J Mol Evol (1991) 33:412–417

Journal of Molecular Evolution

A Quantitative Measure of Error Minimization in the Genetic Code

David Haig[1] and Laurence D. Hurst[2]

[1] Department of Plant Sciences, University of Oxford, South Parks Road, Oxford OX1 3RA, United Kingdom
[2] Department of Zoology, University of Oxford, South Parks Road, Oxford OX1 3PS, United Kingdom

Summary. We have calculated the average effect of changing a codon by a single base for all possible single-base changes in the genetic code and for changes in the first, second, and third codon positions separately. Such values were calculated for an amino acid's polar requirement, hydropathy, molecular volume, and isoelectric point. For each attribute the average effect of single-base changes was also calculated for a large number of randomly generated codes that retained the same level of redundancy as the natural code. Amino acids whose codons differed by a single base in the first and third codon positions were very similar with respect to polar requirement and hydropathy. The major differences between amino acids were specified by the second codon position. Codons with U in the second position are hydrophobic, whereas most codons with A in the second position are hydrophilic. This accounts for the observation of complementary hydropathy. Single-base changes in the natural code had a smaller average effect on polar requirement than all but 0.02% of random codes. This result is most easily explained by selection to minimize deleterious effects of translation errors during the early evolution of the code.

Key words: Genetic code — Complementary hydropathy — Translation

Introduction

As the genetic code was being deciphered in the 1960s, molecular biologists recognized that similar codons often specify similar amino acids. That is, the code is organized such that codons that differ by a single base specify amino acids that are more similar than would be expected if codons had been assigned to amino acids at random. Such a property of the code might have evolved because it reduced the average phenotypic effects of single-base substitutions or of base-pairing errors during transcription and translation (e.g., Sonneborn 1965; Epstein 1966; Goldberg and Wittes 1966; Alff-Steinberger 1969). In this view, the universal code was selected from among a range of variant codes because it was relatively insensitive to the effects of mutational and/or translational errors. At some stage, adaptive evolution of the code ceased because organisms became sufficiently complex that further changes to the code would have been incompatible with survival (Crick 1968).

There have been many models for the evolution of the genetic code, and we will not review them here. Rather, we will briefly discuss two models (Crick 1968; Woese 1973) to illustrate different possible explanations of a tendency for similar codons to specify similar amino acids. Crick (1968) proposed that early versions of the code would have specified many fewer amino acids than the modern code, but that most codons would have specified an amino acid. "In subsequent steps additional amino acids were substituted when they were able to confer a selective advantage, until eventually the code became frozen in its present form." Similar amino acids tended to have similar codons because (1) this diminished the deleterious effects of the initial substitution, and (2) the new tRNA and aminoacyl-tRNA synthetase might have been duplications or modifications of the old tRNA and synthetase, in which case the new amino acid would be likely to be structurally similar to the old amino acid.

Woese and coworkers (Woese 1965, 1973; Woese et al. 1966) argued that similar codons correspond to similar amino acids because the earliest forms of

Table 1. Values for the polar requirement (Woese et al. 1966), hydropathy (Kyte and Doolittle 1982), molecular volume (Grantham 1974), and isoelectric point (Alff-Steinberger 1969) used in this paper

	Polar require-ment	Hydropathy	Molecular volume	Iso-electric point
Ala	7.0	1.8	31	6.00
Arg	9.1	−4.5	124	10.76
Asp	13.0	−3.5	54	2.77
Asn	10.0	−3.5	56	5.41
Cys	4.8	2.5	55	5.07
Glu	12.5	−3.5	83	3.22
Gln	8.6	−3.5	85	5.65
Gly	7.9	−0.4	3	5.97
His	8.4	−3.2	96	7.59
Ile	4.9	4.5	111	6.02
Leu	4.9	3.8	111	5.98
Lys	10.1	−3.9	119	9.74
Met	5.3	1.9	105	5.74
Phe	5.0	2.8	132	5.48
Pro	6.6	−1.6	32.5	6.30
Ser	7.5	−0.8	32	5.68
Thr	6.6	−0.7	61	6.16
Trp	5.2	−0.9	170	5.89
Tyr	5.4	−1.3	136	5.66
Val	5.6	4.2	84	5.96

translation were imprecise, and the distant ancestors of tRNAs were only able to recognize classes of similar codons (an extreme form of wobble) and classes of similar amino acids. In this view, the modern version of the code evolved through a gradual increase in the discrimination of tRNAs for specific amino acids and specific codons within these ancestral sets. Adaptive evolution proceeded through the elimination of less precise versions of the code by their more precise descendants. Woese coupled these ideas with the additional hypothesis that stereochemical associations existed between bases and amino acids such that codon assignments were to some degree predetermined.

Distance Minimization

Amino acids differ from each other in many characters, but the position of an amino acid within the code is most obviously correlated with its hydrophobicity (Epstein 1966; Goldberg and Wittes 1966; Woese et al. 1966). In general, codons with U in the second position specify hydrophobic amino acids, and codons with A in the second position specify hydrophilic amino acids. Multivariate analyses have emphasized the importance of this association between codon assignments and various measures of polarity or hydrophobicity of an amino acid (Sjöström and Wold 1985; Di Giulio 1989a). Polarity and hydrophobicity will be treated as rough synonyms in this article.

None of these studies have satisfactorily quantified the strength of association between position and any amino acid attribute. Some authors (Salemme et al. 1977; Wong 1980) have questioned whether most (nonsynonymous) single-base changes do, in fact, substitute similar amino acids. Wong (1980) estimated that the code has achieved only 45% of the possible distance minimization for an index of similarity that incorporates measures of size, atomic composition, and hydrophobicity. Wong calculated the average distance between neighboring amino acids in the best possible code by minimizing distances for each amino acid separately and then averaging the minimum distances for the 20 amino acids. It is unclear whether such a code can be constructed.

Di Giulio (1989b) estimated that the natural code has achieved 68% minimization of polarity distances. DiGiulio constructed his best code by allowing the polarities of amino acids to vary. Distances between neighboring amino acids were reduced in his code primarily because the code contained fewer strongly hydrophobic and hydrophilic amino acids than occur in the natural code.

In this paper, we adopt a different approach. We do not attempt to derive a best possible code, rather we calculate the average squared change in hydrophobicity, molecular volume, and isoelectric point for all possible single-base changes in the natural code and compare these values to the distribution of similar values calculated for a large number of randomly generated codes. We estimate the efficiency of the natural code by the proportion (P) of random codes that have a smaller average squared difference for single-base changes. The smaller is the value of P, the more efficient is the natural code in minimizing distances between neighboring amino acids.

Methods

We studied four attributes of amino acids: two measures of polarity, a measure of size, and a measure of charge (Table 1). Woese et al.'s (1966) polar requirement is the slope of the line that results when log(1 − R_F)/R_F for free amino acids is plotted against the log mole fraction of water in pyridine solvent. This measure has been used in several analyses of the structure of the genetic code (Alff-Steinberger 1969; Wong 1980; Di Giulio 1989b) and is one of the metrics best correlated with codon position (Sjöström and Wold 1985; Di Giulio 1989a). Kyte and Doolittle's (1982) hydropathy is based on water-vapor transfer free energies, the interior–exterior distribution of amino acid side-chains, and the subjective judgment of the authors. This scale has been used in discussions of complementary hydropathy (see below). Grantham's (1974) molecular volume of side chains is the residue volume minus a constant peptide volume. The values of isoelectric points were taken from Alff-Steinberger (1969).

The mean squared change in an attribute's value was calculated for all single-base substitutions in the first, second, and third codon positions of the natural code (MS_1, MS_2, and MS_3, re-

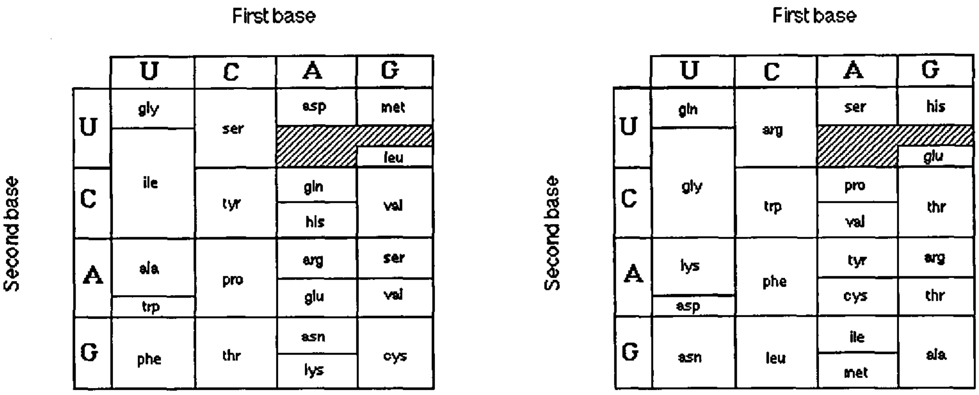

Fig. 1. Randomly generated codes that are more conservative than the natural code (MS_0 = 5.194) with respect to changes in polar requirement. **a** MS_0 = 5.167; **b** MS_0 = 5.189.

Table 2. Mean and standard deviation of MS_1, MS_2, MS_3, and MS_0 for 10,000 randomly generated codes

	MS_1	MS_2	MS_3	MS_0
Polar requirement	12:05 ± 2.77	12.62 ± 2.60	3.58 ± 1.51	9.41 ± 1.51
Hydropathy	17.02 ± 3.43	17.85 ± 3.07	5.05 ± 1.99	13.29 ± 1.70
Molecular volume	3522 ± 765	3690 ± 704	1046 ± 428	2750 ± 399
Isoelectric point	5.956 ± 1.705	6.244 ± 1.657	1.768 ± 0.821	4.651 ± 0.994

spectively), as was the mean squared change for all codon positions combined (MS_0). The change in value was undefined for mutations to and from stop codons, and such mutations were not included in the calculations. Same-sense mutations between synonymous codons were included. The values of mean squared change did not take account of codon usage. Thus, all single-base substitutions within the code were given equal weighting.

For each attribute, MS values were calculated for 10,000 randomly generated codes. The 64 codons of the genetic code were divided into 21 synonymous codon sets. Each set consisted of all the codons specifying the same amino acid in the natural code, plus one set for the three stop codons. In the randomly generated codes, the position of the stop codons remained constant, but the amino acids were assigned at random to the remaining 20 codon sets. There are thus 20! ($>2 \times 10^{18}$) possible codes under our null model. As an illustration, the variant codes with the lowest MS_0 for polar requirement are given in Fig. 1. One code will be described as more conservative than another if it has lower MS_0. P_0 is defined as the proportion of random codes that are more conservative than the natural code. P_1, P_2, and P_3 are similarly defined with reference to MS_1, MS_2, and MS_3.

Our method of generating variant codes does not mimic the evolutionary process. Rather, we use randomly generated codes to derive a probability distribution of MS_0, and use this distribution as a formal device (a null model) to test whether neighboring amino acids in the natural code are more similar with respect to an attribute than would be expected by chance alone. Our null model places strong constraints on the structure of variant codes. All codes have the same level of degeneracy and the same probability of synonymous substitutions as the natural code. Therefore, our results detect error-minimizing features of the code that are additional to third (and second) base redundancy.

Properties of Random Codes

Table 2 presents summary statistics for the mean squared change in the four attributes for 10,000 randomly generated codes. This section uses these data to discuss some general properties of random codes that are a consequence of how the code is divided into synonymous codon sets.

All possible changes from one amino acid to another will occur equally often, when averaged over a very large number of random codes. This is true for changes in all three base positions. Therefore, differences in the average values of MS_1, MS_2, and MS_3 reflect the relative number of synonymous mutations in the three positions. There are 4 synonymous substitutions in the first position, none in the second position, and 126 in the third position (stop → stop not included). Thus, on average, $MS_2 > MS_1 \gg MS_3$.

MS_1 is the most variable among codes, MS_2 slightly less variable, and MS_3 the least variable. The low variance in the third position is easily explained. MS_3 is the average of 126 zeroes (same-sense mutations) and 50 positive values (missense mutations) that vary among codes. The higher variance of MS_1 relative to MS_2 arises because chance juxtapositions of very similar or very dissimilar amino acids can have a greater effect on MS_1 than MS_2. There are 166 missense mutations at the first position (not involving stop codons) that involve 62 different substitutions of one amino acid for another (counting a → b and b → a as different substitutions). Of these 62 substitutions, 4 occur six times, 14 occur four times, 2 occur three times, 38 occur twice, and 4 occur once. On the other hand, all 176 single-base changes at the second position (not involving stop codons) are missense mutations and these involve 82 different substitutions of one

Table 3. MS_1, MS_2, MS_3, and MS_0 for the natural code

	MS_1	MS_2	MS_3	MS_0
Polar requirement	4.88	10.56	0.14	5.19
Hydropathy	5.18	21.78	1.17	9.39
Molecular volume	3272	3458	841	2521
Isoelectric point	9.958	7.352	1.394	6.220

Table 4. The proportion, P, of randomly generated codes (from a sample of 10,000) in which single-base substitutions have a smaller average effect than in the natural code

	P_1	P_2	P_3	P_0
Polar requirement	0.0037	0.2214	0.0002	0.0002
Hydropathy	0.0003	0.9100	0.0142	0.0089
Molecular volume	0.3812	0.3763	0.3503	0.3003
Isoelectric point	0.9828	0.7487	0.3452	0.9281

amino acid for another. Of these substitutions, 12 occur four times, 2 occur three times, 54 occur twice, and 14 occur once.

MS_1 and MS_2 are positively correlated for each of the four attributes, and both are negatively, but more weakly, correlated with MS_3. These correlations arise because MS_1 and MS_2 tend to be large when extreme amino acids are assigned to the six- and four-codon sets, whereas MS_3 tends to be large when extreme amino acids are not assigned to four-codon sets.

Polarity

The natural code is very effective in limiting the average change in polarity caused by single-base substitutions (Tables 3 and 4). Only 2 out of 10,000 random codes were more conservative than the natural code with respect to changes in polar requirement. These two codes are shown in Fig. 1. Similarly, only 89 random codes were more conservative with respect to hydropathy than the natural code. Wong's (1980) improved Code II is actually less conservative than the natural code on both scales.

If the average effect of substitutions is considered for each of the three codon positions, the natural code is very conservative for changes in the first and third position, but less conservative for changes in the second position (Table 4). On the hydropathy scale, more than 90% of randomly generated codes were more conservative than the natural code for second base substitutions, though the equivalent figure was only 22% for polar requirement.

This difference between the polarity scales appears to be a consequence of tyrosine being relatively hydrophobic when judged by polar requirement, but being relatively hydrophilic in terms of the hydropathy scale. In fact, Kyte and Doolittle (1982) subjectively raised the hydrophobicity of tyrosine on their scale because they found it hard to accept that tyrosine was hydrophilic. If this adjustment had not been made, the contrast between the scales would probably have been greater. In terms of polar requirement, tyrosine is the only hydrophobic amino acid among the otherwise hydrophilic amino acids with A in the second position. This reduces P_2 relative to the hydropathy scale because the second position no longer distinguishes so strongly between hydrophobic and hydrophilic amino acids.

The natural code is very conservative with respect to polar requirement ($P_0 = 0.0002$). The striking correspondence between codon assignments and such a simple measure deserves further study.

Complementary Hydropathy

When complementary strands of DNA are read from 5′ to 3′ in the same reading frame, codons for hydrophobic amino acids are generally complemented by codons for hydrophilic amino acids (Blalock and Smith 1984). Brentani (1988, 1990) has proposed that both DNA strands may have had coding capacity during the early evolution of the genetic code. In his view, hydrophobic peptides coded by one strand would have interacted functionally with hydrophilic peptides coded by the complementary strand.

Blalock and Smith (1984) and Brentani (1988, 1990) used Kyte and Doolittle's hydropathy scale. As we have shown, the second codon position discriminates strongly between hydrophobic and hydrophilic amino acids on this scale. Specifically, most strongly hydrophobic amino acids have codons with U in the second position, and most strongly hydrophilic amino acids have codons with A in the second position. As a result, hydrophobic xUy codons are complemented by hydrophilic y′Ax′ codons (where bases x′, y′ are complementary to x, y). This pattern accounts for the observation of complementary hydropathy, and complementary hydropathy is a necessary corollary of any hypothesis that predicts this pattern.

Molecular Volume and Isoelectric Point

The natural code is less conservative with respect to size and charge than it is with respect to polarity. About 30% of random codes are more conservative than the natural code with respect to molecular volume, and over 90% of random codes are more conservative with respect to isoelectric point. On the latter scale, the natural code is particularly nonconservative in the first position ($P_t = 0.98$). This is partly a corollary of the fact that most strongly hydrophilic amino acids share A in the second position. Thus, a first-base change substitutes glutamic acid for lysine. A second factor is that arginine, which has an extreme value for isoelectric point, is assigned a six-codon set in the natural code.

General Discussion

Our results confirm that single-base substitutions are strongly conservative with respect to changes in polar requirement and hydropathy in the first and third codon positions, but much less so in the second codon position. This pattern is more easily accommodated by theories in which the primary selective force is to minimize the effects of codon–anticodon mismatch during translation (or its precursor), rather than to minimize the effects of replication errors. That is, there are many reasons why translation might be initially more error-prone in one position than another, but it is difficult to see why one position should mutate more frequently than another.

Among modern organisms, codon–anticodon mispairing occurs most frequently at those base positions at which pairing errors have least effect (Woese 1965; Goldberg and Wittes 1966; Lagerkvist 1980). Thus, pairing at the third codon position is most error-prone and pairing at the second codon position the least error-prone (Woese 1965). Similarly, abnormal wobble-pairings in the third position are more common when such errors result in synonymous substitutions (Lagerkvist 1980). Such patterns have been used to support the hypothesis that codon assignments evolved to minimize the effects of translational errors. However, the evidence of modern error rates should be treated with caution, unless convincing physico-chemical reasons can be given why certain codon–anticodon pairs are intrinsically less error-prone. This is because natural selection will favor increases in the accuracy of translation, until the costs of further improvements outweigh the benefits. Therefore, the translational apparatus would be expected to evolve an inverse relationship between the frequency and severity of an error, even if such a relationship did not exist in the first place (Kurland 1987; Bulmer 1988).

Our results also suggest that, during the early evolution of the code, the deleterious effects of substituting hydrophilic for hydrophobic amino acids were more severe than the effects of substituting large for small, or acidic for basic, amino acids. Hydrophobic amino acids tend to occupy interior positions within proteins, whereas hydrophilic amino acids tend to occupy exterior positions (e.g., Epstein 1966). Therefore, nonconservative changes in the polarity of an amino acid may have had major effects on the conformation of a protein. It is also possible that the code acquired its major features before the evolution of proteins.

The earliest associations between amino acids and adaptor RNAs probably preceded protein synthesis, and we do not know at what stage in the evolution of the code were amino acids first linked together to form peptides. Adaptors may initially have been used to align substrates (including amino acids) for early metabolic syntheses (Gibson and Lamond 1990), or adaptor-linked amino acids may have had a role in maintaining the correct conformation of catalytic RNAs. For example, hydrophobic amino acids could have been used to anchor ribozymes in membranes. In this latter case, the substitution of one hydrophobic amino acid for another might have little effect, whereas the substitution of a hydrophilic amino acid could be disastrous.

Was the assignment of hydrophobic amino acids to anticodons with A in the second position, and of hydrophilic amino acids to anticodons with U in the second position, completely arbitrary? In a saline solvent system, A nucleotides are the most hydrophobic and U nucleotides the least hydrophobic (Weber and Lacey 1978; data from Garel et al. 1973). Perhaps, chemical affinities between nucleotides and amino acids did contribute to codon assignments (as proposed by Woese et al. 1966). However, a chance correspondence between the hydrophobicities of amino acids and the central base of their anticodon cannot be rejected, given the small number of possible nucleotide pairs. Moreover, given these assignments, the fact that A is complementary to U can account for Blalock and Smith's (1984) observation that complementary strands would encode peptides of opposite hydrophobicity if both strands were translated.

Acknowledgments. This paper has benefited from the thoughtful comments of M. Bulmer, A. Grafen, and A. Pomiankowski. D. Haig is supported by an Endeavour Fellowship from the Royal Society.

References

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

Blalock JE, Smith EM (1984) Hydropathic anti-complementarity of amino acids based on the genetic code. Biochem Biophys Res Comm 121:203–207

Brentani RR (1988) Biological implications of complementary hydropathy of amino acids. J Theor Biol 135:495–499

Brentani RR (1990) Complementary hydropathy and the evolution of interacting peptides. J Mol Evol 31:239–243

Bulmer M (1988) Evolutionary aspects of protein synthesis. Oxf Surv Evol Biol 5:1–40

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

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

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

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

Garel JP, Filliol D, Mandel P (1973) Coéfficients de partage d'aminoacides, nucléobases, nucléosides et nucléotides dans un système solvant salin. J Chromatography 78:381–391

Gibson TJ, Lamond AI (1990) Metabolic complexity in the RNA world and implications for the origin of protein synthesis. J Mol Evol 30:7–15

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

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

Kurland CG (1987) Strategies for efficiency and accuracy in gene expression. 2. Growth optimized ribosomes. Trends Biochem Sci 12:169–171

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

Lagerkvist U (1980) Codon misreading: a restriction operative in the evolution of the genetic code. Am Sci 68:192–198

Salemme FR, Miller MD, Jordan SR (1977) Structural convergence during protein evolution. Proc Natl Acad Sci USA 74:2820–2824

Sjöström M, Wold S (1985) A multivariate study of the relationship between the genetic code and the physical-chemical properties of amino acids. J Mol Evol 22:272–277

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

Weber AL, Lacey JC (1978) Genetic code correlations: amino acids and their anticodon nucleotides. J Mol Evol 11:199–210

Woese CR (1965) On the evolution of the genetic code. Proc Natl Acad Sci USA 54:1546–1552

Woese CR (1973) Evolution of the genetic code. Naturwissenschaften 60:447–459

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

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

Received March 14, 1991/Revised June 7, 1991