R*-values of the SCV matrix are comparable to other unbiased matrices, suggesting the uncontaminated nature of this matrix.

3.3. Correlating the SCV matrix with various amino acid indices

Fig. 1, indicates the correlation coefficients (*R*) obtained from correlating the SCV matrix with 516 other cost measure matrices derived from 516 amino acid indices. *R*-Values are ranged from −0.2 to +0.8 and 62 percent of

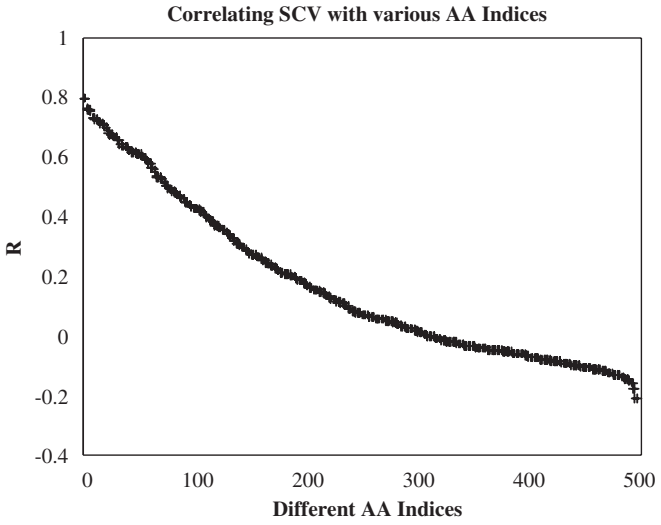

Fig. 1. Correlation coefficients calculated from comparing the SCV matrix with 516 amino acid matrices.

these indices show positive correlation with the SCV matrix (38 percent have an R-value greater than $+0.2$). Those indices that show more than 0.75 correlation with the SCV matrix were identified and tabulated in Table 5.

In the next step, the SCV matrix was compared to these 516 amino acid indices regarding the error-buffering capacity of the canonical genetic code. The calculated z-scores for each of these indices (generating 10^8 random codes from the classic method) are presented as a histogram in Fig. 2 where the corresponding z-score for the SCV matrix is indicated by an arrow (z-score $= -2.130$). As it can be inferred from the data, none of these indices acted as efficiently as the SCV matrix in revealing the load minimization characteristic of the canonical genetic code.

3.4. Results from alignment

By choosing proper values for the parameter of negative exponential function and some other parameters to scale and shift values of the scoring matrix we derived a scoring matrix from SCV matrix satisfies the requirements for local sequence alignments mentioned in Section 2.5 (Tables 6 and 7). Among the 7973 protein families, 328 were screened as "hard-to-discriminate".

Pair-wise sequence alignments were performed between the families' representatives and the corresponding z-scores were calculated for each family. We did this process seven by four times, each time with a different representative or a different scoring matrix. The number of families among these 328 with lowest z-score of greater than 4 is summarized in Table 8. The difference between the highest z-score and the lowest z-score obtained for each family is also presented as another measure for protein family firmness which is also informative about the efficiency of scoring matrices that are shown in Table 9. Based on the data tabulated in these two tables, the "center" representative poses a better delegation for protein families, which is independent of the applied substitution matrix.

These comparisons that are based on previously introduced methods for comparing the scoring matrices (see Henikoff and Henikoff, 1993), highlight the ability of SCV-derived scoring matrix in protein family discrimination which is comparable to that of PAM50 (Table 8 and Table 9). In detail, we ranked the ability of the scoring matrices in discriminating protein families based the results depicted in Tables 8 and 9: PAM250, BLOSUM50, BLOUM65, SCV-derived, PAM50, POLAR-req, 6/−1 Identity; where the leftmost shows the best ability on protein alignment and protein family discrimination in our experience. In terms of alignment, the SCV matrix contains a considerable level of information.

4. Discussion and conclusion

4.1. Solvent accessibility

Table 10 depicts the percentage of exposure of each amino acid (Chen et al., 2004) along with other amino acid indices, namely Polar requirement (Woese et al., 1966), Hydrophobicity scale of Nozaki (Nozaki and Tanford, 1971), Hydrophobicity scale of Engelman (Engelman et al., 1986), and Hydropathic character (Kyte and Doolittle, 1982). We have chosen the solvent accessibility of amino acids to derive a cost measure matrix due to the fact that this property models the impacts of mutations on the structure of proteins more effectively. Solvent accessibility index very much resembles the measurements of hydrophobocity and polarity (considering the correlation coefficients in

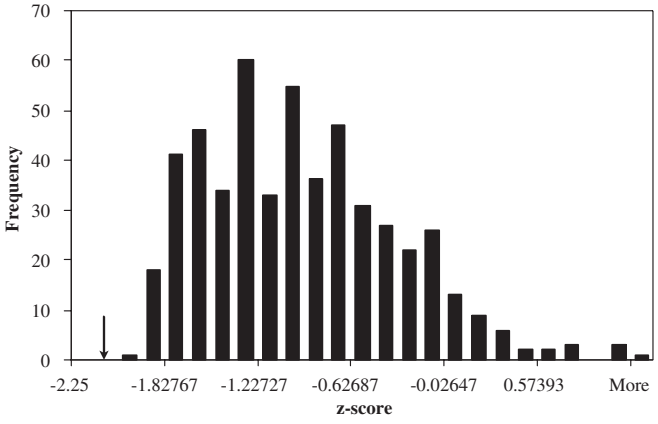

Fig. 2. z-scores calculated for each of the 516 amino acid indices as well as the SCV matrix (indicated by an arrow) generating 10^8 random codes (classic method).

Table 5
The top eight amino acid indices that show more than 0.75 correlation with the SCV matrix accompanied by their corresponding references

Amino acid index	Reference
Hydropathy scale based on self-information values in the two-state model (9%)	Naderi-Manesh et al. (2001)
Information value for accessibility; average fraction (23%)	Biou et al. (1988)
Polarity	Grantham (1974)
Hydropathy scale based on self-information values in the two-state model (16%)	Naderi-Manesh et al. (2001)
Information value for accessibility; average fraction (35%)	Biou et al. (1988)
Hydrophobic parameter pi	Fauchere and Pliska (1983)
Mean fractional area loss	Rose et al. (1985)
14 A contact number	Nishikawa-Ooi (1986)

Table 6
SCV-derived scoring matrix for alignment

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V	B	Z	X
A	7	−5	−1	−5	4	−7	−9	0	−3	3	1	−10	3	3	−3	−1	−1	5	7	−3	−2	−8	−5
R	−5	8	5	4	−5	2	0	−3	0	−9	−8	3	−6	−7	1	0	−2	−7	−3	−13	5	1	−5
N	−1	5	9	5	−2	3	1	1	1	−6	−5	0	−3	−3	4	4	2	−3	1	−9	8	2	−6
D	−5	4	5	8	−5	2	4	−3	−3	−10	−9	0	−6	−7	1	0	−2	−7	−3	−13	7	4	−5
C	4	−5	−2	−5	12	−4	−7	2	0	5	7	−7	9	7	0	2	1	7	7	4	−3	−5	−7
Q	−7	2	3	2	−4	9	3	1	3	−11	−9	0	−3	−8	5	1	3	−8	−5	−6	3	7	−5
E	−9	0	1	4	−7	3	7	−3	−3	−14	−11	2	−6	−11	1	−1	−2	−11	−7	−12	3	7	−5
G	0	−3	1	−3	2	1	−3	7	5	−5	−2	−5	3	−2	4	4	6	−2	2	−3	0	0	−5
H	−3	0	1	−3	0	3	−3	5	10	−7	−5	−2	1	−5	5	4	6	−5	−1	−2	0	2	−6
I	3	−9	−6	−10	5	−11	−14	−5	−7	8	4	−14	2	6	−7	−5	−6	6	3	0	−7	−12	−5
L	1	−8	−5	−9	7	−9	−11	−2	−5	4	6	−12	5	7	−5	−3	−3	4	3	2	−6	−10	−5
K	−10	3	0	0	−7	0	2	−5	−2	−14	−12	8	−9	−11	−2	−2	−4	−11	−8	−15	0	1	−5
M	3	−6	−3	−6	9	−3	−6	3	1	2	5	−9	10	5	1	1	2	5	5	4	−4	−4	−7
F	3	−7	−3	−7	7	−8	−11	−2	−5	6	7	−11	5	9	−5	−2	−3	6	6	2	−4	−9	−6
P	−3	1	4	1	0	5	1	4	5	−7	−5	−2	1	−5	8	5	6	−5	−1	−5	3	3	−6
S	−1	0	4	0	2	1	−1	4	4	−5	−3	−2	1	−2	5	7	5	−2	1	−5	3	0	−5
T	−1	−2	2	−2	1	3	−2	6	6	−6	−3	−4	2	−3	6	5	8	−3	1	−3	1	1	−6
W	5	−7	−3	−7	7	−8	−11	−2	−5	6	4	−11	5	6	−5	−2	−3	12	6	0	−4	−9	−7
Y	7	−3	1	−3	7	−5	−7	2	−1	3	3	−8	5	6	−1	1	1	6	10	−1	−1	−6	−6
V	−3	−13	−9	−13	4	−6	−12	−3	−2	0	2	−15	4	2	−5	−5	−3	0	−1	7	−10	−8	−5
B	−2	5	8	7	−3	3	3	0	0	−7	−6	0	−4	−4	3	3	1	−4	−1	−10	7	3	−5
Z	−8	1	2	4	−5	7	7	0	2	−12	−10	1	−4	−9	3	0	1	−9	−6	−8	3	7	−5
X	−5	−5	−6	−5	−7	−5	−5	−5	−6	−5	−5	−5	−7	−6	−6	−5	−6	−7	−6	−5	−5	−5	−6

Table 7
Scoring matrix for alignment based on polar requirement

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V	B	Z	X
A	7	−2	−5	−19	0	−16	−17	3	3	−2	−3	−6	1	−1	6	5	6	2	1	1	−6	−16	−5
R	−2	8	4	−9	−9	−6	−7	2	6	−10	−11	3	−7	−9	−3	1	−3	−7	−7	−8	3	−7	−5
N	−5	4	9	−5	−12	−2	−3	−1	3	−14	−15	8	−11	−13	−6	−3	−6	−10	−11	−11	7	−3	−6
D	−19	−9	−5	8	−26	6	5	−14	−11	−27	−28	−5	−24	−26	−20	−16	−20	−24	−24	−24	7	6	−5
C	0	−9	−12	−26	12	−23	−24	−4	−5	9	8	−13	9	9	2	−2	2	10	8	6	−14	−24	−8
Q	−16	−6	−2	6	−23	9	8	−12	−8	−25	−25	−2	−22	−24	−17	−14	−17	−21	−22	−22	5	8	−6
E	−17	−7	−3	5	−24	8	7	−13	−9	−25	−26	−3	−22	−24	−18	−14	−18	−22	−22	−23	5	8	−6
G	3	2	−1	−14	−4	−12	−13	7	7	−6	−6	−2	−3	−5	2	5	2	−2	−2	−3	−2	−12	−5
H	3	6	3	−11	−5	−8	−9	7	10	−6	−7	2	−3	−5	2	5	1	−2	−3	−3	1	−9	−6
I	−2	−10	−14	−27	9	−25	−25	−6	−6	8	7	−15	7	8	1	−4	0	9	6	4	−15	−25	−7
L	−3	−11	−15	−28	8	−25	−26	−6	−7	7	6	−16	7	7	0	−5	0	8	6	4	−16	−26	−6
K	−6	3	8	−5	−13	−2	−3	−2	2	−15	−16	8	−12	−14	−7	−4	−7	−11	−12	−12	7	−3	−5
M	1	−7	−11	−24	9	−22	−22	−3	−3	7	7	−12	10	8	4	−1	3	11	9	7	−12	−22	−8
F	−1	−9	−13	−26	9	−24	−24	−5	−5	8	7	−14	8	9	2	−3	1	10	7	5	−14	−24	−7
P	6	−3	−6	−20	2	−17	−18	2	2	1	0	−7	4	2	8	4	8	4	4	3	−7	−17	−6
S	5	1	−3	−16	−2	−14	−14	5	5	−4	−5	−4	−1	−3	4	7	4	0	−1	−1	−4	−14	−5
T	6	−3	−6	−20	2	−17	−18	2	1	0	0	−7	3	1	8	4	8	4	4	3	−8	−18	−6
W	2	−7	−10	−24	10	−21	−22	−2	−2	9	8	−11	11	10	4	0	4	12	10	8	−11	−21	−9
Y	1	−7	−11	−24	8	−22	−22	−2	−3	6	6	−12	9	7	4	−1	4	10	10	8	−12	−22	−7
V	1	−8	−11	−24	6	−22	−23	−3	−3	4	4	−12	7	5	3	−1	3	8	8	7	−12	−22	−6
B	−6	3	7	7	−14	5	5	−2	1	−15	−16	7	−12	−14	−7	−4	−8	−11	−12	−12	7	5	−6
Z	−16	−7	−3	6	−24	8	8	−12	−9	−25	−26	−3	−22	−24	−17	−14	−18	−21	−22	−22	5	8	−6
X	−5	−5	−6	−5	−8	−6	−6	−5	−6	−7	−6	−5	−8	−7	−6	−5	−6	−9	−7	−6	−6	−6	−6

Table 10); yet, three amino acids contribute to most of the remaining differences:

- Pro and Gly are considered to be non-polar residues but the cyclic structure of Pro and the small size of Gly permit them to be exposed in the surface (Nelson and Cox, 2000).
- Cys is a polar residue but two Cys can be readily oxidized to form a disulfide bond which is highly hydrophobic (Chen et al., 2004).

Table 8
Comparison of the ability of different scoring matrices in discriminating between protein families. Entries show the number of families which their lowest *z*-score was greater than 4 (among the 328 hard-to-discriminate families)

Rep-type	fn-blosum65	fn-blosum50	Fn-pam250	fn-pam50	fn-scv	fn-polar_req	fn-ident 6/−1
Longest	56	55	63	51	52	49	47
Furthest 1	79	79	84	65	68	62	61
Furthest 2	66	66	78	56	58	54	51
Center	78	81	86	63	68	61	61

Table 9
Comparison of the ability of different scoring matrices in discriminating between protein families

Rep-type	d-blosum65	d-blosum50	d-pam250	d-pam50	d-scv	d-polar_req	d-ident 6/−1
Longest	5.26	5.20	5.09	5.60	5.41	5.62	5.53
Furthest 1	5.27	5.16	5.00	5.77	5.43	5.74	5.65
Furthest 2	5.43	5.35	5.20	5.86	5.59	5.85	5.79
Center	5.40	5.29	5.12	5.92	5.56	5.90	5.81

The averages of difference between highest and lowest *z*-score for 328 families are tabulated.

Table 10
Different amino acid indices and their pairwise correlation (Pearson) with the solvent accessibility index

	Solvent accessibility	Hydropathic character	Hydrophobicity scale #1	Hydrophobicity scale #2	Polar requirement	Residue charges	Residue volume
Phe	26.8	2.08	2.8	2.8	5.0	0	0.774
Leu	26.7	1.97	1.8	3.8	4.9	0	0.636
Ile	23.9	2.5	1.8	4.5	4.9	0	0.636
Met	33.4	2.3	1.3	1.9	5.3	0	0.613
Val	27.5	1.76	1.5	4.2	5.6	0	0.476
Ser	59.2	−0.63	−0.3	−0.8	7.5	0	0.172
Pro	64.4	−2.96	1.4	−1.6	6.6	0	0.373
Thr	57.0	−0.5	0.4	−0.7	6.6	0	0.334
Ala	40.2	0.38	0.5	1.8	7.0	0	0.17
Tyr	39.3	−0.44	2.3	−1.3	5.4	0	0.796
His	57.7	0.09	0.5	−3.2	8.4	1	0.555
Gln	74.6	−1.3	−0.2	−3.5	12.5	0	0.5
Asn	69.7	−1.21	−0.2	−3.5	10.0	0	0.344
Lys	88.0	−3.51	−3.0	−3.9	10.1	1	0.647
Asp	76.6	−1.27	−2.5	−3.5	13.0	−1	0.304
Glu	82.5	−1.88	−2.5	−3.5	12.5	−1	0.467
Cys	31.5	1.98	1.0	2.5	4.8	0	0.289
Trp	31.2	1.07	3.4	−0.9	5.2	0	1.0
Arg	75.8	−2.35	−3.0	−4.5	9.1	1	0.676
Gly	53.3	0.1	0.0	−0.4	7.9	0	0.0
R		−0.92	−0.87	−0.91	−0.89	0.05	−0.26

The solvent accessibility index is capable of reflecting these properties whereas the other included indices fail to do so. The solvent accessibility index is highly correlated with other polarity-based amino acid indices and including them in one cost measure matrix seems redundant. On the other hand, the residue charge and the residue volume indices show low correlations and can be considered as independent parameters in measuring the cost of substitutions. Based on these assumptions Eq. (3) was chosen to represent the physicochemical differences between the coding amino acids. Besides these analyses, it is fairly expected that polarity, volume, and charge would be independent properties and in order to evaluate the cost of mutation they all should be considered.

4.2. SCV matrix is unbiased towards the genetic code

The performed regression analyses (as in Di Giulio, 2001) suggest the absence of contamination in the SCV matrix as the obtained *R*-values are comparable to those of

other unbiased indices. In particular, the *R*-values obtained for the SCV matrix are even smaller than those of Polar requirement which is considered to be unbiased towards the canonical genetic code.

Apart from the SCV matrix, we have also included the Mutation matrix devised by Gilis et al. (2001) based on the *in silico* study of changes in folding free energies of many sampled proteins. Our data suggest that Mutation matrix is also unbiased and its use in a number of previous studies (Gilis et al., 2001; Goodarzi et al., 2004; Goodarzi et al., 2005a, b, c) might be justified regarding this aspect.

Use of PAM_{74-100} in a number of prior studies to highlight the error-buffering nature of the genetic code (Freeland and Hurst, 1998; Gilis et al., 2001) was criticized based on the biased nature of PAM_{74-100}. Thus, studies were forced to use amino acid indices, which although unbiased towards the genetic code, contain very limited information compared to the high complexity of amino acids as building blocks of proteins. Using SCV matrix for load minimization analysis on the genetic code shows that this matrix results in a higher optimality compared to simple amino acid indices. Hypothetically, if we could model the differences between amino acids perfectly, maybe then the error-buffering nature of the code would have been even more significant.

4.3. SCV matrix and load minimization

Whether the major selectional pressure or solely a byproduct of the evolution of the code, "load minimization" is one of the apparent characteristics of the canonical genetic code (for review see Freeland et al., 2003; Di Giulio, 2005). Substitution matrices such as PAM_{74-100} were used to elicit the extent by which this special characteristic is non-random in the structure of the canonical genetic code (Gilis et al., 2001, Freeland et al., 2000); yet, Di Giulio (2001) discussed the biasness of PAM matrices towards the canonical genetic code and criticized its use in such studies. Although common amino acid indices (e.g. Polar requirement) are as well capable of eliciting load minimization in the structure of the code, applying a more robust and yet unbiased matrix seems crucial for quantifying the extent by which the canonical genetic code is non-randomly structured to minimize the effects of errors. Being a better approximation for differences in general properties of amino acids and also suggested to be unbiased towards the genetic code, the SCV matrix is an acceptable candidate to serve this goal.

As previously mentioned in Section 3.3, using the SCV matrix as a cost measure matrix results in a *z*-score far smaller than any other amino acid index (biased or unbiased). Besides, the calculated *z*-score of -2.13 for the SCV matrix is the closest value to the *z*-score of PAM_{74-100} (i.e. -2.43).

Table 5, includes the amino acid indices that show more than 0.75 correlations with the SCV matrix. All of these indices are polarity or accessibility based, and their high correlation might be attributed to the presence of solvent accessibility index in the SCV matrix. Yet, none of these indices are as efficient in revealing the error-buffering capacity of the canonical genetic code.

4.4. Discriminating the "hard-to-discriminate"

Regardless of the scoring matrix used, the representative called "center" sounds better than the others in implicating protein family properties. Note that in Table 8, greater numbers mean more efficient scoring matrix, while in Table 9, the larger the value, the more indistinguishable it is. Both results in Tables 8 and 9 reveal similar ability for PAM50 and SCV-derived scoring matrices; however the latter seems a bit better. The highest ability for discrimination between protein families by PAM250 and the worst one by 6/−1 identity matrices had been expected. These results suggest that the SCV matrix is not just a random combination of three indices, which accidentally reveals a low *z*-score in load minimization studies. Besides, the SCV matrix is more capable in discriminating protein families when compared to Polar requirement which is a widely used amino acid index.

In short, we have introduced the SCV matrix as a distance-based matrix, which is a linear combination of solvent accessibility, charge and molecular volume of amino acids. Our results show that this matrix is unbiased towards the genetic code as opposed to the frequency-based scoring matrices (e.g. PAM matrices). On the other hand, we have also highlighted the high information content of this matrix through protein family discrimination comparisons with other matrices.

Acknowledgments

Authors are grateful to Mehdi Sadeghi and Hamid Pezeshk for their useful comments and their support.

Appendix A. Supplementary Materials

The online version of this article contains additional supplementary data. Please visit doi:10.1016/j.jtbi.2006.12.014.

References

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

Altschul, S.F., 1991. Amino acid substitution matrices from an information theoretic perspective. J. Mol. Biol. 219, 555–565.

Archetti, M., 2004. Codon usage bias and mutation constraints reduce the level of error minimization of the genetic code. J. Mol. Evol. 59, 258–266.

Bairoch, A., Boeckmann, B., Ferro, S., Gasteiger, E., 2004. Swiss-Prot: juggling between evolution and stability. Brief. Bioinform. 5, 39–55.

Bastien, O., Aude, J.C., Roy, S., Marechal, E., 2004. Fundamentals of massive automatic pairwise alignments of protein sequences: theoretical significance of *Z*-value statistics. Bioinformatics 20, 534–537.

Benner, S.A., Cohen, M.A., Gonnet, G.H., 1994. Amino acid substitution during functionally constrained divergent evolution of protein sequences. Protein Eng. 7 (11), 1323–1332.

Biou, V., Gibrat, J.F., Levin, J.M., Robson, B., Garnier, J., 1988. Secondary structure prediction: combination of three different methods. Protein Eng. 2, 185–191.

Bonnet, E., Van de Peer, Y., 2002. zt: a software tool for simple and partial Mantel tests. J. Stat. Software 7 (10), 1–12.

Chen, H., Zhou, H.X., Hu, X., Yoo, I., 2004. Classification comparison of prediction of solvent accessibility from protein sequences. In: Second Asia-Pacific Bioinformatics Conference, Australian Computer Society, Inc., p. 29.

Di Giulio, M., 1989. Some aspects of the organization and evolution of the genetic code. J. Mol. Evol. 29, 191–201.

Di Giulio, M., 1997a. On the origin of the genetic code. J. Theor. Biol. 187, 573–581.

Di Giulio, M., 1997b. The origin of the genetic code. Trends Biochem. Sci. 22, 49–50.

Di Giulio, M., 1999. The coevolution theory of the origin of the genetic code. J. Mol. Evol. 48, 253–255.

Di Giulio, M., 2000. Genetic code origin and the strength of natural selection. J. Theor. Biol. 205, 659–661.

Di Giulio, M., 2001. The origin of the genetic code cannot be studied using measurements based on the PAM matrix because this matrix reflects the code itself, making any such analyses tautologous. J. Theor. Biol. 208, 141–144.

Di Giulio, M., 2005. The origin of the genetic code: theories and their relationships, a review. BioSystems 80, 175–184.

Ellington, A.D., Khrapov, M., Shaw, C.A., 2000. The scene of a frozen accident. RNA 6, 485–498.

Engelman, D.M., Steitz, T.A., Goldman, A., 1986. Identifying nonpolar transbilayer helices in amino acid sequences of membrane proteins. Annu. Rev. Biophys. Chem. 15, 321–353.

Fauchere, J.L., Pliska, V., 1983. Hydrophobic parameters pi of amino-acid side chains from the partitioning of *N*-acetyl-amino acid amides. Eur. J. Med. Chem. 18, 369–375.

Finn, D.R., Mistry, J., Schuster-Böckler, B., Griffiths-Jones, S., Hollich, V., Lassmann, T., Moxon, S., Marshall, M., Khanna, A., Durbin, R., Eddy, S.R., Sonnhammer, E.L.L., Bateman, A., 2006. Pfam: clans, web tools and services. Nucl. Acids Res., Database Issue 34, D247–D251.

Fitch, W.M., Upper, K., 1987. The phylogeny of tRNA sequences provides evidence for ambiguity reduction in the origin of the genetic code. Cold Spring Harb. Symp. Quant. Biol. 52, 759–767.

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

Freeland, S.J., Knight, R.D., Landweber, L.F., Hurst, L.D., 2000. Early fixation of an optimal genetic code. Mol. Biol. Evol. 17, 511–518.

Freeland, S.J., 2002. The Darwinian code: an adaptation for adapting. J. Gen. Progr. Evol. Mach. 3, 113–127.

Freeland, S.J., Wu, T., Keulmann, T., 2003. The case for an error minimizing standard genetic code. Orig. Life Evol. Biosph. 33, 457–477.

Gilis, D., Massar, S., Cerf, N.J., Rooman, M., 2001. Optimality of the genetic code with respect to protein stability and amino acid frequencies. Genome Biol. 2 (11), 49.1–49.12.

Goldberg, A.L., Wittes, R.E., 1966. Genetic code: aspects of organization. Science 153, 420–424.

Goodarzi, H., Nejad, H.A., Torabi, N., 2004. On the optimality of the genetic code, with the consideration of termination codons. BioSystems 77, 163–173.

Goodarzi, H., Najafabadi, H.S., Nejad, H.A., Torabi, N., 2005a. The impact of including tRNA content on the optimality of the genetic code. Bull. Math. Biol. 67, 1355–1368.

Goodarzi, H., Najafabadi, H.S., Hassani, K., Nejad, H.A., Torabi, N., 2005b. On the optimality of the genetic code, with the consideration of coevolution theory by comparison of prominent cost measure matrices. J. Theor. Biol. 235, 318–325.

Goodarzi, H., Najafabadi, H.S., Torabi, N., 2005c. On the coevolution of genes and genetic code. Gene 362, 133–140.

Grantham, R., 1974. Amino acid difference formula to help explain protein evolution. Science 185, 862–864.

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

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

Henikoff, S., Henikoff, J.G., 1993. Performance evaluation of amino acid substitution matrices. Proteins 17, 49–61.

Kyte, J., Doolittle, R.F., 1982. A simple method for displaying the hydropathic character of a protein. J. Mol. Biol. 157, 105–132.

Naderi-Manesh, H., Sadeghi, M., Arab, S., Moosavi Movahedi, A.A., 2001. Prediction of protein surface accessibility with information theory. Proteins 42, 452–459.

Nelson, D., Cox, M., 2000. Lehninger Principles of Biochemistry, third ed., p. 118.

Nishikawa, K., Ooi, T., 1986. Radial locations of amino acid residues in a globular protein. J. Biochem. 100, 1043–1047.

Nozaki, Y., Tanford, C., 1971. The solubility of amino acids and two glycine peptides in aqueous ethanol and dioxane solutions: establishments of a hydrophobicity scale. J. Mol. Chem. 246, 2111.

Rose, G.D., Geselowitz, A.R., Lesser, G.J., Lee, R.H., Zehfus, M.H., 1985. Hydrophobicity of amino acid residues in globular proteins. Science 229, 834–838.

Sonneborn, T.M., 1965. Degeneracy of the genetic code, extent, nature, and genetic implications. In: Bryson, V., Vogel, H.J. (Eds.), Evolving Genes and Proteins. Academic Press, New York.

Woese, C.R., Dugre, D.H., Dugre, S.A., Kondo, M., Saxinger, W.C., 1966. On the fundamental nature and evolution of the genetic code. Cold Spring Harbor Symp. Quant. Biol. 31, 723–736.

Wong, J.T., 1975. A co-evolution theory of the genetic code. Proc. Natl Acad. Sci. USA 72, 1909–1912.

Zamyatin, A.A., 1972. Protein volume in solution. Prog. Biophys. Mol. Biol. 24, 107–123.

Zhu, W., Freeland, S., 2006. The standard genetic code enhances adaptive evolution of proteins. J. Theor. Biol., available online.

Zhu, C.T., Zeng, X.B., Huang, W.D., 2003. Codon usage decreases the error minimization within the genetic code. J. Mol. Evol. 57, 533–537.

Zuckerkandl, E., Pauling, L., 1965. Evolutionary divergence and convergence in proteins. In: Bryson, V., Vogel, H.J. (Eds.), Evolving Genes and Proteins. Academic Press, New York.