Early Fixation of an Optimal Genetic Code

Stephen J. Freeland, Robin D. Knight,* Laura F. Landweber,* and Laurence D. Hurst†*

*Department of Ecology and Evolution, Princeton University; and †Department of Biology and Biochemistry, University of Bath, Bath, England

The evolutionary forces that produced the canonical genetic code before the last universal ancestor remain obscure. One hypothesis is that the arrangement of amino acid/codon assignments results from selection to minimize the effects of errors (e.g., mistranslation and mutation) on resulting proteins. If amino acid similarity is measured as polarity, the canonical code does indeed outperform most theoretical alternatives. However, this finding does not hold for other amino acid properties, ignores plausible restrictions on possible code structure, and does not address the naturally occurring nonstandard genetic codes. Finally, other analyses have shown that significantly better code structures are possible. Here, we show that if theoretically possible code structures are limited to reflect plausible biological constraints, and amino acid similarity is quantified using empirical data of substitution frequencies, the canonical code is at or very close to a global optimum for error minimization across plausible parameter space. This result is robust to variation in the methods and assumptions of the analysis. Although significantly better codes do exist under some assumptions, they are extremely rare and thus consistent with reports of an adaptive code: previous analyses which suggest otherwise derive from a misleading metric. However, all extant, naturally occurring, secondarily derived, nonstandard genetic codes do appear less adaptive. The arrangement of amino acid assignments to the codons of the standard genetic code appears to be a direct product of natural selection for a system that minimizes the phenotypic impact of genetic error. Potential criticisms of previous analyses appear to be without substance. That known variants of the standard genetic code appear less adaptive suggests that different evolutionary factors predominated before and after fixation of the canonical code. While the evidence for an adaptive code is clear, the process by which the code achieved this optimization requires further attention.

Introduction

All known nonstandard genetic codes appear to be secondarily derived minor modifications of the canonical code (Osawa 1995). The fact that codon reassignment is not always lethal indicates that the amino acid/codon assignments of the canonical code need not be a ''frozen accident'' of history (Crick 1965), but, rather, require explanation. One hypothesis is that the arrangement of amino acid assignments results from natural selection among different codes favoring those that minimize the phenotypic impact of genetic error by maximizing the similarity of amino acids assigned to codons differing by only a single nucleotide (Sonneborn 1965; Woese 1965; Zuckerkandl and Pauling 1965).

Previous evidence for an adaptive code structure derives as follows. The canonical code's susceptibility to error is quantified as a ''code error value,'' Δ_{code}, representing the average change in amino acid meaning (according to some quantitative similarity measure) resulting from all single-nucleotide substitutions in all codons. This is then compared with equivalent values measured for theoretical alternative codes. When amino acid similarity is measured in terms of Polar Requirement (Woese et al. 1966) (essentially a measure of hydrophobicity), the canonical code outperforms all but one in 10,000 randomly generated alternatives (Haig and Hurst 1991; Freeland and Hurst 1998*a*) and appears to be better still when calculation of Δ_{code} is adjusted to incorporate known biases in mutation and mistranslation rates (Ardell 1998; Freeland and Hurst 1998*a*).

Adaptive Evidence as an Artifact of Stereochemistry?

However, this evidence is potentially flawed in several ways. First, when other measures of amino acid similarity, such as volume or charge, are used, the canonical code no longer appears special (Haig and Hurst 1991). Without clear evidence that amino acid polarity is the major defining factor of protein fitness, claims for an adaptive canonical code might be spurious. In particular, the ''stereochemical'' hypothesis proposes that the canonical code originated through specific steric interactions between amino acids and their associated codons (Yarus 1998; Knight, Freeland, and Landweber 1999). Because Polar Requirement values derive from chromatographic partitioning of amino acids in a water/pyridine system (Woese et al. 1966), amino acids that bind nucleotides similarly but act differently in proteins could be responsible for Polar Requirement values, such that a code formed through stereochemical interactions might appear adaptive as an artifact. We address this weakness by employing point accepted mutations (PAM) 74–100 matrix data (Benner, Cohen, and Gonnet 1994), which are derived from the pattern of amino acid substitution frequencies observed within naturally occurring pairs of homologous proteins and thus provide a direct measure of amino acid similarity in terms of protein biochemistry.

Potential Problems with PAM

A subtle and potentially profound problem with using PAM matrix data in this context is that PAM matrix values may simply reflect the structure of the genetic code. Specifically, over a short evolutionary period,

Key words: genetic code, adaptation, evolution, PAM matrix, Polar Requirement.

Address for correspondence and reprints: Stephen J. Freeland, Department of Ecology and Evolution, Princeton University, Princeton, New Jersey 08544. E-mail: freeland@rnaworld.princeton.edu.

Mol. Biol. Evol. 17(4):511–518. 2000

most amino acid substitutions are likely to result from single point mutations within codons: PAM scores for minimally diverged proteins are therefore likely to reflect the arrangement of codon assignments rather than amino acid similarity in terms of protein chemistry, rendering code analysis based on PAM scores a tautologous exercise. The PAM 74–100 uniquely avoids this problem, since it is derived from only highly diverged protein sequences (other PAM matrices are calculated from homologs of low divergence and manipulated mathematically to predict substitution patterns at high divergence). Quantitative analysis of the PAM 74–100 demonstrates its superior measurement of amino acid similarity. PAM matrices derived for homologs of increasing divergence do differ from one another, consistent with the idea that they decreasingly reflect code structure and increasingly reflect amino acid similarity (in terms of protein biochemistry). That PAMs of increasing divergence approach an asymptote below the divergence threshold used to choose proteins for the calculation of the 74–100 indicates that this matrix predominantly reflects amino acid biochemical similarity (i.e., higher divergence would provide no further surprises). Further qualitative observations support this interpretation. For example, within the standard code, Tryptophan is assigned a single codon semisurrounded by termination codons and can only change to arginine by means of a single nucleotide transition error (UGG→CGR). However, tryptophan's side chain is a hydrophobic ring, whereas arginine's side chain is aliphatic and hydrophilic. It is intuitive that tryptophan's protein biochemistry would be more similar to that of, say, phenylalanine (with an aromatic, hydrophobic side chain), which lies two point mutations away in codon space. Low-divergence PAMs imply a greater similarity between Trp and Arg than between Trp and Phe, whereas the PAM 74–100 indicates the opposite relationship, consistent with the idea that the PAM 74–100 accurately records the biochemical similarity of amino acids rather than their assignments within the code. The overall high correlation between the PAM 74–100 and PAMs of low divergence (or theoretical PAM's derived purely from code structure) does not indicate the matrix's dependence on code structure, but can equally well be explained by an adaptive code in which similar codons are assigned to amino acids with similar biochemical properties: it is only the differences, outlined above, that inform the PAM 74–100's suitability for code analysis.

On a different note, it may be argued that because the PAM 74–100 measures amino acid physiochemical similarity in a very general sense (i.e., the degree to which different side chains can operate in similar structure/function roles), we cannot completely rule out the possibility that matrix values would apply to nucleotide-binding affinities of the different amino acids. A stereochemical interpretation of PAM-based analysis thus remains possible, although at present no evidence exists to support this argument; quite simply, we know that the PAM 74–100 is an accurate measure of interchangeability within proteins; the correlation with nucleotide-binding affinities remains at best speculative.

Adaptive Evidence as an Artifact of Biosynthetic Relatedness?

A second weakness in previous adaptive analyses (Wong 1980; DiGiulio 1989, 1994; Haig and Hurst 1991; Goldman 1993; Ardell 1998; Freeland and Hurst 1998*a*) is that most have assumed that each synonymous codon block of the canonical code could have taken any amino acid assignment (fig. 1*a*). A growing body of circumstantial evidence questions this assumption (Knight, Freeland, and Landweber 1999). Many of the 20 canonical amino acids are not plausible products of prebiotic chemistry (Wong and Bronskill 1979) and are only produced in extant organisms as biosynthetic modifications of their plausibly primordial counterparts. Furthermore, biosynthetically related amino acids are often assigned to similar codons within the canonical code (Wong 1975; Taylor and Coates 1989). Taken together, these observations have provoked the code coevolution hypothesis (Wong 1975), proposing that the canonical code evolved from a simpler ancestral form (encoding fewer amino acids with greater redundancy) by successively reassigning subsets of synonymous codons to incorporate novel amino acid biosynthetic derivatives. Although detailed perceived patterns (Wong 1975) are untrustworthy because of the biosynthetic interrelatedness of most amino acids within present-day metabolism (Amirnovin 1997), it does appear that amino acids from the same biosynthetic pathway are generally assigned to codons sharing the same first base (Taylor and Coates 1989) (fig. 1*b*). If this reflects a history of biosynthetic expansion from some primordial code, then the implied restrictions on code evolution would reduce the number of possible codes so greatly as to render previous adaptive results meaningless (Freeland and Hurst 1998*b*). We investigate this possibility by constructing a set of possible codes that allows interchange of amino acids only within each biochemical pathway (fig. 1*c*).

Other Analyses Indicate a Nonadaptive Code

Third, while previous analyses suggest that the canonical code outperforms most alternatives, a comprehensive search of possible code structures suggests that far better alternatives are possible (Wong 1980; DiGiulio 1989, 1994; Goldman 1993). Indeed, the canonical code achieves between 45.3% (Wong 1980) and 78% (DiGiulio 1994) of the possible error minimization, depending on precise assumptions. Unfortunately, this alternative method of estimating code optimality, cited as evidence against an adaptive canonical code, has coincided with further methodological differences which could either explain or obscure the qualitatively different results. In particular, previous reports of a highly optimized code structure have measured the average squared difference in amino acid Polar Requirement, whereas those reporting a less adaptive code structure have used the modular difference. One possible explanation of qualitatively different results is that the adaptive arrangement of just a few key amino acids with extreme Polar Requirement values could account for most apparent optimization of the code: squaring the Polar Re-

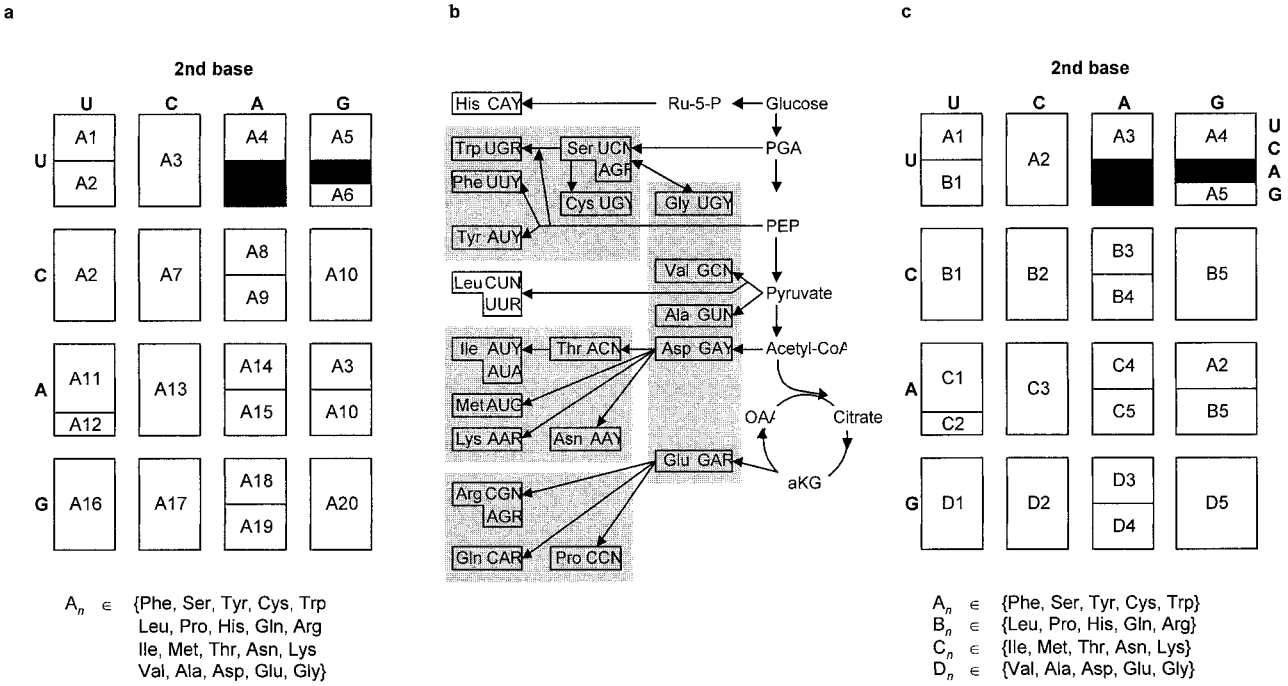

FIG. 1.—Definitions of the unrestricted and restricted sets of theoretically possible codes. Both sets maintain the pattern of synonymous codon blocks and the assignments of termination codons found in the canonical code (implying a fixed relationship between codons and tRNA anticodons). *a*, The unrestricted set; the 20 amino acids are randomly assigned to the 20 synonymous codon blocks with no further mapping restrictions. *b*, The correlation between amino acid biosynthetic pathways and amino acid assignments within the code reported by Taylor and Coates (1989). *c*, The restricted set; codon assignments are divided into four groups (A–D), each containing five members (1–5). Each group takes the same five assignments as in the canonical code, although assignments within a group are allowed to vary randomly. While the best treatment of Leu, His, and the extra codon pairs assigned to Ser and Arg remains unclear, our results show minimal change when these codon assignments are allowed to vary freely.

quirement differences between amino acids would exaggerate the importance of the arrangement of outlier amino acids within the code, providing an exaggerated estimate of code optimality. We compare estimates of code optimality according to both interpretations of optimality over a range of scaling values and transition/transversion biases.

What of Nonstandard Genetic Codes?

Finally, previous analyses have concentrated exclusively on the canonical genetic code, ignoring secondarily derived nonstandard variants. We present a comparative analysis of all known nonstandard codes to investigate their relationship to the adaptive hypothesis.

Materials and Methods

The process of testing the canonical code for evidence of adaptation in terms of error minimization may be divided into three stages: (1) summarizing a particular code's susceptibility to genetic error as a code error value, (2) defining a set of theoretically possible codes from which the canonical code evolved, and (3) measuring the optimality of the canonical code by comparing its error value with those of randomly selected plausible alternatives.

Calculating a Code's Error Value

A particular code's error value, Δ_i, is the weighted mean of all possible changes in codon meaning resulting from all possible single-nucleotide changes in all codons. With errors subdivided into transitions and transversions, Δ_i is calculated as

$$\Delta_i = \frac{\sum_{i=1}^{210} (w\alpha_i + \beta_i)\epsilon_i}{\sum_{i=1}^{210} (w\alpha_i + \beta_i)}, \quad (1)$$

where ϵ_i is the error magnitude associated with amino acid substitution i (see below), w represents the transition/transversion bias weighting under consideration, and α_i and β_i represent the total numbers of times error ϵ_i occurs within a particular code as the result of single-nucleotide transitions and transversions, respectively. For example, three single nucleotide changes cause substitution of Ile for Met in the canonical code, the transition AUG→AUA, and the two transversions AUG→AUY; thus, $\alpha_{\text{Met}\to\text{Ile}} = 1$, whereas $\beta_{\text{Met}\to\text{Ile}} = 2$. Many amino acid substitutions will not occur as the result of any single-nucleotide substitutions for any given code; e.g., $\alpha_{\text{Met}\to\text{Gly}} = \beta_{\text{Met}\to\text{Gly}} = 0$ for the canonical code. The summation terms reflect that 210 different errors are possible for an amino acid alphabet of 20

members, given error symmetry (i.e., the error caused by substitution of amino acid a_1 for amino acid a_2 is equivalent to that caused by the substitution of a_2 for a_1). Although abundant empirical data indicate that transitions occur more frequently than transversions (i.e., $w > 1$), an exact value is not known for primordial evolution. We therefore test a range of w values.

Calculation of individual codon error (ϵ_i) values varies according to the measure of amino acid similarity used. For data from the PAM 74–100 matrix, we use the transformation

$$\epsilon = \begin{cases} (10^{-(M_{a1,M_{a2}}/10)})^p & a_1 \neq a_2 \\ 0 & a_1 = a_2, \end{cases} \quad (2)$$

where ϵ_i is the error value associated with a change from amino acid a_1 to amino acid a_2 (the ''correct'' and ''incorrect'' amino acids) and $M_{a1,a2}$ is the PAM matrix score associated with amino acids a_1 and $a_2 M_{a1,a2} = M_{a2,a1}$ because of error symmetry), and p is the power to which the modular difference in amino acid property is raised (''modular power function'').

For comparison with previous analyses, we also use Polar Requirement (Woese et al. 1966) data, a measure of hydrophobicity, for which the codon error value is calculated as

$$\epsilon = |A_1 - A_2|^p, \quad (3)$$

where A_1 and A_2 are the values for the ''correct'' and ''incorrect'' amino acid meanings.

Neither similarity measure defines values for Ter codons, which are ignored. Since the most biologically realistic value of the scaling function p is unknown, we test robustness over differences in this arbitrarily chosen parameter ($1 \leq p \leq 5$).

Defining the Set of Possible Codes

The level of error minimization achieved by the canonical genetic code is found by assessing the position of its error value (Δ_{code}: eq. 1) relative to those of possible alternatives: this entails definition of the set of possible codes. Minimal requirements are that all variants comprise 64 codons, with assignments divided between the 20 amino acids and the translation termination signal. We further retain both the pattern of redundancy and the position of ''stop'' codons found in the canonical genetic code. Although redundancy patterns vary in secondarily derived nonstandard codes, their relevance to the evolution of the canonical code is unclear. For example, the precanonical codes probably utilized a minimal set of tRNAs, usually one per amino acid, whereas extant code variation has involved further tRNA diversification (Osawa 1995). Furthermore, it is plausible that the canonical code reached its present form through incorporation of novel amino acids via subdivision and reassignment of synonymous codon blocks (Dillon 1973; Wong 1975). Under any of these models, our assumptions represent possible code variation with good accuracy. More importantly, if our restrictions underestimate possible code variation, then they actually bias our analysis against the adaptive hypothesis: most variation in redundancy (especially individual codon reassignments) would reduce the remarkable symmetry of the canonical code, leading to an increased error value. For example, a code in which individual codons are randomly reassigned produces a sample of one million variant codes with a mean error value (Δ_{mean}) between 23% ($p = 1$, $w = 1$) and 33% ($p = 1$, $w = 5$) higher than the mean of a sample produced by our method. Given these considerations, in asking whether the arrangement of amino acid assignments within the canonical code is adaptive, retaining the synonymous codon block structure for all variants represents the best compromise.

Our rules permit $20! \approx 2.43 \times 10^{18}$ different codes (fig. 1*a*), which we refer to as the ''unrestricted'' set. The incorporation of further restrictions to reflect the observation that biosynthetically related amino acids are often assigned codons with the same first base identity (fig. 1*b*) produces a ''restricted'' set of $(5!)^4 \approx 2 \times 10^9$ codes (fig. 1*c*).

Measuring the Optimality of the Natural Genetic Code

Our analysis considers an optimal code as one in which the arrangement of amino acid assignments to synonymous codon blocks minimizes the average amino acid difference resulting from single-nucleotide changes within all codons. Previous analyses of code structure have interpreted this criterion, estimating optimality of the canonical code relative to that of plausible theoretical alternatives in one of two different ways: (1) analyses indicating a highly optimized code have used a statistical measure of efficiency, generating a sample of possible codes and calculating the proportion that have a lower error value than the canonical code (Haig and Hurst 1991; Ardell 1998; Freeland and Hurst 1998*a*, 1998*b*), and (2) those indicating a less optimized code have used an engineering approach, measuring Δ_{code} relative to the lowest code error value (Δ_{opt}) and the mean error value (Δ_{mean}) of all possible codes (Wong 1980; DiGiulio 1989, 1994; Goldman 1993; DiGiulio and Medugno 1999):

$$\%\ optimization = \frac{\Delta_{mean} - \Delta_{code}}{\Delta_{mean} - \Delta_{opt}} \times 100\%. \quad (4)$$

Identifying Δ_{opt} within ''possible code space'' under any particular set of assumptions has previously been approached analytically (Wong 1980; DiGiulio 1989) and by heuristic computer search algorithms (Goldman 1993; DiGiulio 1994: DiGiulio and Medugno 1999). The latter outperform the former (DiGiulio 1994), and the most comprehensive search to date used simulated annealing (SAN) (Goldman 1993). We use a powerful alternative known as ''the Great Deluge algorithm'' (GDA) (Dueck 1992). The GDA outperforms SAN in classical optimization problems, requiring fewer operational parameters to locate better optima in less computing time (Dueck 1992). For each point in parameter space, the GDA procedure was run 100 times, each start-

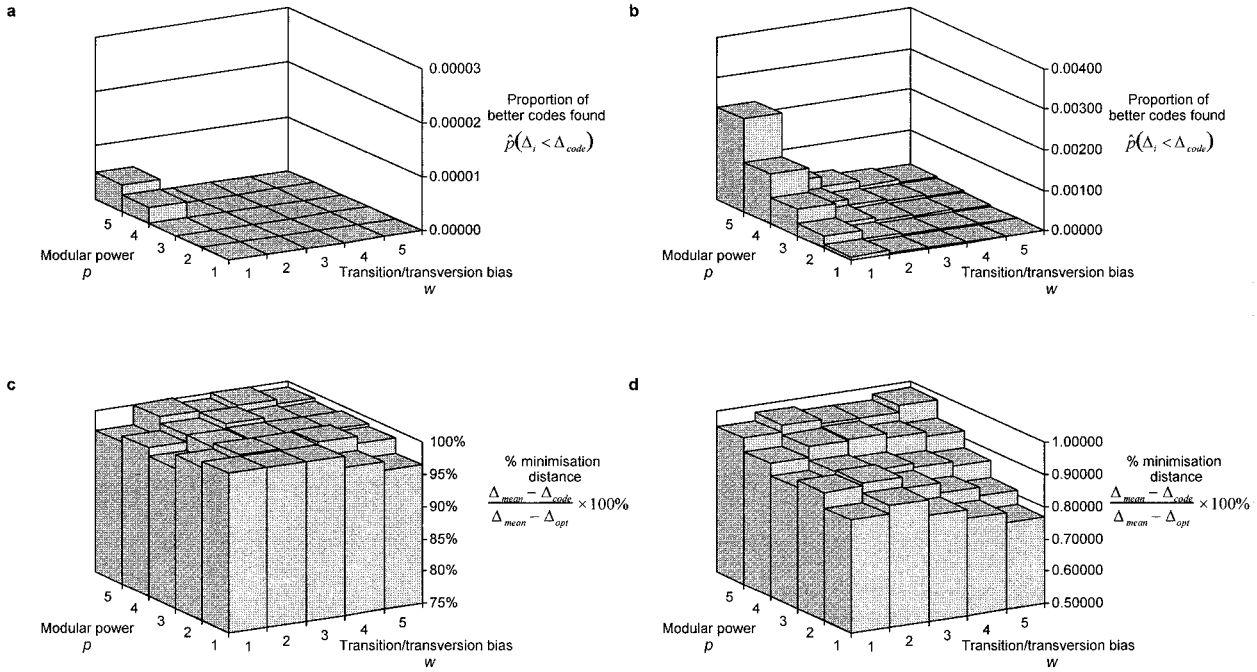

FIG. 2.—Estimates of code optimality. *a*, The proportion of better codes found in a sample of one million random variants drawn from the ''restricted'' set of codes using PAM matrix data. *b*, The same analysis repeated using Polar Requirement to measure amino acid similarity. *c*, Percentage distance minimization estimates for the ''restricted'' set of codes using PAM data. *d*, the same analysis repeated for the ''unrestricted'' set of codes.

ing from a random code configuration. Optima located through the GDA were used to calculate percentage distance minimization estimates (eq. 4) with Δ_{mean} calculated for random samples of one million variant codes.

Results

The Adaptive Code Is No Artifact of Stereochemistry, Analytical Methodology, or Biosynthetic Restrictions

Our analysis shows that when the canonical code is tested against a sample of one million random variants using PAM matrix data to measure amino acid dissimilarity, the code appears to be extremely highly optimized at all transition weightings and modular power functions. For the unrestricted set of codes, no better alternatives are found anywhere (data not shown). This pattern is repeated for the restricted set of possible codes excepting the least plausible extremes of parameter space (no transition bias [$w = 1$] and at a high-modular-power mapping function [$4 \leq p \leq 5$]), where three and five better variants are found, respectively (fig. 2*a*). Far from explaining reports of a highly adapted code structure as an artifact, high-modular-power scaling functions actually cause the code to appear less adaptive. This suggests that overall code optimality is not the result of careful arrangement of a few key outlier amino acids, but is indeed a reflection of a complex and intricate adaptive arrangement.

When the analysis is repeated using Polar Requirement as a similarity measure, the results are remarkably similar: once again, no better alternatives are found in a sample of one million codes drawn from the unrestricted set (data not shown), and the only exceptions for the restricted set of codes are once again found in the absence of a transition bias ($w = 1$) (fig. 2*b*). It is noteworthy that where results differ, those based on PAM matrix data provide consistently higher estimates of code optimality (by around two orders of magnitude) than those based on Polar Requirement; the better the definition of amino acid similarity (in terms of selection), the better the canonical code appears. These observations vindicate previous adaptive evidence as a robust interpretation of code evolution rather than an artifact of, say, a stereochemically determined code.

The Best of All Possible Codes?

When the error value of the standard code is compared with the lowest error value of any code found in an extensive search of parameter space, results are somewhat more variable. Estimates based on PAM data for the restricted set of codes indicate that the canonical code achieves between 96% and 100% optimization relative to the best possible code configuration (fig. 2*c*). If our definition of biosynthetic restrictions are a good approximation of the possible variation from which the canonical code emerged, then it appears at or very close to a global optimum for error minimization: the best of all possible codes.

Previous Reports of a Less Optimized Code are Based on a Flawed Measurement System

Equivalent calculations based on the unrestricted set of possible codes are much more variable over parameter space (fig. 2*d*), placing code optimality between

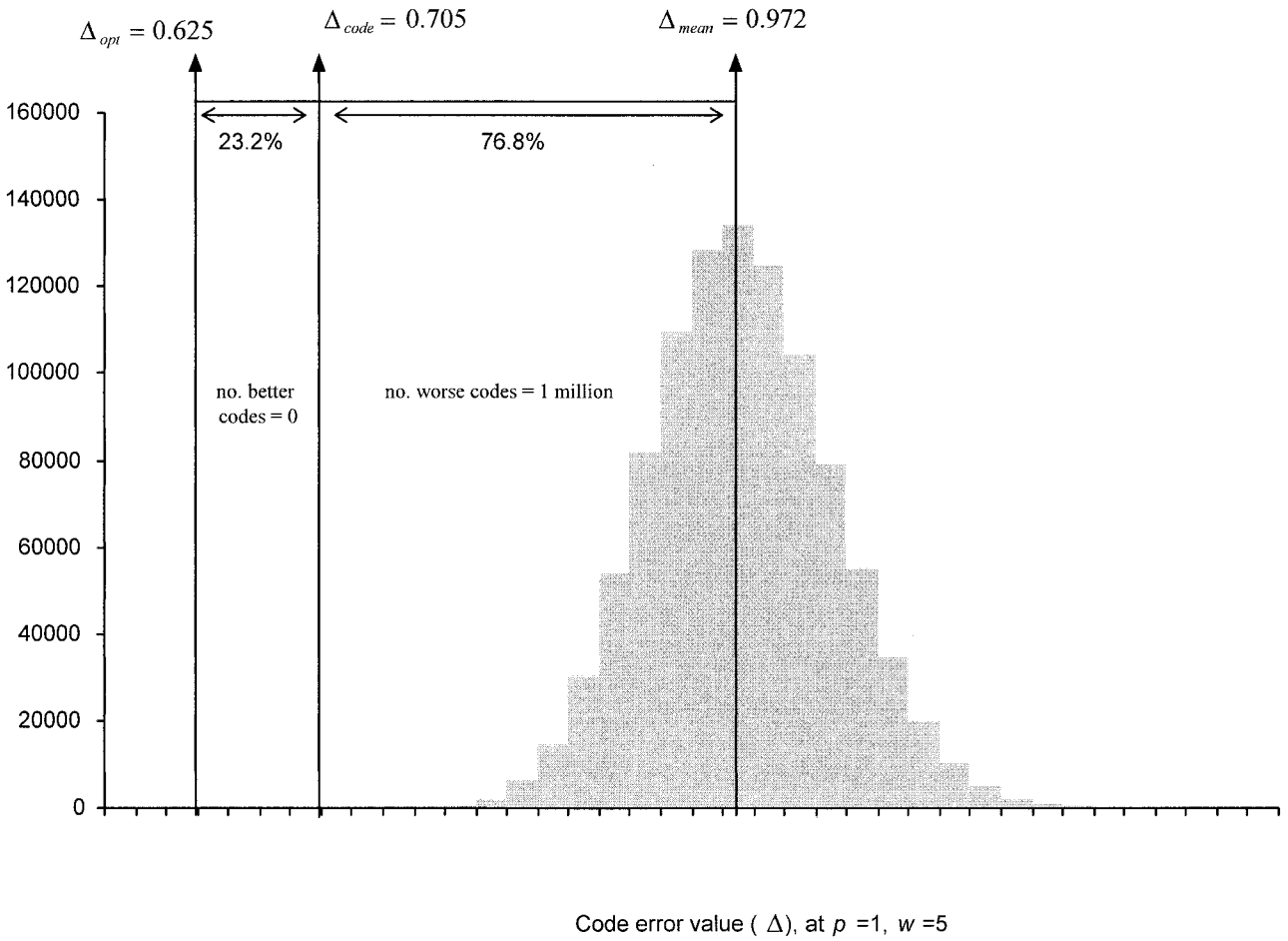

FIG. 3.—Comparison of methods for estimating code optimality: percentage of distance error minimization (''engineering'') versus proportion of better codes found (''statistical''). The distribution shown is for a sample of one million codes drawn from the restricted set using PAM matrix data for a measure of amino acid similarity $p = 1$, $w = 5$. The comparison remains qualitatively unchanged over parameter space.

76% and 97% relative to the global optimum arrangement of codon assignments. However, where sampling and percentage distance minimization provide very different estimates (fig. 3), the sampling method is more biologically relevant: the ''better'' codes implied by percentage error minimization scores are not found within the random sample simply because they are so rare. Percentage error minimization estimates misleadingly measure code optimality on a simple linear scale; the distribution of possible error values is approximately Gaussian rather than uniform. Increasingly fit codes are increasingly rare. Evolutionary changes within the code would thus not follow a linear path of successive fitness increments, but, rather, would approach the global optimum asymptotically. This becomes particularly important where strong conclusions are drawn from the precise values produced by this metric under questionable assumptions (e.g., see DiGiulio and Medugno 1999). Effectively, such studies rely on the flawed assumption that a normally distributed variable observed, say, two standard deviations from the mean is twice as significant as one observed one standard deviation from the mean. Quite simply, only the sampling method of estimating code optimality presents an accurate and robust picture of the strength of natural selection in determining codon assignments.

The observed variation in optimization estimates for the canonical code thus indicates it to be highly optimized under any set of assumptions, but ''the best of all possible codes'' only if biosynthetic restrictions and a moderate transition bias are assumed.

Naturally Occurring Nonstandard Codes Are Less Adaptive

The canonical code may be highly adaptive in terms of error minimization, but what of the secondarily derived nonstandard codes? The error values of all nonstandard codes are equal to or slightly higher than that of the canonical code (fig. 4), indicating that none are more adaptive in this respect. This observation agrees well with a detailed review of explanations for secondary code variation (Osawa 1995). Other factors, including codon usage patterns (associated with fluctuations in genome GC content), genome simplification, and founder effects, probably dominate here. This difference can be understood in terms of the timing of canonical code evolution relative to secondary code divergence. Extant

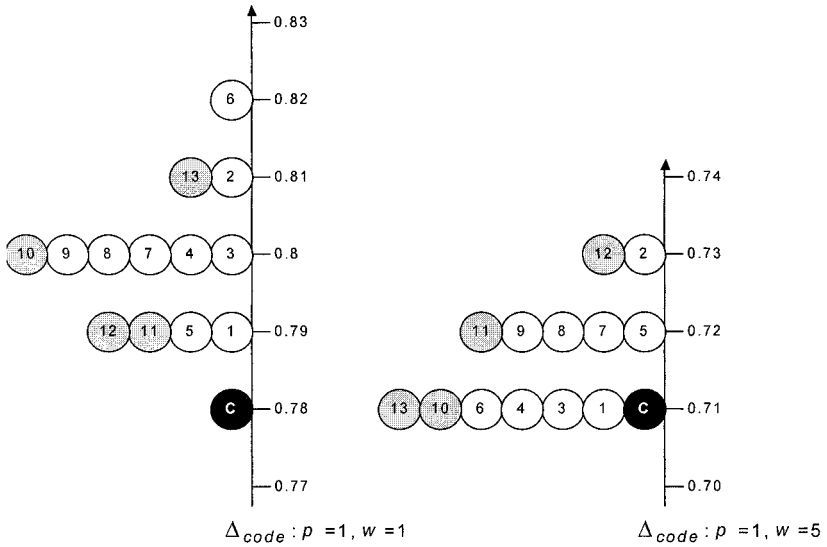

FIG. 4.—A comparison of code error values, Δ, for the canonical genetic code (black circle), the nine mitchondrial nonstandard codes (white circles), and the four nuclear nonstandard codes, as recorded in GenBank (Elzanowski and Ostell 1996). C = canonical code; 1 = vertebrate mitochondrial code; 2 = yeast mitochondrial code; 3 = mold/protist/coelenterate mitochondrial code; 4 = invertebrate mitochondrial code; 5 = echinoderm mitochondrial code; 6 = pyrudae (sea squirt) mitochondrial code; 7 = platyhelminth mitochondrial code; 8 = chlorophycean mitochondrial code; 9 = trematode mitochondrial code; 10 = ciliate nuclear code; 11 = euplotid nuclear code; 12 = yeast nuclear code; 13 = *Blepharisma* spp. (ciliate) nuclear code.

genomes operate in a remarkably error-free environment, by combining use of DNA for genetic information storage with sophisticated protein machinery for replication, translation, and error checking. This is unlikely to have been true for primordial organisms in which the canonical code evolved (Freeland, Knight, and Landweber 1999). In particular, new evidence (Logan et al. 1999) supports the hypothesis that the genetic code emerged in an RNA world before the evolution of DNA (Reichard 1993). Not only is RNA intrinsically more error prone, by several orders of magnitude, than DNA (Lazcano et al. 1988), but where protein enzymes seem to have replaced ribozyme forerunners (Nagel and Doolittle 1995; Wetzel 1995), they are likely to have done so by affording greater catalytic sophistication (Szathmary 1999). Our results are consistent with a model for primordial evolution in which genetic error played a significantly greater role in defining the relative fitness of organisms.

Discussion

Taken together, our results provide strong evidence that the structure of the canonical code was strongly influenced by natural selection for error minimization. Analysis based on PAM matrix values demonstrates not only that results of previous analyses are unlikely to be an artifact of stereochemistry, but that if biosynthetic pathways limited codon assignments, then the code is very near to (and quite possibly at) a global optimum for error minimization. While our implementation of biosynthetic restrictions on possible code evolution may not be entirely accurate, it is the best available at present and is representative of general patterns within the code. Importantly, then, analysis of biosynthetically restricted codes indicates that code coevolution, far from explaining adaptive code structure as an artifact, actually precludes the few ''better'' alternative codon arrangements found in previous analyses.

The Mechanism of Adaptive Code Evolution

This leads to the question of the evolutionary mechanisms responsible for an adaptive canonical code. The many models of precanonical code evolution, reviewed extensively elsewhere (Knight, Freeland, and Landweber 1999), permit two major possibilities: that an adaptive code was selected from a large pool of variants, or that an adaptive code arose de novo by code expansion (or simplification) within adaptive, error-minimizing constraints. Individual codon reassignments, necessary for adaptive code shuffling, are certainly possible, but the question remains unresolved, and two lines of evidence increasingly favor the latter explanation.

First, the notion of code expansion from a simpler primordial form, although still lacking in detail, is now associated with a diverse body of empirical and phylogenetic evidence (Knight, Freeland, and Landweber 1999). It seems unlikely that clear patterns of biosynthetic relatedness would be found in a code which had undergone extensive codon assignment shuffling. Additionally, while adaptive code structure is unlikely to be an artifact of a stereochemically determined code, empirical evidence suggests that stereochemistry is not without a role. For example, RNA molecules artificially selected to bind Arginine contain disproportionately many *CGN*/*AGR* codons (Knight and Landweber 1998). If all or most amino acids show stereochemical affinities for their corresponding codons, this would suggest that natural selection worked in concert with stereochemical interactions and biosynthetic expansion to produce the canonical code de novo, ''choosing'' the current 20 ami-

no acids as those that satisfied criteria for both stereochemical affinity and error minimization. This interpretation would thus offer a novel insight into the selection of the proteinaceous amino acids from the near-infinite possibilities of both prebiotic syntheses and biosynthetic modification.

Conclusions

We have presented comprehensive evidence that the standard genetic code is a product of natural selection to minimize the phenotypic impact of genetic error; the arrangement of codon assignments meets, to an extraordinary degree, the predictions of the adaptive hypothesis and cannot be explained as an artifact of stereochemistry, biosynthetically mediated code expansion, or analytical methodology. However, the process by which an adaptive code evolved at present remains unclear, and yet its resolution may be of key importance to our understanding of the amino acid components universal to life.

Acknowledgments

We would like to thank Nick Goldman, Adam Eyre Walker, Gill McVean, John Barrat, and Dawn Brooks for helpful discussion of this manuscript. This work has been supported by an HFSP fellowship to S.J.F.

LITERATURE CITED

AMIRNOVIN, R. 1997. An analysis of the metabolic theory of the origin of the genetic code. J. Mol. Evol. **44**:473–476.

ARDELL, D. H. 1998. On error minimisation in a sequential origin of the genetic code. J. Mol. Evol. **47**:1–13.

BENNER, S. A., M. A. COHEN, and G. H. GONNET. 1994. Amino acid substitution during functionally divergent evolution of protein sequences. Protein Eng. **7**:1323–1332.

CRICK, F. H. C. 1965. The origin of the genetic code. J. Mol. Biol. **38**:367–379.

DIGIULIO, M. 1989. The extension reached by the minimisation of polarity distances during the evolution of the genetic code. J. Mol. Evol. **29**:288–293.

———. 1994. On the optimisation of the physiochemical distances between amino acids in the evolution of the genetic code. J. Theor. Biol. **168**:43–51.

DIGIULIO, M., and M. MEDUGNO. 1999. Physiochemical optimization in the genetic code origin as the number of codified amino acids increases. J. Mol. Evol. **49**:1–10.

DILLON, L. S. 1973. The origins of the genetic code. Bot. Rev. **39**:301–345.

DUECK, G. 1992. New optimisation heuristics: the great deluge algorithm and record to record travel. J. Comp. Phys. **104**: 86–92.

ELZANOWSKI, A., and J. OSTELL. 1996. The genetic codes. (http://www.ncbi.nlm.nih.gov/Taxonomy/taxpage2.html) National Center for Biotechnology Information (NCBI), Bethesda, Md.

FREELAND, S. J., and L. D. HURST. 1998*a*. The genetic code is one in a million. J. Mol. Evol. **47**:238–248.

———. 1998*b*. Load minimisation of the genetic code: history does not explain the pattern. Proc. R. Soc. Lond. B. Biol. Sci. **265**:2111–2119.

FREELAND, S. J., R. D. KNIGHT, and L. F. LANDWEBER. 1999. Do proteins predate DNA? Science **286**:690–692.

GOLDMAN, N. 1993. Further results on error minimisation in the genetic code. J. Mol. Evol. **37**:662–664.

HAIG, D., and L. D. HURST. 1991. A quantitative measure of error minimisation in the genetic code. J. Mol. Evol. **33**: 412–417.

KNIGHT, R. D., S. J. FREELAND, and L. F. LANDWEBER. 1999. Selection, history and chemistry: the three faces of the genetic code. Trends Biochem. Sci. **24**:241–247.

KNIGHT, R. D., and L. F. LANDWEBER. 1998. Rhyme or reason: RNA-arginine interactions and the genetic code. Chem. Biol. **5**:R215–R220.

LAZCANO, A., R. GUERRERO, L. MARGULIS, and J. ORO. 1988. The evolutionary transition from RNA to DNA in early cells. J. Mol. Evol. **27**:283–290.

LOGAN, D. T., J. ANDERSSON, B. M. SJOBERG, and P. NORDLUND. 1999. A glycyl radical site in the crystal structure of a class III ribonucleotide reductase. Science **283**:1499–1504.

NAGEL, G. M., and R. F. DOOLITTLE. 1995. Phylogenetic analysis of the aminoacyl-tRNA synthetases. J. Mol. Evol. **40**: 487–498.

OSAWA, S. 1995. The evolution of the genetic code. Oxford University Press, Oxford, England.

REICHARD, P. 1993. From RNA to DNA, why so many ribonucleotide reductases? Science **260**:1773–1777.

SJOSTROM, M., and S. WOLD. 1985. A multivariate study of the relationship between the genetic code and the physiochemical properties of the amino acids. J. Mol. Evol. **22**: 272–277.

SONNEBORN, T. M. 1965. Degeneracy in the genetic code: extent, nature and genetic implications. Pp. 377–397 *in* V. BRYSON and H. J. VOGEL, eds. Evolving genes and proteins. Academic Press, New York and London.

SZATHMARY, E. 1999. The origin of the genetic code: amino acids as co-factors in an RNA world. Trends Genet. **15**:223–229.

TAYLOR, F. J. R., and D. COATES. 1989. The code within the codons. Biosystems **22**:177–187.

WETZEL, R. 1995. Evolution of the aminoacyl-tRNA synthetases and the origin of the genetic code. J. Mol. Evol. **40**: 545–550.

WOESE, C. R. 1965. On the evolution of the genetic code. Proc. Natl. Acad. Sci. USA **54**:1546–1552.

WOESE, C. R., D. H. DUGRE, S. A. DUGRE, M. KONDO, and W. C. SAXINGER. 1966. On the fundamental nature and evolution of the genetic code. Cold Spring Harb. Symp. Quant. Biol. **31**:723–736.

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

———. 1980. Role of minimisation of chemical distances between amino acids in the evolution of the genetic code. Proc. Natl. Acad. Sci. USA **77**:1083–1086.

WONG, J. T.-F., and P. M. BRONSKILL. 1979. Inadequacy of prebiotic synthesis as the origin of proteinaceous amino acids. J. Mol. Evol. **13**:115–125.

YARUS, M. 1998. Amino acids as RNA ligands: a direct-RNA-template theory for the code's origin. J. Mol. Evol. **47**:109–117.

ZUCKERKANDL, E., and L. PAULING. 1965. Evolutionary divergence and convergence in proteins. Pp. 97–167 *in* V. BRYSON and H. J. VOGEL, eds. Evolving genes and proteins. Academic Press, New York and London.

PEKKA PAMILO, reviewing editor

Accepted December 10, 1999