J Mol Evol (1998) 47:1–13

JOURNAL OF MOLECULAR EVOLUTION

On Error Minimization in a Sequential Origin of the Standard Genetic Code

David H. Ardell

Department of Biological Sciences, Stanford University, Stanford, CA 94305, USA

Received: 28 July 1997 / Accepted: 23 January 1998

Abstract. Distances between amino acids were derived from the polar requirement measure of amino acid polarity and Benner and co-workers' (1994) 74-100 PAM matrix. These distances were used to examine the average effects of amino acid substitutions due to single-base errors in the standard genetic code and equally degenerate randomized variants of the standard code. Second-position transitions conserved all distances on average, an order of magnitude more than did second-position transversions. In contrast, first-position transitions and transversions were about equally conservative. In comparison with randomized codes, second-position transitions in the standard code significantly conserved mean square differences in polar requirement and mean Benner matrix-based distances, but mean absolute value differences in polar requirement were not significantly conserved. The discrepancy suggests that these commonly used distance measures may be insufficient for strict hypothesis testing without more information. The translational consequences of single-base errors were then examined in different codon contexts, and similarities between these contexts explored with a hierarchical cluster analysis. In one cluster of codon contexts corresponding to the RNY and GNR codons, second-position transversions between C and G and transitions between C and U were most conservative of both polar requirement and the matrix-based distance. In another cluster of codon contexts, second-position transitions between A and G were most conservative. Despite the claims of previous authors to the contrary, it is shown theoretically that the standard code may have been shaped by position-invariant forces such as mutation and base content. These forces may have left heterogeneous signatures in the code because of differences in translational fidelity by codon position.

A scenario for the origin of the code is presented wherein selection for error minimization could have occurred multiple times in disjoint parts of the code through a phyletic process of competition between lineages. This process permits error minimization without the disruption of previously useful messages, and does not predict that the code is optimally error-minimizing with respect to modern error. Instead, the code may be a record of genetic process and patterns of mutation before the radiation of modern organisms and organelles.

Key words: Error minimization — G/C bias — Transition bias — Translational error — Amino acid substitution matrices — Codon context — RNY hypothesis — RNA world

Introduction

The Error-Minimization Hypothesis of Standard Code Origin

All modern organisms bequeath to their offspring genes encoding proteins and machinery to decode those genes. Transmission errors of encoded protein information (''message errors'') occur because of translational misreading or misacylation, because the gene was altered by mutagenic damage in either the parent or the offspring or

Correspondence to: D.H. Ardell; *e-mail:* ardell@charles.stanford.edu

by replicative errors.[1] A lineage with a code such that these transmission errors result in chemically conservative amino acid substitutions may be more fit than a lineage with a less conservative code.

The hypothesis that message errors influenced the assignment of amino acid meaning to codons is as old as our knowledge of the standard code itself (Nirenberg et al. 1963; Sonneborn 1965; Woese 1965; Goldberg and Wittles 1966; Epstein 1966). The error-minimization hypotheses states that the standard code evolved an inverse relationship between the severity and the frequency of message errors (Swanson 1984; Haig and Hurst 1991) *at the time of fixation of codon meaning* (strictly speaking, it is not errors but rather their consequences that are minimized). A formal statement of this argument is given in Appendix A, with a proof of optimality of the inverse relationship between error frequency and severity.

Representing the Fitness Effects of Substitution Using Amino Acid Distances

A surrogate for directly measuring the fitness effects of amino acid substitutions is the use of distances derived either from sequence-alignment-based amino acid substitution matrices or from differences in amino acid physicochemical properties shown to correlate with such matrices. Generally, amino acid polarity and volume have been identified as most related to patterns of amino acid substitution (Grantham 1974; French and Robson 1983; Swanson 1984; Benner et al. 1994; Tomii and Kanehisa 1996), although different matrices and methods identify different specific measures of these properties as most explanatory of the data. Interestingly, amino acid volume measures (Grantham 1974) seem less related than polarity measures to the pattern of amino acids in the standard code (Haig and Hurst 1991; Di Giulio 1994).

Because amino acid substitution matrices are derived from alignments of modern proteins, it is assumed in both methods of deriving distances for testing hypotheses about the code that the fitness-relevant chemistry of proteins has not changed since the time the code—or its parts—has fixed. Distances derived directly from substitution matrices are sensitive to amino acid frequencies in the sequence set from which they are generated (Altschul 1991), which are likely to have changed over the history of life. However, such distances incorporate information from the averaged effect of many different protein contexts weighted by frequency of occurrence.

As reviewed in the next section, both squared (Haig and Hurst 1991; Goldman 1993) and absolute-value differences (Alff-Steinberger 1969; Di Giulio 1989a; Szathmary and Zintzaras 1992; Di Giulio 1995a) in polar requirement[2] ($|\Delta PR|^2$ and $|\Delta PR|$, respectively) have been used to test hypotheses about the genetic code. Although other measures of polarity have been shown to be more explanatory of matrix substitution data than polar requirement (Tomii and Kanehisa 1996; Koshi and Goldstein 1997), the differences are slight. As is shown in this study, statistical results may depend critically on how the distances are transformed before analysis.

What Kinds of Message Errors Are Thought to Have Influenced Code Evolution?

On average, errors in the first position of the code conserve chemical polarity much more than in the second position (Woese 1965; Alff-Steinberger 1969; Kimura 1980; Swanson 1984; Haig and Hurst 1991). This is true with different measures of chemical polarity, including polar requirement, and Kyte–Doolittle hydropathy (Haig and Hurst 1991). Consequently, the second codon position determines the chemistry of the encoded amino acid (Wolfenden et al. 1979; Sjöström and Wold 1985; Di Giulio 1989b; Taylor and Coates 1989).

The translational error model for code evolution (Woese 1965) explains this positional discrepancy in conservation of chemical polarity in the code as a consequence of the greater frequency of translational misreading in the first codon position relative to the second position as observed in vitro (Davies et al. 1964, 1966). More recent in vivo studies of translational error support this positional trend in general (Parker 1989). Assuming that studies of extant translational error can tell us about translational error at the time the code originated, the correspondence between translational error frequencies by codon position and their resulting average conservation of chemical polarity in the standard code is consistent with translational error minimization, but not—it has been claimed—mutation minimization (Alff-Steinberger 1969; Swanson 1984; Haig and Hurst 1991; Goldman 1993). These authors have argued that the frame invariance of mutations should have resulted in equal conservation of amino acid distance in each codon position.

The main thesis of this paper is that error minimization with respect to frame-invariant error processes such as mutation is *not* inconsistent with positional discrepancies in conservation. Message mutations must first be translated to affect code fitness. Therefore, positional differences in translational fidelity might have caused a positional asymmetry in the effect of message mutations on code evolution. For example, if primordial genomes were G/C-rich (evidence reviewed below), the greater

[1] Error is defined here with reference to Watson–Crick fidelity and is, thus, restricted to the first or second codon position in translation.

[2] A chromatographic measure of hydrophilicity, namely, the logarithm of the slope of the line resulting from a plot of log $[(1 - R_F)/R_F]$ against log mole fraction water in a series of pyridine solvents of increasing polarity.

translational fidelity of the second codon position would lead to a greater mutational signature of this bias in the pattern of amino acid assignments in the second codon position than in the first position. To explore this hypothesis, the average effects of transitions and transversions were examined in the first and second codon positions, and the effects of all possible errors in different codon contexts were examined and compared to one another.

The History of Ancient Genomes and the Genetic Code

There has been independent evidence that primordial genomes were G/C-rich. Eigen and Schuster argue for primordial G/C-richness as a consequence of selection for replicative fidelity. Before well-adapted replicases, the G · C pair is thought to have replicated with greater fidelity because of its greater relative thermodynamic stability (Eigen and Schuster 1979). Recently, an in vitro-evolved telomerase-like ribozyme was shown to polymerize nucleotides 10–40 times more efficiently and with a higher fidelity when directed by G and C template residues than with A and U residues (Ekland and Bartel 1996).[3] Furthermore, phylogenetically reconstructed tRNA ancestors have a higher G/C content than is the average in extant tRNA molecules (Eigen and Winkler-Oswatitsch 1981; Fitch and Upper 1987; Di Giulio 1995b). In addition, the GNC-encoded amino acids (glycine, alanine, aspartate, and valine) are the most abundant amino acids produced in the Miller–Urey reactions and found in the Murchison Meteorite (Miller 1987) and are also at the roots of the amino acid biosynthetic pathways (Wong 1975; Taylor and Coates 1989). A recent phylogenetic analysis of acceptor domains of tRNA molecules suggests that all tRNA molecules are descended from GNC codon-recognizing adapters (Rodin et al. 1996).

What kinds of errors may have affected the code through error minimization? Did the code evolve its amino acid meaning all at once, or in successive stages? How can error minimization be reconciled with the notion, due to Crick (1968), that genetic codes cannot change without drastic losses in fitness due to the disruption of message meaning? Evidence is presented here to suggest that the code evolved in at least two distinct stages of error minimization, each in response to different distributions of error. Furthermore, a phyletic scheme is suggested for how this could have occurred without disrupting message meaning.

Methods

Amino Acid Distances: Criteria and Methods

Values of Woese's polar requirement[4] and Benner's 74-100 PAM amino acid substitution matrix were taken from the AAINDEX database (Nakai et al. 1988; Tomii and Kanehisa 1996).

Both $|\Delta PR|^2$ and $|\Delta PR|$ were used here in comparing average distance conservation of transitions and transversions in the first and second codon positions. In this study, the third position was neglected because its high degree of degeneracy makes it incomparable to the other two positions.

A third distance, D_B, was generated from Benner and co-workers' (1994) 74-100 PAM amino acid substitution matrix **M.** This is a log-odds matrix, the entries of which must be exponentiated for arithmetic averaging. The distance $D_B(i,j)$ between amino acid i and amino acid j is calculated from the matrix entry $M_{i,j}$ using the formula

$$D_B(i,j) = \begin{cases} 10^{-(M_{i,j}/10)} & \text{if} \quad i \neq j \\ 0 & \text{if} \quad i = j \end{cases} \tag{1}$$

which for $i \neq j$ is identical to the transformation used by Benner et al. (1994) to define distances in their study. With reference to the sequence set that generated the matrix **M,** $D_B(i, j)$ will vary directly with the product of the frequencies of amino acids i and j in the sequence set and inversely with their frequency of pairwise alignment (Altschul 1991).

The Benner 74-100 PAM matrix was used because it is based on a set of sequences of large estimated PAM distances within a specific range. In comparison with matrices derived from less-diverged sequences, it is demonstrably less affected by the genetic code, and more correlated with physicochemical properties of amino acids (Benner et al. 1994; Tomii and Kanehisa 1996). Furthermore, the matrix entries are significant to two digits instead of one, increasing the accuracy of calculated distances when transformed by Eq. (1).

Comparisons of Transitions and Transversions in the Standard Code and in Randomized Variants

Average values of conservation of the two polar requirement-based distances and the matrix-based distance D_B were calculated for transitions and transversions in the first and second codon positions of the standard code. The equal weighting of each codon in the calculation of these averages is consistent with the assumption of a mutation model with a uniform stationary distribution, such as the Kimura two-parameter model used in Appendix B (assuming no other evolutionary forces).

To estimate the probability that the results would have occurred by chance alone, distributions of averages and their standard errors for each error category and position were empirically calculated using randomized codes in the method of Haig and Hurst (1991). In this method, distributions are generated by sampling from the 20! possible permutations of amino acids within the standard code, leaving constant the size and location of codon sets that encode for the same amino acid and the location and number of stop codons. The proportion of randomized codes with an average distance less than or equal to that of the standard code for a given chemical criterion, codon position, and base type defines a P value, an estimate of difference from the null model

[3] This could reflect the evolutionary history of the ribozyme, as it was evolved from a ligase that used a tethered G-rich template. However, that the descendent ribozyme also has higher fidelity with a C template suggests that the bias in fidelity is extrinsic to its evolutionary history.

[4] There are two published sets of values with this name (Woese 1966, 1973), differing in values for Cys, Trp, and Tyr (at most by nearly 15% for Cys). The more recent set of values is used here, but the results do not differ appreciably from those with the older values.

Table 1. Means of the distances $|\Delta PR|^2$, $|\Delta PR|$, and D_B in the standard code $\overline{D}_{SGC}$, their averages $\overline{\overline{D}}$ in 10^6 randomized codes, and the proportion P of such codes at least as conservative as the standard code by position and type of error

		Position I		Position II	
Distance	Statistic	Transitions	Transversions	Transitions	Transversions
$D = \|\Delta PR\|^2$	$\overline{D}_{SGC}$	3.21	4.94	5.54	13.00
	$\overline{\overline{D}} \pm$ SE	11.3 ± 3.8	11.7 ± 3.2	12.1 ± 3.9	12.1 ± 2.8
	P	5.6×10^{-3b}	6.2×10^{-3b}	3.2×10^{-2a}	6.5×10^{-1}
$D = \|\Delta PR\|$	$\overline{D}_{SGC}$	1.32	1.43	1.94	3.10
	$\overline{\overline{D}} \pm$ SE	2.57 ± 0.51	2.66 ± 0.43	2.76 ± 0.53	2.76 ± 0.38
	P	6.1×10^{-3b}	1.1×10^{-3b}	5.9×10^{-2}	8.2×10^{-1}
$D = D_B$	$\overline{D}_{SGC}$	0.80	0.91	1.13	1.34
	$\overline{\overline{D}} \pm$ SE	1.31 ± 0.17	1.36 ± 0.13	1.41 ± 0.17	1.41 ± 0.12
	P	2×10^{-4c}	1×10^{-4c}	4.2×10^{-2a}	2.8×10^{-1}

[a] $P < 0.05$.
[b] $P < 0.01$.
[c] $P < 0.001$.

used. Clearly the choice of null model used will influence the estimate of statistical significance (Goldman 1993). In this null model, biases arise from nonrandomness in the location of sites of large and small degeneracies (Haig and Hurst 1991).

In the initial comparisons between transitions and transversions, five tests were run with each distance measure, with 10^6 randomized codes examined in each test. All statistics were rounded to the level of significance determined from these five repeated measurements. Identical calculations were also performed with randomized codes constrained to be as conservative as the standard code in first-position transitions and transversions, except that only a single run was performed for each estimate. All calculations described above were performed in ANSI C.

Detailed Analysis of the Effects of Error in Different Codon Contexts

Values of D_B and signed values of ΔPR were calculated for every amino acid substitution that would result from a single-base error in the first or second position of the standard code. Each of the $\binom{4}{2}$ undirected errors between two bases was organized by the position in which it occurred and by the context of the error. The *codon context* of an error classifies the error at a given codon position by the identity of the bases unchanged by the error. For example, the set of all second-position errors in the context of a first-position A and a third-position pyrimidine is denoted ''A*Y.''

Signed values of ΔPR were generated to assess similarities among different codon contexts in both the directions and the magnitudes of distances. This was not possible with the symmetrical D_B distance. For the correlation analysis described below, signed values were used. Otherwise, only the ranks of magnitudes were compared in different contexts to infer rates of errors between the different pairs of bases in those contexts, assuming an error-minimization process.

Values of ΔPR and D_B for errors involving the codons AUA (isoleucine) and AUG (methionine) were averaged (in the case of ΔPR, they were of the same sign), and errors to and from stop codons were excluded from all calculations. All calculations above were performed in ANSI C.

To summarize similarities among the patterns of the data in different contexts, Pearson product–moment correlations were calculated between the different contexts for the second position alone and for the first and second positions combined. The two contexts containing errors exclusively to and from stop codons (*AR and U*R) were excluded from these correlation calculations. A hierarchical cluster analysis was performed on the correlations in S-Plus (Version 3.4. MathSoft, Inc., 1997), using the ''connected'' (also called ''single-linkage'') method, which defines the correlation between two clusters as the highest correlation among two members each from a different cluster.

Results

Transitions Are More Conservative Than Transversions in the Second Codon Position

Table 1 shows that with both the $|\Delta PR|^2$ and the D_B distances, less than 5% of randomized codes were more conservative than the standard code in second-position transitions, while second-position transversions were not significantly conservative in this measure. In the first position, transitions and transversions were both highly and significantly conservative for all distances examined. Transitions were not significantly different from random in the second codon position with the $|\Delta PR|$ distance, by a 0.05 criterion. The quantities calculated with the $|\Delta PR|^2$ distance were consistent in magnitude with those of Haig and Hurst (1991), who pooled transitions and transversions together. As with their results, differences in $\overline{\overline{D}}$ and their standard errors for a given distance among categories may be explained by the uneven pattern of degeneracy within each class of error and the different numbers of each type of error entering the averaging. Both transitions and transversions in the first position were averaged with two values of zero because of the degeneracy pattern of the standard code. Approximately twice as many nonzero transversions as transitions entered the calculation of averages, excluding errors to and from stop codons (29 transitions versus 58 transversions in the first position and 30 transitions versus 58 transversions in the second position).

To test whether error minimization in the first position could cause the transition-biased pattern in the second position, the calculations were repeated with randomized codes constrained to be as conservative as the standard

Table 2. Statistics of randomized codes conditioned to be at least as conservative as the standard code in first-position transitions and transversions (N gives the number of codes examined and sampled)

		Position I		Position II	
Distance	Statistic	Transitions	Transversions	Transitions	Transversions
$D = \|\Delta PR\|^2$	$\overline{\overline{D}} \pm$ SE	3.0 ± 0.6	4.3 ± 0.8	12.6 ± 4.2	11.6 ± 2.1
(N = 853 of 10^6)	P	1.0	1.0	5×10^{-2}	7×10^{-1}
$D = \|\Delta PR\|$	$\overline{\overline{D}} \pm$ SE	1.2 ± 0.13	1.3 ± 0.12	2.8 ± 0.6	2.6 ± 0.3
(N = 882 of 5×10^6)	P	1.0	1.0	1×10^{-1}	9×10^{-1}
$D = D_B$	$\overline{\overline{D}} \pm$ SE	0.77 ± 0.03	0.87 ± 0.03	1.62 ± 0.22	1.46 ± 0.11
(N = 172 of 7×10^7)	P	1.0	1.0	1×10^{-2}	7×10^{-2}

code in first-position transitions and transversions. Large numbers of randomized codes were needed to develop a sample size sufficient for crude estimates, and different distances required the examination of different numbers of codes to get approximately the same sample size. Table 2 shows the results of this analysis. For $D = |\Delta PR|^2$, the second-position statistics were quite similar to those in Table 1 but only borderline significant. $|\Delta PR|$ was more strongly affected, indicating a possible interaction in the two dimensions of organization in the code. Only 172 matrices of 7×10^7 examined satisfied the condition with the D_B distance, in keeping with the extreme P values of first-position transitions and transversions obtained with this distance (Table 1).

Patterns of ΔPR *and* D_B *Within Different Contexts of Error and Their Correlations in the Standard Code*

Tables 3 and 4 present the first and second positions of signed differences in polar requirement ΔPR and unsigned Benner matrix distances, D_B, respectively. Columns show values for errors labeled at the top in the contexts listed at the left. The sole purpose in showing signed values of ΔPR was to display the unique consistency of second-position errors, such that most values could be made positive in the second position by fixing the direction of errors in a specific way. As may be seen from the columns of oppositely signed values in the upper half of Table 3, this was impossible in the first codon position.

Several trends were apparent.

1. The second-position pyrimidine context was disproportionately responsible for the conservative nature of first-position errors with respect to both distances examined. These contexts correspond to errors within the hydrophobic and small polar residues.
2. With both distances in each of the A*Y, G*Y, and G*R contexts, G ↔ C errors were most conservative and A ↔ U errors were least or second-least conservative. C ↔ U errors were also conservative in these contexts.
3. In the A*R, C*Y, and C*R contexts, A ↔ G errors were most conservative, and with ΔPR in these contexts, both transition-type errors were more conservative than any transversion-type error.
4. With the ΔPR distance, the A*Y context of second-position errors was more similar to the G*Y and G*R contexts than to the A*R context, which in turn was more similar to the C*R and C*Y contexts. Closer inspection of the data showed that the difference between A*Y and A*R was due largely to arginine in the AGR codons.

Table 3. Signed differences in polar requirement ΔPR in the first and second positions of the standard code, by context and error type

	G → C	C → U	G → U	A → G	A → C	A → U
Position I						
*GY	−1.20	3.60	2.40	−0.40	−1.60	2.00
*GR	−1.20	3.80	2.60	1.20	0.00	3.80
*AY	4.60	2.70	7.30	−3.00	1.60	4.30
*AR	3.90	Stop	Stop	−2.40	1.50	Stop
*CY	0.40	−0.90	−0.50	−0.40	0.00	−0.90
*CR	0.40	−0.90	−0.50	−0.40	0.00	−0.90
*UY	0.70	−0.10	0.60	−0.70	0.00	−0.10
*UR	0.70	0.00	0.70	−0.50	0.20	0.20
Position II						
G*Y	0.90	1.40	2.30	5.10	6.00	7.40
G*R	0.90	1.40	2.30	4.60	5.50	6.90
A*Y	0.90	1.70	2.60	2.50	3.40	5.10
A*R	2.50	1.50	4.00	1.00	3.50	5.00
C*Y	2.50	1.70	4.20	−0.70	1.80	3.50
C*R	2.50	1.70	4.20	−0.50	2.00	3.70
U*Y	−2.00	2.50	0.50	0.20	−1.80	0.70
U*R	2.20	2.60	0.40	Stop	Stop	Stop

There were also notable differences between the results with ΔPR and those with D_B. In addition to the second-position pyrimidine contexts of first-position error, the *AY and *AR contexts were also very conservative of the D_B distance (Table 4, upper half). A ↔ G errors in the second position were conservative in every context using the D_B distance, but in polar requirement were only especially conservative in the A*R, C*Y, and C*R contexts. With both distances, the U*Y and U*R contexts of second-position errors appeared similar to each other and different from the other second-position

Table 4. Values of the Benner 74-100 PAM matrix-based distance D_B for the first and second positions of the standard code, by context and error type

	C ↔ G	U ↔ C	U ↔ G	A ↔ G	A ↔ C	U ↔ A
Position I						
*GY	1.26	1.66	1.58	0.91	1.05	0.98
*GR	1.26	1.45	2.57	1.26	0.00	1.45
*AY	0.91	0.56	1.91	0.60	0.76	1.38
*AR	0.68	Stop	Stop	0.76	0.68	Stop
*CY	0.91	0.89	0.78	0.85	0.98	0.72
*CR	0.91	0.89	0.78	0.85	0.98	0.72
*UY	0.65	0.62	0.98	0.48	0.52	0.81
*UR	0.65	0.00	0.65	0.57	0.52	0.52
Position II						
G*Y	0.87	0.98	2.04	0.95	1.07	1.95
G*R	0.87	0.98	2.04	1.12	1.02	1.62
A*Y	0.72	1.07	1.51	0.81	0.91	1.91
A*R	1.07	1.08	1.67	0.51	0.98	1.56
C*Y	1.26	1.66	1.74	0.79	1.26	1.55
C*R	1.26	1.66	1.74	0.69	1.05	1.48
U*Y	0.98	1.82	1.17	1.10	1.55	0.30
U*R	2.19	1.66	1.23	Stop	Stop	Stop

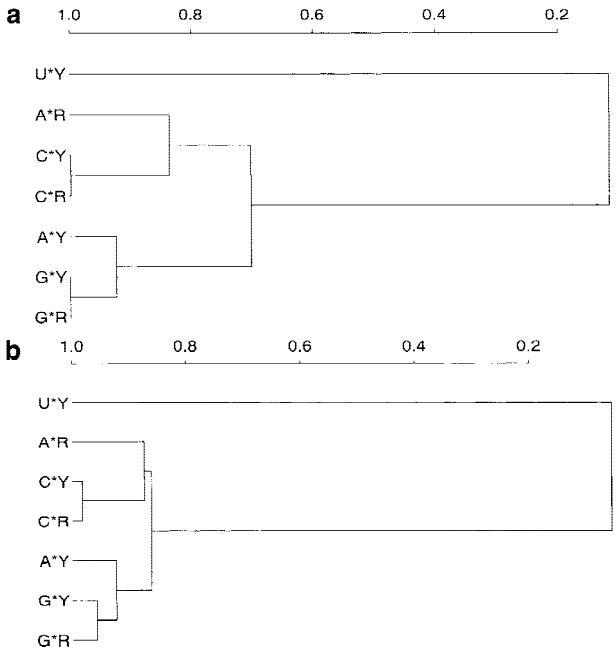

Fig. 1. Hierarchical cluster analyses of correlations among seven contexts of second-position errors (rows in Tables 3 and 4, bottom halves). **a** Correlations between different contexts of second position errors using signed differences in polar requirement ΔPR measure. **b** Correlations derived from the substitution matrix-based distance D_B among contexts of errors in position two. In both a and b, the U*R context containing errors to and from stop codons was excluded. The correlations between clusters are single-linkage (given by the maximum correlation between two members, one from each cluster). By giving an upper bound on correlation between clusters, single-linkage gives a sense of the strength of separation of the A*Y and A*R contexts; this separation is stronger with the ΔPR measure but is maintained in both.

error contexts, where the two contexts could be compared numerically.

A hierarchical cluster analysis of the second-position errors grouped by context is shown in Figs. 1a and b. The A*Y and A*R contexts clustered separately in the dendrograms, such that A*Y was closer to G*Y/G*R, and A*R was closer to C*Y/C*R. When first and second-position errors were pooled, the relative clustering relationships among the second-position errors were robust for correlations based on ΔPR but not D_B (Figs. 2a and b).

Discussion

Variation in the Results Based on Distances Examined

In comparing transitions and transversions with the two measures of difference in polar requirement, *P* values varied extremely depending on the transformation used (Tables 1 and 2). There was no a priori basis on which to decide which of the squared or absolute value differences in polar requirement was more appropriate for estimating the average effect of amino acid substitutions on fitness. Some way is needed to ascertain the importance of polar requirement to the functioning of ancient proteins (Crick 1968). Until this is possible, these statistics are not sufficient for strict hypothesis testing.

Most trends emphasized in this report were consistent when analyzed with other polarity measures and, also, with distances derived from other substitution matrices (data not shown).

Some discrepancies between the results in Table 3 and those in Table 4 were due to the influence of amino acid volume on the Benner matrix-based distance. For example, the differences between the two tables in errors involving U in the G*N/A*Y contexts are attributable to amino acid differences in volume, when measured by Grantham's amino acid volume index (data not shown). Either amino acid volume was not as important to the functioning of proteins at the time of code origin, or it was and D_B is better for estimates of ancient error, or the hypothesis is fundamentally flawed. In the following, greater emphasis is given to results obtained with both distances, and special mention is given to results inferred from the polar requirement measure alone.

A Second-Position-Pyrimidine Context May Promote First-Position Translational Error

Table 3 shows that the second-position pyrimidine contexts (*CY, *CR, *UY, and *UR) were primarily responsible for the conservative nature of first-position errors with respect to polar requirement. The pattern is also seen with the matrix-based distance in Table 4, but not as distinctly (the *AY and *AR contexts are also conservative).

One possible explanation for these data is that, given error minimization, a pyrimidine context relatively pro-

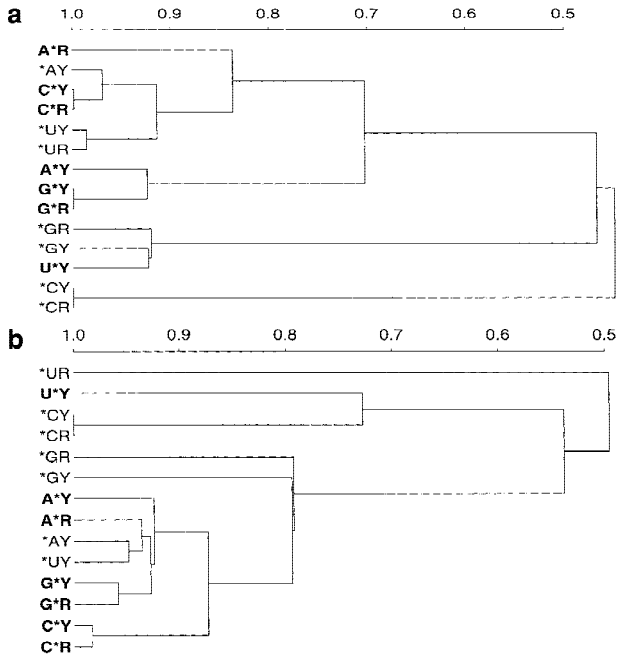

Fig. 2. Hierarchical cluster analyses of correlations among the pooled contexts of errors in the first and second positions (rows in Tables 3 and 4). **a** Correlations are calculated between different contexts of first- and second-position errors using signed differences in polar requirement ΔPR as a ''distance'' measure. **b** Correlations derived from the substitution matrix-based distance D_B among contexts of errors in positions one and two. In both a and b, two contexts containing errors exclusively to and from stop codons were excluded (*AR and U*R). Correlations between clusters are ''single-linkage'' as in Fig. 1. The relative clustering of second-position errors (in *boldface*) is preserved (with reference to Fig. 1) using the polar requirement-derived distance, despite the addition of data from the first position, but is disrupted using the matrix-derived distance. The more segregating ''group-linkage'' method restored these clusters with the matrix-based distance (data not shown).

moted first-position translational error at the time the code originated. This is consistent, assuming that modern rates of translational error may be extrapolated backward in time, with evidence that second-position pyrimidine contexts promote translational error in a cell-free extract taken from *Escherichia coli* under the action of streptomycin (Negre et al. 1988).[5] More evidence regarding context effects on first-position errors would be desirable to assess this claim. Unfortunately, in vitro studies of tRNA/tRNA binding affinities (Grosjean et al. 1978) do not examine first-position non-Watson–Crick pairs in different second-position contexts. In vivo translational error studies have relied mostly on labeled cysteine or charge heterogeneity in proteins to detect error and, thus, have been restricted to second-position purine contexts (Parker 1989).

[5] This same study did not find a significant difference in translational error between the two codon positions (Negre et al. 1988), but the balance of evidence suggests that the first position is more error-prone than the second codon position (Parker 1989).

Also consistent with a relatively stabilizing effect of second-position purine contexts on first-position translational error is the placement of stop codons in this context. Nonsense errors have been claimed to be more detrimental on average than any missense error in terms of fitness effects (Sonneborn 1965). Finally, if first-position translational error is infrequent in second-position purine contexts, then these contexts should reflect the characteristics of mutation more than other contexts of first-position error (Appendix B). Figure 2a shows that with the ΔPR distance, the *AY context was more than 0.95 correlated with the C*N contexts. However, the *GY and *GR contexts did not cluster near these contexts, and the *UY and *UR contexts were also more than 0.90 correlated with this cluster.

Transition-Biased Conservation in the Second Codon Position Could Be Explained by Translational Error Minimization

Why are the first and second codon positions different with respect to transitions and transversions? Translational error could have been more transition biased in the second codon position when the code originated. This explanation could be accommodated by preferential stability of either the G · U or the C · A non-Watson–Crick pair in the second codon position. A second-position G · U pair was presumably responsible for the observed in vitro translation of UGU as tyrosine under a high Mg^{2+} concentration (Nishimura et al. 1968). A second-position G · U pair may also explain observed misreading among nonsense suppressors (Strigini and Brickman 1973). Although the thermodynamics of base-pairing are not the sole determinant of RNA secondary structure, and of codon/anticodon interactions in particular (Grosjean et al. 1978), some assessment of the thermodynamics of RNA base-pairing under the nearest-neighbor model may be relevant (Turner 1996). The C · A pair is thermodynamically less stable in RNA than is the widely observed G · U pair (Wu et al. 1995). One in vitro system using tRNA/tRNA annealing found that the G · U pair was an order of magnitude more stable in the wobble position than in the middle position, where it has only marginal stability (Grosjean et al. 1978). Because this study used a tRNA/tRNA binding affinity assay, one tRNA's wobble position was another's first position. Thermodynamic measurements confirm that the G · U pair has a lower stability in the middle of an RNA helix (Turner et al. 1988) than at its end. The formation of middle-position G · U pairs may be promoted by an adjacent U nucleotide context (Friedman, in Davies 1966; Strigini and Brickman 1973).

On the basis of this evidence, translational error minimization cannot be excluded as at least contributing to the pattern of transition-biased distance conservation in the second codon position. A prediction from the data in

Tables 3 and 4 is that, since errors between A and G were generally more conservative than those between C and U, and so long as C · A pairs are much less important, if translational error is responsible for the pattern, then a template G initiates the G · U pair more often than a template U.

Mutation Minimization Is Consistent with Positional Discrepancies in Distance Conservation

Alternatively, minimization of the combined effects of mutation and translational error could explain the discrepancy in distance conservation of transitions in the first and second codon positions. It was observed early on that transitions conserve chemical polarity more than do transversions in the standard code (Goldberg and Wittes 1966), although the positional dependency of this in the code was not explored. This was interpreted to reflect an influence of mutation on the code. Mutations resulting from both replication errors and DNA damage are more often transitions than transversions, which led to the introduction of the two-parameter model for genetic sequence evolution (Kimura 1980). Given that the code may have originated in an RNA world, it is interesting to note that modern RNA replicases from polio virus demonstrate a transition bias in enzymatic RNA replication (Kuge et al. 1989).

Appendix B shows heuristically how positional differences in translational fidelity may lead to an influence of message mutations on error minimization in code evolution that depends on codon position. Under this model, if at least parts of the code have retained their original amino acid assignments [contrary to Crick's (1968) suggestion that the primitive code may have been completely wiped out], then the translational fidelity of the second codon position may have shaped the code in a way that could yield information about ancient genomes and their mutation processes.

The Organization of a Subset of the Code Is Consistent with Primordial Genomic G/C-Richness

Tables 3 and 4 show that among second-position errors in the G*N and A*Y contexts (or, equivalently, R*Y and G*R contexts), C ↔ G errors resulted in the most conservative amino acid substitutions. This was true when calculated with the D_B distance (Table 4) despite the large protein frequency of glycine, alanine, serine, and threonine (King and Jukes 1969), which tends to enlarge these distances. Furthermore, A ↔ U errors were least conservative of the Δ*PR* distance (Table 3) and second least conservative of D_B (Table 4) compared to any other error in these same contexts. The present interpretation of these data is that C ↔ G transversions (from mutation) may have been very frequent, and A ↔ U transversions very infrequent, in coding regions at the time the amino acid assignments of the RNY and GNR codons became fixed. This is consistent with independent evidence for G/C-content bias in ancient genomes introduced above, provided that some explanation can be made for why the code exhibits this pattern in only certain second-position contexts and not others.

Suppose selection for replication fidelity favored a high G/C content in the era that amino acids were assigned to the GNN and ANY codons. G ↔ C errors in a coding region would have occurred at a high frequency for two distinct reasons. First, a high G/C content would have caused such errors to be intrinsically more frequent than A ↔ U transversions [cf. the source distribution term, $p(\cdot)$ in Appendix A]. But if only a high G/C content were responsible, and not selection or mutation pressure for G/C richness, then error minimization according to a Kimura two-parameter model would predict transitions to be more conservative than transversions (Appendix B, assuming that translational fidelity is high), as they are for Δ*PR* in the C*N and A*R contexts of Table 3. Alternatively, directional pressure for G/C richness would have caused errors that maintained or increased the G/C content to have been more frequently translated than errors that increased or maintained the A/U content. It is interesting that A-involved errors in the second position are among the least conservative in the G*N and A*Y contexts in Tables 3 and 4, despite adenine's being thought to have been one of the most abundant prebiotic nucleotides (Eigen and Schuster 1979).

Evidence for Minimization of the Translational Consequences of Genetic Damage in a G/C-Rich RNA World

Before the advent of an ozone layer and genomic repair, far-UV radiation must have produced constant and extensive genetic damage in the early history of life. It is thought that UV mutagenesis in RNA is similar to that in DNA because such damage does not involve the sugar–phosphate backbone. For example, a DNA photolyase can catalyze photolysis of U–U dimers in RNA. It has a lower affinity for the RNA substrate, but once bound, the enzyme has the same quantum yield of reaction as it does in DNA (Kim and Sancar 1991). In DNA, the deamination of cytosine to uracil is extremely enhanced by its UV-induced photodimerization with an adjacent pyrimidine (Friedberg et al. 1995). Of special interest in this context is the claim for primordial G/C richness, given that cytosine is especially susceptible to UV-induced damage in DNA. Therefore, the C ↔ U transition may have been especially frequent at the time of the origin of the code, both because of a higher source frequency of C and a greater transition frequency of the C ↔ U transition.

Consistent with the above reasoning is the fact that the C ↔ U transition is relatively conservative in the G*N and A*Y contexts in comparison with its substitutional effect in the C*N and A*R contexts, and also with the A ↔ G transition in these contexts.

The Consequences of Second-Position Error in Different First-Position Contexts Are Consistent with a Multiple-Stage Origin of the Genetic Code

Hierarchical cluster analysis (Figs. 1a and b) showed that the A*Y and A*R contexts were more like other contexts than each other. With the $|\Delta PR|$ measure, the A*Y and A*R contexts correlated at about 0.70 (they were the closest-correlated pair between the two clusters), despite that distances from errors involving AUA and AUG were averaged, tending to make the A*Y and A*R contexts more alike. In contrast, the C*Y and C*R and the G*R and G*Y pairs of contexts were in perfect correlation to two significant digits. This discrepancy might have been expected because three of the four sets of codons with first-position A are split and have different amino acid meaning, compared to two sets of eight in the CNN and GNN codons combined. However, the difference between A*Y and A*R is surprising given the generally conservative nature of third-position errors (Haig and Hurst 1991). Furthermore, the high correlation of the A*Y codon context with the G*N contexts was unexpected.

The amino acid assignments of the GNN and ANY codons may have evolved under distinct conditions of error from those of the ANR and CNN codons. That is, the GNN and ANY codons evolved in an era of directional pressure for high G/C content in coding regions, after which the CNN and ANR codons evolved in a less composition-biased era, wherein A ↔ G transitions were more frequent. In addition, under the evolutionary model discussed below, it is possible that variation in the coding of the GNN and ANY codons was exhausted at the time when error minimization for the disjoint set of ANN and CNR codons occurred. In this case, the meaning of the GNN and ANY codons could not have changed very easily during the subsequent fixation of meaning to the ANR and CNN codons. Error minimization could have occurred more than once in disjoint parts of the code. This hypothesis is referred to as a ''sequential model for evolution of the code by error minimization.''

The high proportion of twofold degenerate family boxes in the first-position A codons may have resulted from their distinct times of fixation of meaning. Furthermore, the discrepancy between the A*Y and the A*R contexts is explained largely by arginine in the AGR codons. If the fixation of arginine to the AGR codons was to minimize the consequences of an increasingly frequent A ↔ G transition, this could explain the excess degeneracy of arginine relative to its abundance in modern proteins (Goldberg and Wittes 1966; King and Jukes 1969) as well as its anomalous assignment to the AGR codons relative to its metabolic (Taylor and Coates 1989) and physicochemical (Tolstrup et al. 1994) relationships to other amino acids in the code.

The fact that the ANY and GNR codons are a superset of the RNY codons argued to have been primordial in the code (see, e.g., Eigen and Schuster 1979) is noted here in passing, although for reasons of space, the full debate and evidence for and against different versions of the RNY hypothesis cannot be reviewed. I note only that the currently RNY- and GNR-encoded amino acids are among the most abundant amino acids produced in the so-called Miller–Urey reactions and found in the Murchison Meteorite (reviewed by Miller 1987).

Hypotheses that the code expanded in meaning are not necessarily inconsistent with the view that all codons encoded amino acids throughout the existence of the code and that parts of the code became successively refined and distinct in amino acid meaning (Woese 1965; Fitch 1966; Crick 1968; Fitch and Upper 1987). These authors emphasize, as does Sonneborn (1965), the deleterious nature of a high frequency of noncoding codons, suggesting that very quickly, almost all codons came into use. While most codons may have quickly come into use, they need not all have been translated with the same consistency at each stage of evolution in the code. This model of code evolution could be called ''consistency expansion'' to emphasize that translational (acylation) ambiguity in different parts of the code may have been reduced at different times. In particular, translational ambiguity in the RNY and GNR codons may have been reduced first, accompanied by one round of error minimization, followed by an expansion of consistency with concomitant error minimization in the CNN and ANR codons in an era of genomic composition more like that of modern organisms.

I have not found an explanation for the pattern of amino acid assignments of the UNN codons within the current model.

The sequential model for code evolution by error minimization has implications for the use of null models in estimating the likelihood that certain code features arose by chance alone. This is because the possible variation in codes is substantially constrained if one permits only subsets of codons and amino acids to be permuted at a given time. Related to this is a possible answer to Juncgk's (1978) objection to error-minimization hypotheses, that the space of possible codes is too large to have been effectively searched by ancient populations. Finally, if there is no variation left in parts of the genetic code that have already ''frozen,'' then those parts of the code cannot be reminimized with respect to a changed

error distribution. Therefore, it should not be expected that the code minimizes modern message errors in a globally optimal way.

Error Minimization as an Evolutionary Process in Code Origin

As we have seen, there are two ways to explain some, but not all, of the results with an error-minimization hypothesis of code evolution: one invokes translational error alone; the other invokes mutation and translational error combined. While genetic damage and translational error more directly affect individuals, the necessary conditions for minimization of the effect of replication errors are not yet well defined. The stationary mutational distribution around a fit wild-type is called a molecular quasispecies (Eigen and Schuster 1979). Selection at the level of quasispecies has been demonstrated in various theoretical and numerical studies (see, e.g., Huynen et al. 1994 and references therein). The relevant argument here would be that through error minimization, the encoded catalytic potential of the entire quasispecies is *focused*—more similar to that of the wild-type. It has been shown that quasispecies with error-minimizing codes are competitively superior (Figureau 1989).

The observation of near-universality of genetic codes supports the principle that they must evolve very slowly. The slow rate of evolutionary change in the code, first discussed by Crick (1968) as a component of the frozen accident hypothesis, should be reconciled with error-minimization hypotheses for its origin. In general, to change a code without disruption of the meaning of preexisting messages requires either than the change be conservative with respect to meaning (Crick 1968) or that the symbols, the sense of which will be altered by the change, are infrequent in messages. A conservative change in a code is measured by the relative neutrality of its concomitant substitutional effect with respect to overall message meaning. A disruption in message meaning may also be offset by the advantage of an expanded symbol set (Wong 1980). In genomes, low codon frequencies may arise through a skewed base content (Szathmary 1991; Osawa et al. 1992; Maynard-Smith and Szathmary 1995). Generally, once reliance upon certain symbols increases, alterations of their sense will be deleterious.

Contrary to what Crick (1968) and Woese (1973) have written, the above considerations do not preclude the action of natural selection on amino acid assignments, even in frequently used codons the precise meaning of which are essential to fitness. The near-frozenness of code assignments impedes *phenetic change* of assignments, that is, the reassignment of essential codons *within a lineage*. But suppose the standard code evolved through successive stages of expansion (Eigen and Schuster 1979), ambiguity reduction (Woese 1965; Fitch 1966; Crick 1968), metabolic coevolution (Wong 1975; Taylor and Coates 1989), amino acid intrusion (Crick 1968; Jukes 1973; Wong 1980), or some combination, such as consistency expansion. If certain codons are underused or neutrally ambiguous, their meanings are free to be reassigned or refined, and potentially differently in different lineages. As these codons become more frequently used and the precision of their meaning more essential in variant lineages, the lineages might no longer share genes that use these codons without sharing also the means to decode them properly—a barrier to gene flow [but see Crick (1968) for a discussion of an alternative possibility of fusion of individuals with different codes]. Initially neutral variation in codes may have contributed to the differential success of these lineages, enabling a *phyletic* process of error minimization in the code.

A successive process of radiation and bottleneck events may have occurred in code origin, during which newly defined parts of the code froze in epistasis with useful messages, but differently in competing lineages. A so-called ‘‘sequential’’ model of code evolution with error minimization does not predict that the code reached a global optimum with respect to modern message errors, because each successive stage may have been characterized by different distributions of error. Only one bottleneck is necessary to explain the near-universality of the code in modern organisms and organelles.

Conclusions

Detailed evidence is now available for the error-minimization theory of standard code evolution. Contrary to prior interpretations of error minimization, the standard code may have been shaped by position-invariant forces such as mutation and base content. These forces may have left heterogeneous signatures in the code because of differences in translational fidelity by codon position.

A scenario for the origin of the code is presented wherein selection for error minimization could have occurred multiple times in disjoint parts of the code through a phyletic process of competition between lineages. This process permits error minimization without the disruption of previously useful messages, and does not predict that the code is optimally error-minimizing with respect to modern error. Finally, through a scheme of heterogeneous ambiguity reduction (‘‘consistency expansion’’), an expansion of code meaning need not imply a concomitant reduction in stop codon number.

Corollary observations in this paper are testable, such as the promotion of first-position translational error by a second-position pyrimidine context. Statistical hypothesis testing of code features should account for differences that may arise by the amino acid distance and the type of null model used.

By virtue of its evolutionary inertia as well as its likelihood of having been an object of natural selection, the standard code may be a record of evolution before the radiation of modern organisms and organelles.

Acknowledgments. Marcus Feldman and Michael Lachmann made numerous suggestions on style and substance during the preparation of the manuscript. Jonathan Eisen and Philip C. Hanawalt made helpful suggestions regarding the influence of mutagenesis. J. Eisen and Jonathan Pritchard first suggested the use of PAM matrices. Charles Yanofsky, Carl Bergstrom, Aviv Bergman, Mark Tanaka, Lauren Ancel, Sarah Otto, Eric Ekland, and David Zapol participated in helpful discussions. The comments of anonymous reviewers improved the review of prior literature. This research was supported in part by NIH Grants GM28016 and GM28428 to M.W. Feldman. This is Contribution No. 3 from the Center for Computational Genetics and Biological Modeling, Stanford University.

References

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

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

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

Crick F (1968) The origin of the genetic code. J Mol Biol 38:367–379

Davies J (1966) Streptomycin and the genetic code. Cold Spring Harbor Symp Quant Biol 31:670

Davies J, Gilbert W, Gorini L (1964) Streptomycin, suppression and the code. Proc Natl Acad Sci USA 51:883–890

Davies J, Jones D, Khorana H (1966) A further study of misreading of codons induced by streptomycin and neomycin using ribopolynucleotides containing two nucleotides in alternating sequence as templates. J Mol Biol 18:48–57

Di Giulio M (1989a) 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 (1989b) Some aspects of the organization and evolution of the genetic code. J Mol Evol 29:191–201

Di Giulio 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 (1995a) The phylogeny of tRNAs seems to confirm the predictions of the coevolution theory of the origin of the genetic code. Orig Life Evol Biosph 25:549–564

Di Giulio M (1995b) Was it an ancient gene codifying for a hairpin RNA that, by means of direct duplication, gave rise to the primitive tRNA molecule? J Theor Biol 177:95–101

Eigen M, Schuster P (1979) The hypercycle: a principle of natural self-organization. Springer, Berlin

Eigen M, Winkler-Oswatitsch R (1981) Transfer-RNA: the early adaptor. Naturwissenschaften 68:217–228

Ekland E, Bartel DP (1996) RNA-catalyzed RNA polymerization using nucleoside triphosphates. Nature 382(6589):373–376

Epstein C (1966) Role of the amino-acid ''code'' and of selection for conformation in the evolution of proteins. Nature 210(5031):25–28

Figureau A (1989) Optimization and the genetic code. Orig Life Evol Biosph 19:57–67

Fitch W (1966) Evidence suggesting a partial, internal duplication in the ancestral gene for heme-containing globins. J Mol Biol 16:1

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

French S, Robson B (1983) What is a conservative substitution? J Mol Evol 19:171–175

Friedberg E, Walker G, Siede W (1995) DNA repair and mutagenesis. ASM Press, Washington, DC

Goldberg AL, Wittes R (1966) Genetic code: aspects of organization. Science 153:420–424

Goldman N (1993) Further results on error minimization in the genetic code. J Mol Evol 37:662–664

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

Grosjean H, de Henau S, Crothers D (1978) On the physical basis for ambiguity in genetic coding interactions. Proc Natl Acad Sci USA 75(2):610–614

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

Huynen M, Hogeweg P (1994) Pattern generation in molecular evolution: exploitation of the variation in RNA landscapes. J Mol Evol 39:71–79

Jukes T (1973) Arginine as an evolutionary intruder into protein synthesis. Biochem Biophys Res Commun 53(3):709–714

Juncgk J (1978) The genetic code as a periodic table. J Mol Evol 11:211–224

Kim S-T, Sancar A (1991) Effect of base, pentose, and phosphodiester backbone structures on binding and repair of pyrimidine dimers by *Escherichia coli* dna photolyase. Biochemistry 30(35):8623–8630

Kimura M (1980) A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. J Mol Evol 16:111–120

King J, Jukes T (1969) Non-darwinian evolution. Science 164:788–798

Koshi J, Goldstein R (1997) Mutation matrices and physical-chemical properties: correlations and implications. Proteins 27:336–344

Kuge S, Kawamura N, Nomoto A (1989) Strong inclination toward transitions in nucleotide substitutions by poliovirus replicase. J Mol Biol 207(1):175–182

Maynard-Smith J, Szathmary E (1995) The major evolutionary transitions in evolution. W.H. Freeman, Oxford

Miller S (1987) What organic compounds could have occurred on the prebiotic earth? Cold Spring Harbor Symp Quant Biol 52:17–27

Nakai K, Kidera A, Kanehisa M (1988) Cluster analysis of amino acid indices for prediction of protein structure and function. Protein Eng 2:93–100

Negre D, Cenatiempo Y, Cozzone A (1988) Differential pattern of misreading induced by streptomycin *in vitro*. J Mol Biol 204:213–216

Nirenberg M, Jones O, Leder P, Clark B, Sly W, Pestka S (1963) On the coding of genetic information. Cold Spring Harbor Symp Quant Biol 28:549–558

Nishimura S, Harada F, Hirabayashi M (1968) Nature of magnesium-induced miscoding: the *in vitro* synthesis of valine-tyrosine copolypeptide directed by poly-(U-G). J Mol Biol 40:173–186

Osawa S, Jukes T, Watanabe K, Muto A (1992) Recent evidence for evolution of the genetic code. Microbiol Rev 56(1):229–264

Parker J (1989) Errors and alternatives in reading the universal genetic code. Microbiol Rev 53(3):273–298

Rodin S, Rodin A, Ohno S (1996) The presence of codon-anticodon pairs in the acceptor stem of tRNAs. Proc Natl Acad Sci USA 93:4537–4542

Sitaramam V (1989) Genetic code preferentially conserves long-range interactions among the amino acids. FEBS Lett 247(1):46–50

Sjöström M, Wold S (1985) Genetic code preferentially conserves long-range interactions among the amino acids. FEBS Lett 247(1): 46–50

Sonneborn T (1965) Degeneracy of the genetic code: extent, nature, and genetic implications. In: Bryson V, Vogel H (eds.) Evolving genes and proteins. Academic Press, New York, pp 377–397

Strigini P, Brickman E (1973) Analysis of specific misreading in *Escherichia coli*. J Mol Biol 75:659–672

Swanson R (1984) A unifying concept for the amino acid code. Bull Math Biol 46(2):187–203

Szathmary E (1991) Codon swapping as a possible evolutionary mechanism. J Mol Evol 32:178–182
Szathmary E, Zintzaras E (1992) A statistical test of hypotheses on the organization and origin of the genetic code. J Mol Evol 35:185–189
Taylor F, Coates D (1989) The code within the codons. BioSystems 22:177–187
Tolstrup N, Toftgård J, Engelbrecht J, Brunak S (1994) Neural network model of the genetic code is strongly correlated to the GES scale of amino acid transfer free energies. J Mol Biol 243:816–820
Tomii K, Kanehisa M (1996) Analysis of amino-acid indexes and mutation matrices for sequence comparison and structure prediction of proteins. Protein Eng 9(1):27–36
Turner DH (1996) Thermodynamics of base pairing. Curr Opp Struct Biol 6(3):299–304
Turner DH, Sugimoto N, Freier SM (1988) RNA structure prediction. Annu Rev Biophys Biophys Chem 17:167–192
Woese C (1965) On the evolution of the genetic code. Proc Natl Acad Sci USA 54:1546–1552
Woese C (1973) Evolution of the genetic code. Naturwissenschaften 60:447–459
Woese C, Dugre D, Dugre S, Kondo M, Saxinger W (1966) On the fundamental nature and evolution of the genetic code. Cold Spring Harbor Symp Quant Biol 31:723–736
Wolfenden R, Cullis P, Southgate C (1979) Water, protein folding and the genetic code. Science 206(2):575–577
Wong J (1975) A co-evolution theory of the genetic code. Proc Natl Acad Sci USA 72(5):1909–1912
Wong J (1980) Role of minimization of chemical distances between amino acids in the evolution of the genetic code. Proc Natl Acad Sci USA 77(2):1083–1086
Wu M, McDowell J, Turner DH (1995) A periodic table of symmetric tandem mismatches in RNA. Biochemistry 34(10):3204–3211

Appendix A

A Formalization of the Error-Minimization Principle for Standard Genetic Code Evolution

The error-minimization principle for genetic code evolution is an optimality argument. The quantity to be optimized, which might be called ''load'' (Maynard-Smith and Szathmary 1995), incorporates several factors:

1. a set A of amino acids, with a real-valued ''distance'' $d\colon A \times A \rightarrow \Re+$ between them [the distance need not be transitive nor symmetric, but $d(x, y) \geqslant 0$ with equality if $x = y$];
2. a set Y of *codons* of length n;
3. a family of *codes*, $\mathbf{C} = \{c \in C \mid c\colon Y \rightarrow A\}$, deterministic functions taking codons $y \in Y$ to amino acids $a \in A$; and
4. a *transmission distribution* $P(S, T)$, $S,T \in Y$, giving the joint probability of a *source or input codon* S and a *target or output codon* T after transmission through some *channel.*

The weighted average load is a functional $\mathbf{L}\colon\mathbf{C} \rightarrow \Re+$ and may be defined as

$$\mathbf{L}(c) = \sum_{y_{\text{in}},y_{\text{out}} \in Y} p(y_{\text{in}})p(y_{\text{out}} \mid y_{\text{in}})d[c(y_{\text{in}}),c(y_{\text{out}})] \tag{2}$$

Optimization of (2) is specified by

$$\min_{\mathbf{C}}\{L(c),c \in \mathbf{C}\} \tag{3}$$

In (2), the transmission distribution is decomposed into a *source distribution* of input, $p(\cdot)$, and a *transition distribution* conditional on that input, $p(\cdot \mid \cdot)$. The source distribution corresponds to the translational frequency of each codon and the transition distribution gives the error spectrum associated with each codon through mutation and translation. The load depends on codon frequencies and, therefore, indirectly considers the effects of base content.

Note that in (2), the load does not depend on the absolute placement of amino acids in the code table, since for deterministic codes, $d[c(y_{\text{in}}), c(y_{\text{out}})] = 0$ when $y_{\text{in}} = y_{\text{out}}$. For nondeterministic codes, for example, if there is misacylation, the weighted average load could depend on the absolute position of amino acids in the code table.

Suppose that $|A|$ amino acids are encoded by $|Y|$ codons at a level of degeneracy fixed for each amino acid.[6] In this case, Eq. (2) is minimized by pairing, through changes in c, the largest joint probabilities of y_{in}, and y_{out} with the smallest distances $d[c(y_{\text{in}}],c(y_{\text{out}})]$. A proof of this statement follows.

For two real-valued k-vectors α and β and the set of $k \times k$ permutation matrices $\mathbf{P}$, one can show that the inner product $(\alpha \cdot P\beta)$, $P \in \mathbf{P}$ is minimized over $\mathbf{P}$ when α and $P\beta$ satisfy the condition

$$(\forall i,j\colon 1 \leqslant i,j \leqslant k), \qquad \alpha_i \geqslant \alpha_j \Rightarrow (P\beta)_j \geqslant (P\beta)_i \tag{4}$$

where the ith entry of vector x is x_i.[7] Condition (4) is called ''pseudo-orthogonalization.'' The proof is by induction.

Case $k = 2$. Define two real-valued vectors $\alpha = \langle\alpha_1,\alpha_2\rangle$ and $\beta = \langle\beta_1,\beta_2\rangle$ with $\alpha_2 > \alpha_1$ and $\beta_2 > \beta_1$. Let I be the 2×2 identity matrix and R be the permutation matrix $\left(\begin{smallmatrix}0 & 1\\ 1 & 0\end{smallmatrix}\right)$. Then $\alpha \cdot I\beta - \alpha \cdot R\beta = (\alpha_1 - \alpha_2)(\beta_1 - \beta_2) > 0$. Since these are the only permutations of a 2-vector, pseudo-orthogonalization ($P = R$) minimizes $(\alpha \cdot P\beta)$ in this case.

Case $k = n$. Assume that (4) is true for case $k = (n - 1)$. Let α and $R\beta$ be n-vectors pseudo-orthogonalized by a permutation matrix R. R will be shown to minimize each of several inner products of $(n - 1)$-vectors whose sum is equal to $(\alpha \cdot R\beta)$. R therefore minimizes $(\alpha \cdot P\beta)$ over all $\mathbf{P}$.

Let J_i be the $(n - 1) \times n$ projection matrix equal to the $n \times n$ identity matrix with the ith row deleted, $1 \leqslant i \leqslant n$. Since

$$n(\alpha \cdot R\beta) = \sum_{i=1}^{n} [(J_i\alpha \cdot J_iR\beta) + \alpha_i(R\beta)_i] \tag{5}$$

an inner product of n-vectors may be reduced to a sum of inner products of $(n - 1)$-vectors:

$$\alpha \cdot R\beta = \frac{1}{n-1}\sum_{i=1}^{n} (J_i\alpha \cdot J_iR\beta) \tag{6}$$

Because the projection matrices J_i preserve the order of entries in α and $R\beta$,

$$(\forall r,s\colon 1 \leqslant r,s \leqslant (n-1), \forall i\colon 1 \leqslant i \leqslant n),\\ (J_i\alpha)_r \geqslant (J_i\alpha)_s \Rightarrow (J_iR\beta)_s \geqslant (J_iR\beta)_r \tag{7}$$

[6] Allowing the degeneracy to vary permits trivial solutions; the present theory cannot address the evolution of amino acid degeneracy or the size of $|A|$.

[7] It is claimed but not shown that a permutation matrix P satisfying (4) is unique up to equality of entries in β.

That is, each term on the right-hand side of (6) is pseudo-orthogonalized by the pseudo-orthogonalization of α and $R\beta$. By n applications of the inductive hypothesis, $(\alpha \cdot P\beta)$ is minimized by $P = R$.

This has been claimed by numerous authors in the past (Woese 1965; Sitaramam 1989; Haig and Hurst 1991) but never formally demonstrated in this context.

Appendix B

Mutation Affects Code Evolution Through the Filter of Translational Error

In this Appendix, I derive a simple representation of a *transition distribution* (the conditional probability of a source codon in a coding region being translated as another codon; see Appendix A) that incorporates both mutation and translational error. This heuristic argument has three simplifications. First, the transition distribution is derived per codon position and, therefore, neglects context effects. Second, the frequencies of translational error are symmetric with respect to the different bases. These simplifications are irrelevant to the argument so long as the net effect of all influences on translation is such that, on average, one position is translated with higher Watson–Crick fidelity than another. The third simplification is more restrictive. Mutation is represented as occurring within an individual. This treatment may describe genetic damage better than misreplication.

Suppose that ancient mutation were transition-biased. The Kimura two-parameter mutation model (Kimura 1980) is expressed in terms of the matrix $\mathbf{M} = \|M_{ij}\|$ of probabilities of transmission of the source base i as base j:

$$\mathbf{M}(\alpha,\beta) = \begin{pmatrix} 1-\alpha-2\beta & \alpha & \beta & \beta \\ \alpha & 1-\alpha-2\beta & \beta & \beta \\ \beta & \beta & 1-\alpha-2\beta & \alpha \\ \beta & \beta & \alpha & 1-\alpha-2\beta \end{pmatrix} \tag{8}$$

where α and β are the transition and transversion probabilities, respectively ($\alpha > \beta$).

Now suppose for the sake of argument that translational error in any codon position is symmetric, i.e., provided mistranslation occurs, the source base is equally likely to be translated as any of the three other bases. Symmetry is not strictly necessary for the argument to proceed, but it simplifies the exposition by requiring only one parameter to describe translational error. A matrix representation for symmetrical translation in terms of the translational fidelity per position p is

$$\mathbf{T}(p) = \begin{pmatrix} p & (1-p)/3 & (1-p)/3 & (1-p)/3 \\ (1-p)/3 & p & (1-p)/3 & (1-p)/3 \\ (1-p)/3 & (1-p)/3 & p & (1-p)/3 \\ (1-p)/3 & (1-p)/3 & (1-p)/3 & p \end{pmatrix} \tag{9}$$

where $T_{i,j}(\cdot)$ is the probability of the source base i being translated as base j.

Now consider the transmission of a source base through one round of mutation followed by one round of translation at a given codon position. Both processes are error-prone and considered to occur independently. The matrix representing the transmission of a base through such a process in a given codon position is the product of matrices **M** and **T:**

$$\mathbf{MT}(\alpha,\beta,p) = \begin{pmatrix} \gamma p + (1-\gamma)q & \alpha p + (1-\alpha)q & \beta p + (1-\beta)q & \beta p + (1-\beta)q \\ \alpha p + (1-\alpha)q & \gamma p + (1-\gamma)q & \beta p + (1-\beta)q & \beta p + (1-\beta)q \\ \beta p + (1-\beta)q & \beta p + (1-\beta)q & \gamma p + (1-\gamma)q & \alpha p + (1-\alpha)q \\ \beta p + (1-\beta)q & \beta p + (1-\beta)q & \alpha p + (1-\alpha)q & \gamma p + (1-\gamma)q \end{pmatrix} \tag{10}$$

where $\gamma = 1 - \alpha - 2\beta$, and $q = [(1 - p)/3]$.

From Eq. (10) we may see how translational fidelity exposes mutational error. At maximum translational noise, as $p \to 0.25$, all dependency on α and β is lost, and all entries of matrices (9) and (10) reduce to ¼. The exact result will depend on the form of the translational error matrix in (9). However, as $p \to 1$, matrix (10) reduces to matrix (8), the matrix for mutation. This is less sensitive to the exact form in (9).