

F-H 238 M. Rausher UPG 287

The Genetic Code Is One in a Million

Stephen J. Freeland,¹ Laurence D. Hurst²

¹ Department of Genetics, Downing Street, Cambridge CB2 3EH, UK

² Department of Biology and Biochemistry, University of Bath, Claverton Down, Bath BA27AY, UK

Received: 25 July 1997 / Accepted: 9 January 1998

NOTICE: THIS MATERIAL MAY BE
PROTECTED BY COPYRIGHT LAW
(TITLE 17 U.S. CODE)

Abstract. Statistical and biochemical studies of the genetic code have found evidence of nonrandom patterns in the distribution of codon assignments. It has, for example, been shown that the code minimizes the effects of point mutation or mistranslation: erroneous codons are either synonymous or code for an amino acid with chemical properties very similar to those of the one that would have been present had the error not occurred. This work has suggested that the second base of codons is less efficient in this respect, by about three orders of magnitude, than the first and third bases. These results are based on the assumption that all forms of error at all bases are equally likely. We extend this work to investigate (1) the effect of weighting transition errors differently from transversion errors and (2) the effect of weighting each base differently, depending on reported mistranslation biases. We find that if the bias affects all codon positions equally, as might be expected were the code adapted to a mutational environment with transition/transversion bias, then any reasonable transition/transversion bias increases the relative efficiency of the second base by an order of magnitude. In addition, if we employ weightings to allow for biases in translation, then only 1 in every million random alternative codes generated is more efficient than the natural code. We thus conclude not only that the natural genetic code is extremely efficient at minimizing the effects of errors, but also that its structure reflects biases in these errors, as might be expected were the code the product of selection.

Key words: Genetic code — Error minimization — Mistranslation — Transition/transversion bias — Evolution — Natural selection

Introduction

The genetic code is not random. Wong (1975), for example, has argued that the assignment of codons to amino acids was guided by the biosynthetic relationships between amino acids (but see Amirnovin 1997). Significantly, then, it has been shown that codons specifying amino acids that share the same biochemical synthetic pathway tend to have the same first base (Taylor and Coates 1989). Codons of the amino acid belonging to the shikimate, pyruvate, aspartate, and glutamate families tend to have U, G, A, and C in the first position, respectively.

That amino acids in the same biochemical pathway are coded by related codons does not necessarily explain the observation that amino acids that have similar physicochemical properties also have similar codons (Di Giulio 1997). It was noted early on that the natural genetic code appeared to be arranged such that amino acids with similar chemical properties are coded by similar codons. This may well be the result of selection favoring those codes that minimized the average phenotypic effects of single-point mutations or mistranslations (see, e.g., Alf-Steinberger 1969; Epstein 1966; Goldberg and Wittes 1966; Sonneborn 1965; Woese 1965, 1973; Woese et al. 1966).

An attempt to quantify this effect by Haig and Hurst

Correspondence to: S.J. Freeland; e-mail: s.freeland@gen.cam.ac.uk

(1991) found that of 10,000 randomly generated codes, only 2 performed better at minimizing the effects of error, when polar requirement was taken as the amino acid property (see also Di Giulio 1989; Goldman 1993; Szathmari and Zintzaras 1992). Were this due to the fact that biosynthetically related amino acids share the same first codon position, changes at the second and third base should account for nearly all the efficiency of the code, as these conserve first-base identity. However, first and third bases are very efficient, while the second base proved unexceptional (Haig and Hurst 1991). Put another way, it is the second base that determines the polar requirement of the amino acid. It may hence be concluded that the code is very efficient at minimizing the effects of errors, and this is probably the result of selection between alternative codes, with selection favoring those that minimize the effects of errors on fitness.

The above analysis (Haig and Hurst 1991) made no allowance for any biases in errors. All bases were assumed to be equally prone to error and all forms of error at each base were assumed to be equally likely. However, both mutation and mistranslation are biased processes. Transition mutations tend to occur more frequently than transversion mutations (e.g., see Collins 1994; Kumar 1996; Moriyama and Powell 1997; Morton 1995). Likewise, mistranslation appears to have transition/translation biases as well as biases by codon position (Friedman and Weinstein 1964; Parker 1989; Woese 1965). If the code has evolved to minimize the effects of either, then we might expect that if one allows for biases in the direction of mutation and/or mistranslation, the natural code should, in terms of relative efficiency, perform even better than randomly generated codes. Here, then, we extend the previous analysis so as to ask how these biases effect the relative efficiency of the natural code.

Methods

The Mean Square (MS) Measure

Work presented here uses the "mean square" (MS) measure (Haig and Hurst 1991) to quantify the relative efficiency of any given code. This measure calculates the mean squared change in an amino acid property resulting from all possible changes to each base of all codons within a given code. Any one change is calculated as the squared difference between the amino acid coded for by the original codon and the amino acid coded for by the new (mutated) codon. Synonymous changes are included in the calculation, but changes to and from stop codons are ignored.

The use of a mean square value avoids a problem with negatives, but it also, unavoidably, introduces a form of weighting of the relative importance of differences in chemical property. Many alternative weightings are imaginable. Code fitness might perhaps be a linear function of chemical distance or, perhaps, large changes in chemical property have disproportionately large effects on code fitness. It is hard to know what would be the best model relating chemical distance to code fitness and the mean square seems as valid as any.

The MS measure for any particular genetic code is calculated as four separate values: MS1, MS2, and MS3 correspond to all possible single-base substitutions in the first, second, and third codon positions, respectively, of all codons in a given genetic code; MS0 corresponds to all possible single base changes in all codon positions.

The Weighted Mean Square (WMS) Measure

In an extension to the methods of Haig and Hurst, MS values for each code were calculated at 20 different weightings (weights of 1, 2, . . . , 20) of transition/transversion bias. Because the nature of the MS measure is to incorporate every possible (i.e., transition and transversion) mutation of a given base, it was not possible to weight the *probability* of transitions as opposed to transversions, so our tests weighted the squared difference in polar requirement resulting from transitions differently from that resulting from transversion bias, thus turning the MS measures into WMS measures.

At a weighting of 1, all possible mutations are weighted equally when calculating the MS values for each position of each codon. At a weighting of 2, the differences in amino acid attribute resulting from transition error (i.e., U to C, C to U, A to G, or G to A) were weighted twice as heavily as those resulting from transversion errors. For example, the possible errors used to calculate the MS2 value of codon UUU (Phe) are UCU (Ser), UAU (Tyr) and UGU (Cys), of which only UCU (Ser) represents a transition error, Tyr and Cys resulting from transversion errors.

All MS calculations presented here consider a single attribute of amino acids, namely, the "polar requirement" (Woese et al. 1966), which may be considered a measure of hydrophobicity: amino acids with a high polar requirement are more strongly hydrophobic than those with a low polar requirement. This particular measure was chosen because previous work (Haig and Hurst 1991) on optimization of the genetic code found it to give the most significant evidence of load minimization from an array of four amino acid properties (also tested were hydrophobicity, molecular volume, and isoelectric point).

Rules for Forming Variant Genetic Codes

In general, MS values for a given transition/transversion bias were calculated for a large number of randomly generated variant genetic codes. Our criteria for creating plausible alternative codes are the same as those used by Haig and Hurst (1991).

1. The "codon space" (i.e., the 64 possible codons) is divided into the 21 nonoverlapping sets of codons observed in the natural code, each set comprising all codons specifying a particular amino acid in the natural code (20 sets for the amino acids and 1 set for the 3 stop codons).
2. Each alternative code is formed by randomly assigning each of the 20 amino acids to one of these sets. All three stop codons remain invariant in position for all alternative codes.

The MS measures of each sample of codes generated by this process form a probability distribution against which the real code MS values may be compared. As noted by Haig and Hurst, variant codes produced by this method retain the level of redundancy inherent to the natural code, i.e., this method controls for the redundancy inherent in the code.

These methods were incorporated into an ANSI "C" program which calculated MS values for all randomly generated codes over a range of weightings for transition/transversion bias. The program also generated basic descriptive statistics (mean, range, standard deviation) for each of the MS distributions formed, explored the nature and behavior of superficially "better" (lower WMS0) codes, and explored different models for weighting bias.

Table 1. Basic descriptive statistics for the distributions of possible MS values from which the natural genetic code is drawn: comparison of our results (sample size, 1 million) with those reported by Haig and Hurst (sample size, 10,000)

Measure	Our calculations ($n = 1,000,000$)	Haig and Hurst (1991) calculations ($n = 10,000$)
Mean \pm SD		
MS0	9.41 \pm 1.51	9.41 \pm 1.51
MS1	12.04 \pm 2.80	12.05 \pm 2.77
MS2	12.63 \pm 2.60	12.62 \pm 2.60
MS3	3.59 \pm 1.50	3.58 \pm 1.51
Proportion of better codes found		
MS0	0.0001	0.0002
MS1	0.0030	0.0037
MS2	0.2216	0.2214
MS3	0.0001	0.0002

Results

Equal Transition/Transversion Bias

Haig and Hurst (1991) calculated MS values for 10,000 randomly generated codes and found only 2 variants which were more conservative (i.e., gave a lower MS0) than the natural code. They thus estimated that the chance that a code as conservative (for polar requirement) as the natural code arose by chance was 0.00002 and, therefore, concluded that the natural code was a product of natural selection for load minimization.

Our results, based on a sample size of 1,000,000, found 114 "better" (lower MS0) codes (a proportion of 0.000114), indicating a refinement to the previous estimate for relative code efficiency such that the code be considered almost twice as conservative as suggested previously. A full comparison of MS values found during our test and those described by Haig and Hurst (1991) is shown in Table 1.

Figures 1a–d show the distribution of MS values of the 1 million random variants generated by our program. In each plot, the appropriate MS value of the natural genetic code is indicated by an arrow: the area under the curve to the left of the arrow thus indicates the number of variant codes which are more conservative than the natural code and may, thus, be used to estimate the probability of a code as efficient as the natural genetic code arising through chance alone.

These plots give a graphical context in which to view the load minimization of the natural genetic code (in terms of polar requirement) reported by Haig and Hurst (1991). As reported previously, both the first and the third bases show strong evidence for adaptation to load minimization but that the second base shows no significant evidence of optimization (indeed, MS2 varies little from the mean of the random sample).

One previously unreported feature seen in these plots is the grainy or "spiky" nature of the frequency distri-

bution, most pronounced for the MS1 and MS3 measurements. Several tests were applied to investigate the cause of this pattern. First, the programs were run with an entirely different pseudo-random number generator to verify that the observed pattern was not an artifact of the period or mechanism of the algorithm used to create variant genetic codes. This had no effect on the observed distribution. Having ruled out this possibility, further investigations used "fake" codes (with different codon block structures) and "fake" amino acid sets (with differing distributions of polar requirement). Both types of test changed the smoothness of the distribution, indicating the observed phenomenon to be a combined result of the discrete, clumped distribution of amino acid polar requirement (Fig. 2) and of the patterns of codon blocks in the first and third bases. For example, there are $4!$ variant codes in which all codons beginning with C have a polar requirement of either 4.9 or 6.7, meaning that there are $4!$ variant codes which all give the same MS1 and MS3's for codons CNN. In effect, the MS distributions show some characteristics reminiscent of the normal distribution (as might be expected from the central limit theorem) but differ in their detail because our methodology for creating variant codes maintains the codon block structure and the distribution of amino acid polar requirement.

Introducing a Transition/Transversion Bias

Previous MS calculations have been based on the assumption that transition errors (i.e., $C \leftrightarrow T$ and $A \leftrightarrow G$) and transversion errors (i.e. $C, U \leftrightarrow A, G$) are equally likely to occur. In contrast, we carried out a second set of simulations in which we generated 100,000 variant codes and tested each of these at 20 different weightings of transition/transversion bias (see below).

The results of these tests are summarized in Figs. 3 and 4. In each plot, the Y axis represents the proportion of random variant codes found which were more conservative (lower WMS) than the actual genetic code, and the X axis represents different weightings of transition:transversion bias. The first plot shows the number of more conservative codes for all WMS measures, while the second plot shows the same, but with the Y axis rescaled to clarify the behavior of WMS0, WMS1, and WMS3.

The most startling feature of these plots is the dramatic effect of transition:transversion bias on the relative efficiency of the second codon base: the number of better codes (i.e., variant codes for which the WMS2 measure is smaller than that of the natural code) decreases almost sixfold as the transition:transversion bias increases from 1 to 5. This increase in second-base relative efficiency is most pronounced at low weightings, appearing to reach an asymptote as the weighting bias becomes extreme. Even at a high transition/transversion bias, however, the second base remains an order of magnitude less relatively efficient than the first and third bases.

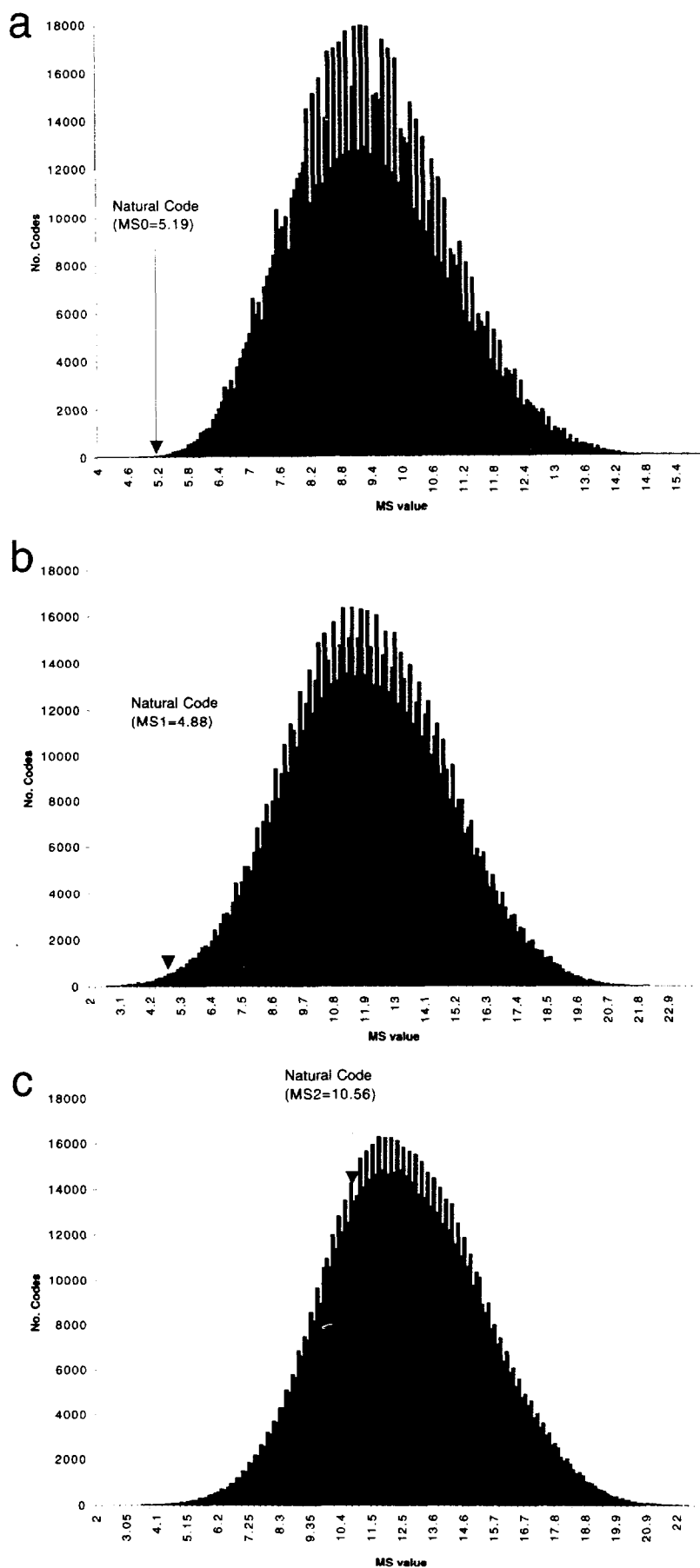


Fig. 1. Histograms for the MS values obtained from 1 million randomly generated variants of the natural genetic code. In each plot, the X axis gives a particular range of categories of MS values, and Y axis gives the number of random variant codes generated with an MS value in that category (from a sample of 1 million random variant codes tested). In addition, the arrow in each plot shows the category into which the appropriate MS calculation for the natural code falls: the cumulative frequency to the left of this arrow therefore indicates the number of more conservative codes found among the random variants and, thus, is used to estimate the probability of a code as efficient as the natural code arising by chance alone: **a** MS0, 114 "better" codes found ($P = 0.0001$); **b** MS1, 2964 better codes found ($P = 0.0030$); **c** MS2, 221,633 better codes found ($P = 0.2216$); **d** MS3, 88 better codes found ($P = 0.00009$).

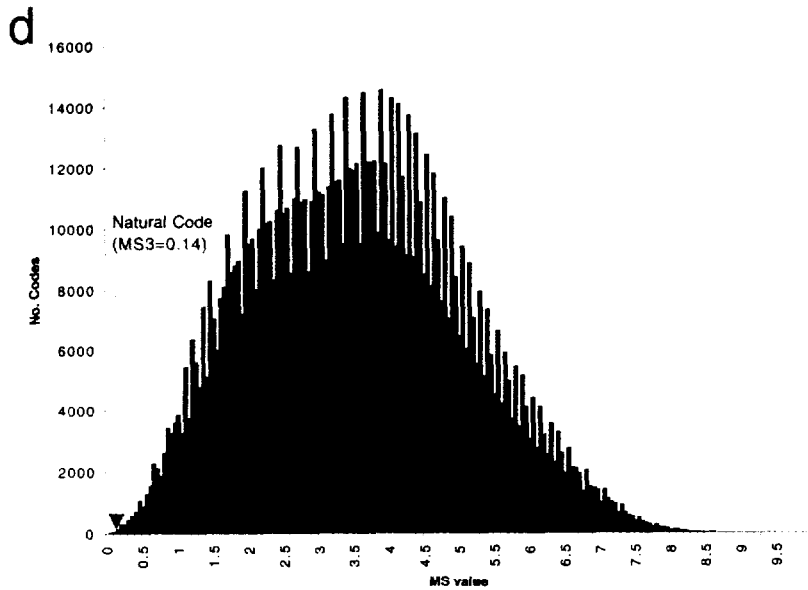


Fig. 1. Continued.

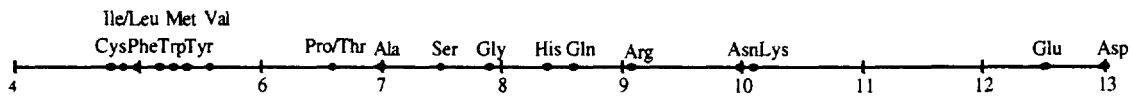


Fig. 2. The distribution of polar requirement values for the 20 naturally occurring amino acids (data from Woese et al. 1966).

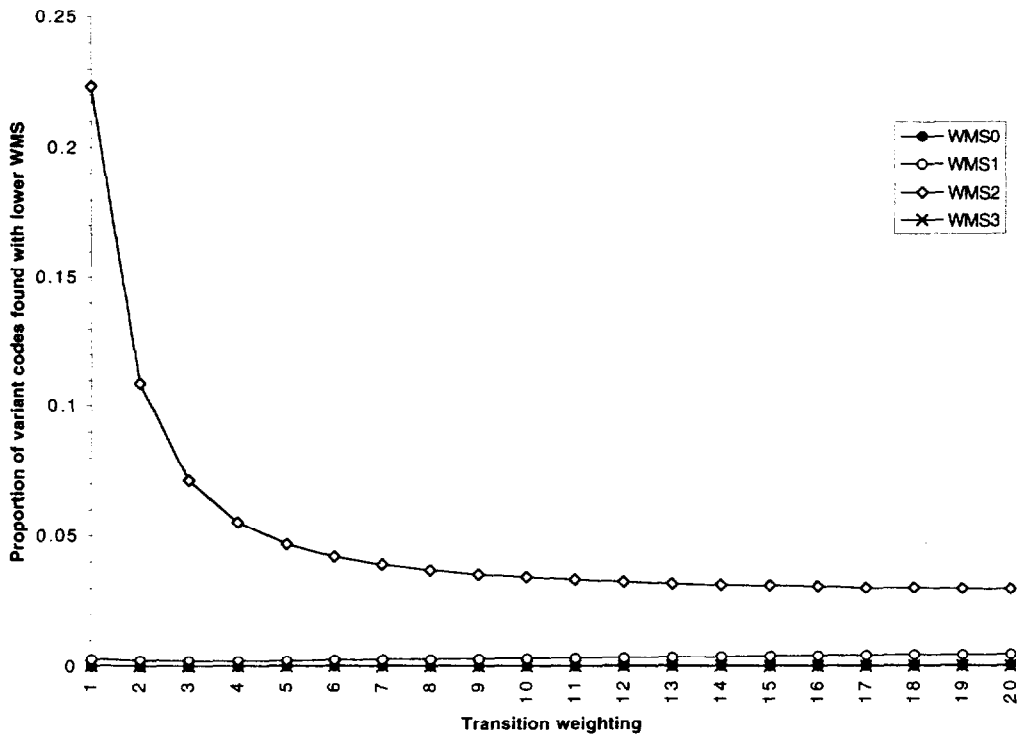


Fig. 3. The proportion (of a sample of 100,000 variants) of "better" (lower WMS) codes found at each of 20 weightings for transition/transversion bias for the first, second, and third and all codon bases combined.

In contrast to the second base, the relative efficiency of the first base improves as a low transition bias is applied (optimizing at a bias of about 3) and then steadily worsens with increasing bias (from 3 to 20). In fact, our

results show that the value of WMS1 decreases (i.e., the base 1 relative efficiency increases) as the transition bias increases from 1 to 20 (i.e., it behaves in a qualitatively similar manner to WMS2) but that this increase in effi-

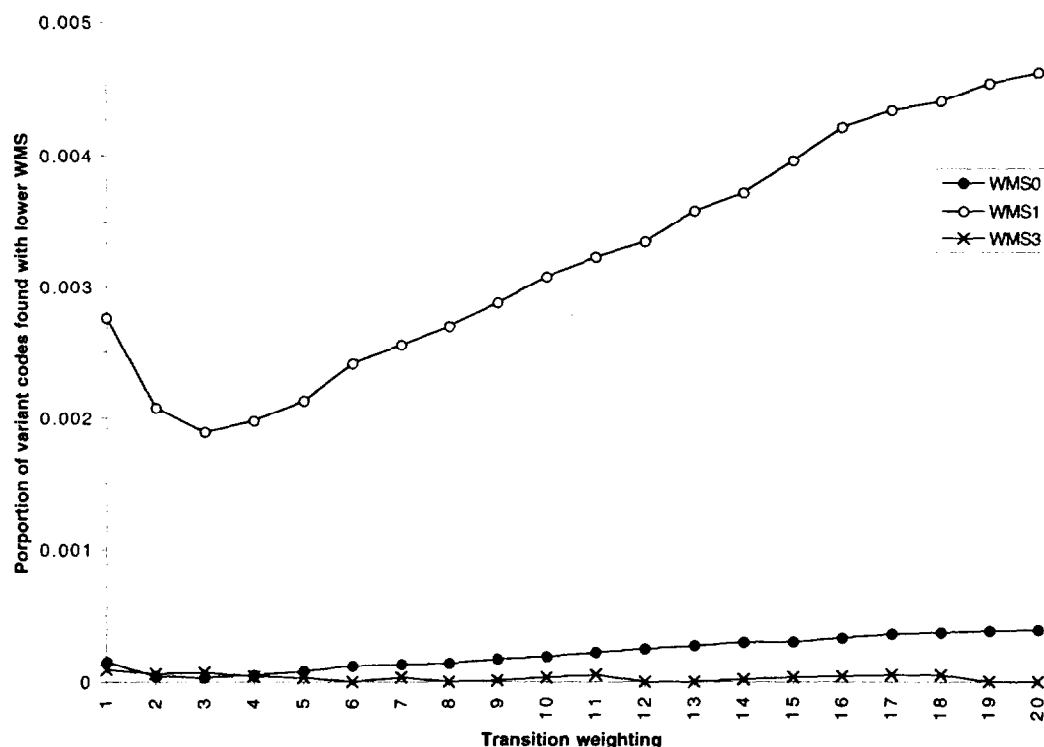


Fig. 4. The proportion (of a sample of 100,000 variants) of "better" codes (lower WMS) found at each of 20 weightings for transition/transversion bias for the first and third and all codon bases combined.

ciency is less pronounced than that of many random variants, so that its comparative efficiency decreases.

The relative efficiency of the third base changes very little across the entire range of transition biases applied. This might be expected from the pattern of redundancy inherent within both the natural code and our variants, in that redundancy is limited almost-exclusively to the third-base positions: the structure of the code is such that half of all possible codons exist as members of a "family box" of four, so that both transition and transversion mutations in the third position of such codons are all synonymous.

These effects in individual bases combine in such a way that the overall relative efficiency of the natural code (as measured by WMS0) increases with increasing transition/transversion ratio up to a bias of approximately 3 (i.e., transition errors being weighted as three times more likely than transversion errors). As the transition bias increases still further, however, the relative efficiency of the natural code decreases: at a bias between 8 and 9, the code is as efficient as it is with no bias at all, and thereafter it is relatively less efficient. This observation coincides quite well with typical empirical data, which reveal general transition/transversion biases of between 1.7 and 5 (e.g., see Collins 1994; Kumar 1996; Moriyama and Powell 1997; Morton 1995). Although these data represent the biases observed in extant taxa, the cause of the bias is primarily physiochemical in that the two purines are of a similar size and shape as are the pyrimidines, but the two groups are different from each

other. It therefore seems reasonable to suppose that the biases observed now were present to a similar extent during the early evolution of life.

Testing "More Conservative" Codes with a Transition Bias

Given that the WMS values calculated for the natural genetic code vary considerably as the transition/transversion bias is altered, we next addressed the question of how seemingly "better" codes (i.e., lower MS0) behave as the transition/transversion bias is increased. Specifically we calculated WMS0 values at transition biases of 1..20 for the first 15 better codes found. Figure 5 shows the behavior of WMS0 for each code at different transition weightings. The 15 superficially better codes used for this study are given in Fig. 6.

Two clear observations may be made of the WMS calculations. First, it may be seen that of the 15 superficially better codes, only 1 remains more efficient (lower WMS0) at transition weightings of 3 or above. This is in keeping with the observations of improvement in WMS values for all bases of the natural code with increasing transition bias. Second, a related but more general observation is that MS0 (i.e., WMS0 at a transition/transversion bias of 1) is in no way a predictor of WMS0 behavior at higher transition weightings: code 13 is more conservative when no transition bias is applied (MS0 = 4.73) but is less efficient than four other codes

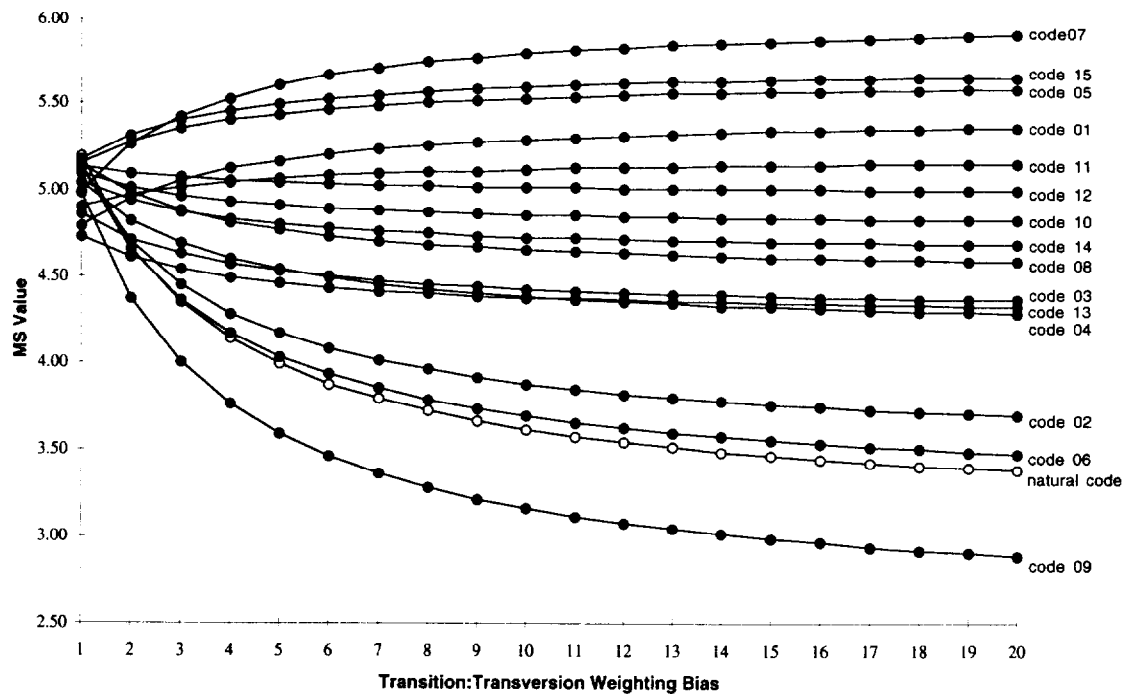


Fig. 5. The behavior of WMS0 values to 15 superficially "better" codes at each of 20 transition/transversion weightings.

(including the natural code) even at a transition weighting of 3.

Mapping Translational Error Data to MS Calculations

Both the work carried out by Haig and Hurst and that presented above assume that mistakes are equally likely to be made at any of the three codon positions: it is under this assumption that the second base remains an order of magnitude less efficient than either the first or the third base (Fig. 3). This assumption is plausible if we are considering point mutations to DNA which are accurately translated via mRNA into an erroneous amino acid(s) but must be reconsidered if we are considering mistranslation of accurate mRNA. In essence there is no reason to suspect that rates of point mutations vary depending upon codon base number, but translation machinery acts upon mRNA in a specific reading frame, reading bases in triplets. The possibility thus exists that translation accuracy does indeed vary in a consistent manner according to base position within a codon (e.g., the "wobble" rules).

Empirical data for the rate and nature of translational errors indicate that the detail of their occurrence varies in a complex manner [according to the specific codon under consideration (Parker 1989)] but, in general, appears to support this assertion (see Friedman and Weinstein 1964; Parker 1989; Woese 1965). Investigation of the polypeptide product resulting from *in vitro* translation of poly(U) mRNA suggests the following patterns (Friedman and Weinstein 1964).

(a) Mistranslation of the second codon position is much less frequent than mistranslation of either the first or

the third codon position. Mistranslation of the first codon position is less frequent than mistranslation of the third codon position.

- (b) Those mistranslations which do occur at the second codon position appear to be almost-exclusively transitional (as opposed to transversional) in nature.
- (c) At the first codon position, mistranslations appear to be fairly heavily biased toward transitional errors.
- (d) At the third codon position, there is very little (if any) transition bias.

We therefore carried out a further series of MS calculations which reflected these observations (we term these measures *tMS*), both for the natural code and for a sample of 1 million random variant codes. The precise quantification of mistranslation data used in creating weighted MS values is given in Table 2. The distribution of results obtained in this manner, together with the relative position of the natural code, is shown in Fig. 7, and simple descriptive statistics of the distribution of plausible codes are given in Table 3.

From Fig. 7, it is apparent that even under our rather crude estimations of the relative rate and nature of mistranslations at each codon position, the natural genetic code shows startling evidence of optimization, two orders of magnitude higher than has been suggested previously. Though the precise quantification used here may be questioned, the overall result seems fairly clear: under our model, of 1 million random variant codes produced, only 1 was better (i.e., had a lower *tMS*) than the natural code—our genetic code is quite literally "1 in a million." The single more conservative code found is shown in Fig. 8 alongside the natural genetic code; superficially

UUU	01	UCU	06	UAU	10	UGU	17
UUC		UCC		UAC		UGC	
UUA	02	UCA		UAA	TER	UGA	TER
UUG		UCG		UAG		UGG	18
CUU		CCU	07	CAU	11	CGU	
CUC	02	CCC		CAC		CGC	19
CUA		CCA		CAA	12	CGA	
CUG		CCG		CAG		CGG	
AUU		ACU	08	AAU	13	AGU	06
AUC	03	ACC		AAC		AGC	
AUA		ACA		AAA	14	AGA	19
AUG	04	ACG		AAG		AGG	
GUU		GCU	09	GAU	15	GGU	
GUC	05	GCC		GAC		GGC	20
GUA		GCA		GAA	16	GGA	
GUG		GCG		GAG		GGG	

MEANING	01	02	03	04	05	06	07	08	09	10	11	12	13	14	15	16	17	18	19	20
Nat. code	<i>Phe</i>	<i>Leu</i>	<i>Ile</i>	<i>Met</i>	<i>Val</i>	<i>Ser</i>	<i>Pro</i>	<i>Thr</i>	<i>Ala</i>	<i>Tyr</i>	<i>His</i>	<i>Gln</i>	<i>Asn</i>	<i>Lys</i>	<i>Asp</i>	<i>Glu</i>	<i>Cys</i>	<i>Trp</i>	<i>Arg</i>	<i>Gly</i>
code 1	<i>Phe</i>	<i>Pro</i>	<i>Ala</i>	<i>Cys</i>	<i>Gln</i>	<i>Gly</i>	<i>Trp</i>	<i>Leu</i>	<i>Arg</i>	<i>Val</i>	<i>Tyr</i>	<i>Ile</i>	<i>Thr</i>	<i>Met</i>	<i>His</i>	<i>Asn</i>	<i>Asp</i>	<i>Glu</i>	<i>Ser</i>	<i>Lys</i>
code 2	<i>Trp</i>	<i>Thr</i>	<i>Cys</i>	<i>Val</i>	<i>Leu</i>	<i>Ala</i>	<i>Met</i>	<i>Tyr</i>	<i>Pro</i>	<i>Arg</i>	<i>Asn</i>	<i>Lys</i>	<i>Ser</i>	<i>Gly</i>	<i>Asp</i>	<i>Glu</i>	<i>Phe</i>	<i>Ile</i>	<i>His</i>	<i>Gln</i>
code 3	<i>Arg</i>	<i>Pro</i>	<i>Trp</i>	<i>Phe</i>	<i>Ala</i>	<i>Tyr</i>	<i>Met</i>	<i>Val</i>	<i>Thr</i>	<i>Lys</i>	<i>His</i>	<i>Cys</i>	<i>Ser</i>	<i>Ile</i>	<i>Asn</i>	<i>Gln</i>	<i>Glu</i>	<i>Asp</i>	<i>Leu</i>	<i>Gly</i>
code 4	<i>Leu</i>	<i>Phe</i>	<i>Met</i>	<i>Tyr</i>	<i>Arg</i>	<i>Thr</i>	<i>Ala</i>	<i>Ile</i>	<i>Ser</i>	<i>Trp</i>	<i>Cys</i>	<i>Gln</i>	<i>Val</i>	<i>His</i>	<i>Gly</i>	<i>Glu</i>	<i>Asn</i>	<i>Asp</i>	<i>Pro</i>	<i>Lys</i>
code 5	<i>Ala</i>	<i>Ser</i>	<i>Phe</i>	<i>Trp</i>	<i>Pro</i>	<i>Asn</i>	<i>Glu</i>	<i>Lys</i>	<i>His</i>	<i>Tyr</i>	<i>Thr</i>	<i>Met</i>	<i>Ile</i>	<i>Leu</i>	<i>Val</i>	<i>Cys</i>	<i>Gln</i>	<i>Asp</i>	<i>Arg</i>	<i>Gly</i>
code 6	<i>Leu</i>	<i>Arg</i>	<i>Lys</i>	<i>Asp</i>	<i>Asn</i>	<i>Met</i>	<i>Ala</i>	<i>Gly</i>	<i>Thr</i>	<i>Tyr</i>	<i>Ile</i>	<i>Phe</i>	<i>Trp</i>	<i>His</i>	<i>Val</i>	<i>Gln</i>	<i>Cys</i>	<i>Glu</i>	<i>Pro</i>	<i>Ser</i>
code 7	<i>Tyr</i>	<i>Cys</i>	<i>Leu</i>	<i>Val</i>	<i>Pro</i>	<i>Lys</i>	<i>Ala</i>	<i>Arg</i>	<i>Gly</i>	<i>Ser</i>	<i>Phe</i>	<i>Trp</i>	<i>His</i>	<i>Thr</i>	<i>Met</i>	<i>Ile</i>	<i>Glu</i>	<i>Asp</i>	<i>Asn</i>	<i>Gln</i>
code 8	<i>Ile</i>	<i>Phe</i>	<i>Trp</i>	<i>Cys</i>	<i>Met</i>	<i>Gly</i>	<i>Gln</i>	<i>Pro</i>	<i>His</i>	<i>Val</i>	<i>Lys</i>	<i>Asp</i>	<i>Thr</i>	<i>Tyr</i>	<i>Arg</i>	<i>Asn</i>	<i>Leu</i>	<i>Glu</i>	<i>Ser</i>	<i>Ala</i>
code 9	<i>Asn</i>	<i>Gln</i>	<i>Met</i>	<i>Trp</i>	<i>Thr</i>	<i>His</i>	<i>Lys</i>	<i>Ala</i>	<i>Pro</i>	<i>Asp</i>	<i>Gly</i>	<i>Val</i>	<i>Phe</i>	<i>Cys</i>	<i>Ile</i>	<i>Leu</i>	<i>Glu</i>	<i>Arg</i>	<i>Ser</i>	<i>Tyr</i>
code 10	<i>Arg</i>	<i>Thr</i>	<i>Asn</i>	<i>Asp</i>	<i>Ile</i>	<i>Ser</i>	<i>Gln</i>	<i>His</i>	<i>Leu</i>	<i>Tyr</i>	<i>Gly</i>	<i>Phe</i>	<i>Met</i>	<i>Val</i>	<i>Trp</i>	<i>Cys</i>	<i>Lys</i>	<i>Glu</i>	<i>Pro</i>	<i>Ala</i>
code 11	<i>Val</i>	<i>Thr</i>	<i>Cys</i>	<i>Phe</i>	<i>Gln</i>	<i>Trp</i>	<i>Tyr</i>	<i>Met</i>	<i>Gly</i>	<i>Asn</i>	<i>Pro</i>	<i>Lys</i>	<i>Ala</i>	<i>Arg</i>	<i>Glu</i>	<i>Asp</i>	<i>Ser</i>	<i>Leu</i>	<i>Ile</i>	<i>His</i>
code 12	<i>Glu</i>	<i>Arg</i>	<i>Lys</i>	<i>Asp</i>	<i>Ala</i>	<i>Pro</i>	<i>His</i>	<i>Val</i>	<i>Cys</i>	<i>Thr</i>	<i>Ser</i>	<i>Leu</i>	<i>Phe</i>	<i>Trp</i>	<i>Met</i>	<i>Ile</i>	<i>Asn</i>	<i>Gly</i>	<i>Gln</i>	<i>Tyr</i>
code 13	<i>Glu</i>	<i>Arg</i>	<i>Asn</i>	<i>Asp</i>	<i>Lys</i>	<i>Ser</i>	<i>Ala</i>	<i>Pro</i>	<i>Met</i>	<i>Gln</i>	<i>Trp</i>	<i>Leu</i>	<i>Val</i>	<i>Phe</i>	<i>Gly</i>	<i>Cys</i>	<i>His</i>	<i>Ile</i>	<i>Thr</i>	<i>Tyr</i>
code 14	<i>Cys</i>	<i>Ala</i>	<i>Asn</i>	<i>Glu</i>	<i>Phe</i>	<i>Leu</i>	<i>Pro</i>	<i>Ser</i>	<i>Met</i>	<i>Val</i>	<i>Arg</i>	<i>His</i>	<i>Gly</i>	<i>Lys</i>	<i>Tyr</i>	<i>Ile</i>	<i>Trp</i>	<i>Asp</i>	<i>Gln</i>	<i>Thr</i>
code 15	<i>Gly</i>	<i>Gln</i>	<i>Asn</i>	<i>Lys</i>	<i>Ala</i>	<i>Thr</i>	<i>Phe</i>	<i>Val</i>	<i>Cys</i>	<i>Tyr</i>	<i>Ile</i>	<i>Pro</i>	<i>Met</i>	<i>Leu</i>	<i>Trp</i>	<i>Arg</i>	<i>Asp</i>	<i>Glu</i>	<i>Ser</i>	<i>His</i>

Fig. 6. The first 15 random variant codes found with a lower MS0 (= WMS0 at a transition bias of 1) than that of the natural code. These codes were used in further tests which explored their behavior under increasing transition bias (Fig. 5 and its legend).

Table 2. Quantification of translational errors used to measure the relative efficiency of the natural genetic code in terms of mistranslation

	First base	Second base	Third base
Relative frequency	0.5	0.1	1
Transition weighting	2	5	1
Combined weighting			
For transitions	1	0.5	1
For transversions	0.5	0.1	1

it bears little similarity to the natural genetic code other than in its calculated tMS value (natural code tMS = 2.63, better code tMS = 2.61).

Discussion

Prior to the work presented here, use of the MS measure has provided strong evidence that natural selection has

shaped the genetic code to minimize the effects of mutation and mistranslation but has suggested that this adaptation is limited to codon positions 1 and 3.

It is widely accepted that bias exists in the rate of transition/transversion mutations (i.e., that mutations $C \leftrightarrow T$ and $A \leftrightarrow G$ occur more frequently than mutations $C, U \leftrightarrow A, G$) (see, e.g., Fitch 1967; Kimura 1983). If this bias is incorporated into our calculations of the relative efficiency of the code (transforming a MS measure of the effect of measure/translation errors into a WMS measure), then the overall effect of a mild bias (up to the point where transitions are considered approximately three times more likely than transversions) is to increase the relative efficiency of the code and, at a higher bias, to decrease the relative efficiency of the code. This overall effect may be partitioned into different effects at each of the three codon positions: the first base mirrors the overall effect, increasing in relative efficiency up to a bias of 3 and decreasing in relative efficiency thereafter; the

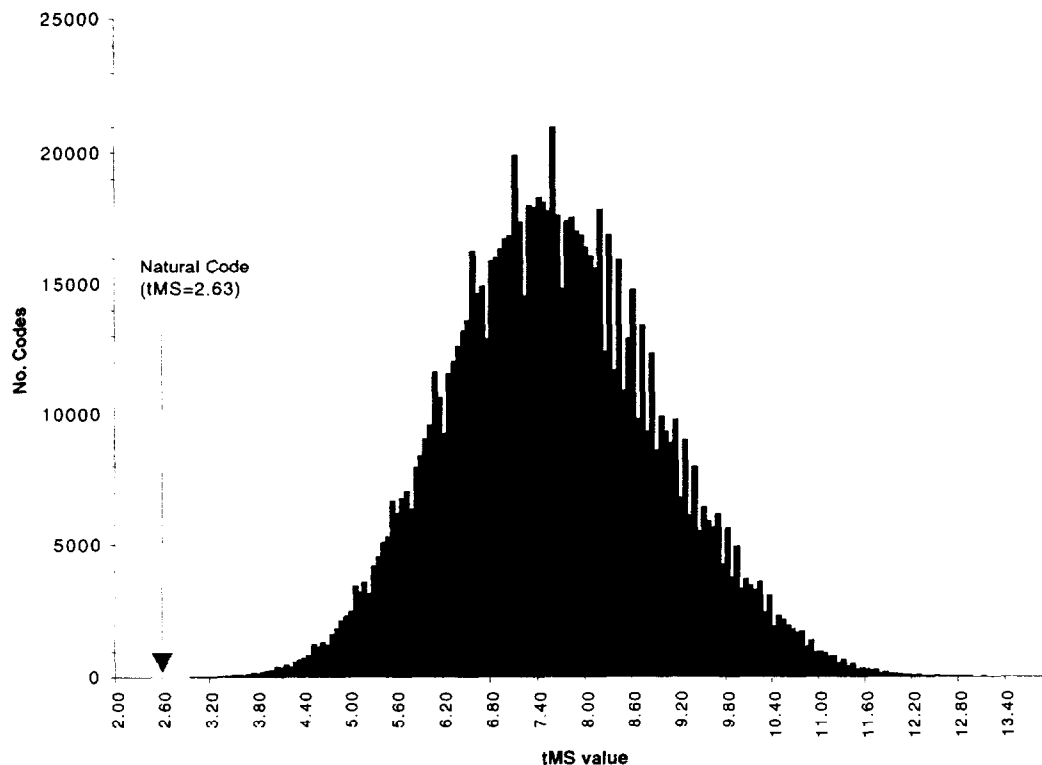


Fig. 7. Frequency distribution for the tMSO (=MSO adjusted for mistranslation parameters) values obtained from 1 million randomly generated variants of the natural genetic code. The X axis gives a particular range of categories of MS values, and the Y axis gives the number of random variant codes generated with an MS value in that category (from a sample of 1 million random variant codes tested). In addition, the arrow indicates the category into which the tMSO calculation for the natural code falls: the cumulative frequency to the left of this arrow therefore indicates the proportion of more conservative codes found among the random variants. This cumulative frequency is in fact 1 (i.e., only 1 of the 1 million variants had a lower tMS value), indicating that under our quantification of mistranslation parameters, the probability of a code as efficient as or more efficient than the natural code evolving by chance alone is 0.000001.

Table 3. tMS values calculated for the natural code and for a sample of 1 million random variants

Natural code tMS value	Sample of 1 million random variants		
	Mean	SD	Number of better codes found
2.63	7.63	1.35	1

third base is affected very little by changes in bias; the second base, however, shows a dramatic, significant, and consistent improvement with increasing bias. The relative efficiency of the second base does not peak at any intermediate bias but does appear to asymptote at an improvement of an order of magnitude. One might reasonably suppose that as bases 1 and 3 are well adapted even in the absence of any mutational bias, it is only base 2 that could show any significant improvement. While probably true, it is unexpected that base 2 should show such a dramatic improvement with even mild transition-transversion bias.

In addition to these results, close examination of variant codes which would have appeared "better" under Haig and Hurst's (1991) original model (i.e., produced a lower MSO value than the natural code) show them mostly to behave in the opposite manner to the real code

Ile	Ala	Gln	His	Phe	Ser	Tyr	Cys
Cys		TER	TER	Leu		TER	TER
			Gly				Trp
Cys	Leu	Thr	Ser	Leu	Pro	His	Arg
		Phe				Gln	
Trp	Pro	Asp	Ala	Ile	Thr	Asn	Ser
Val		Glu	Ser	Met		Lys	Arg
Tyr	Met	Asn	Arg	Val	Ala	Asp	Gly
		Lys				Glu	

Fig. 8. The single random variant code (of a sample of 1 million) found to have a lower tMS value than the natural code. The natural genetic code is shown on the right; the single "better" variant is shown on the left.

when transition bias is applied. Of 15 better codes tested, only 1 consistently outperforms the natural code as transition bias is increased. In other words, there is good reason to suspect that the observed behavior of the natural code really does represent a biologically important feature.

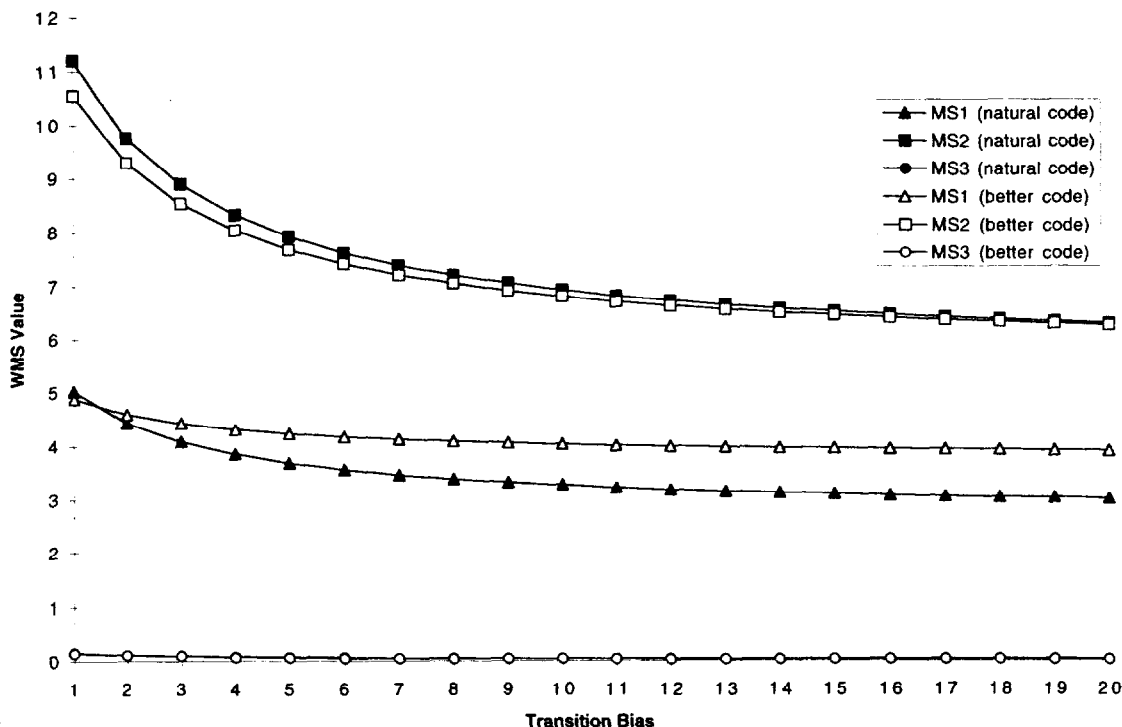


Fig. 9. The single random variant code found (of a sample of 1 million) compared to the natural code in terms of WMS measures calculated for each base at each of 20 transition weightings.

While it is thus tempting to view the code as having been shaped by natural selection to minimize the effects of point mutation, there is a subtly different possibility. The observed behaviors of the three bases under increasing transition bias exactly match the apparent order of translation efficiency. Limited empirical data for the relative rates and bias in translational errors at the three codon positions may be mapped into MS calculations for the natural code in order to estimate its relative efficiency in terms of translational efficiency. Under this model, the relative efficiency of the natural code increases another two orders of magnitude: of a sample of 1 million random variant codes generated, only 1 variant code was found to be of greater efficiency under these criteria. It is necessary to caution, however, that the data on mistranslational biases are limited, and while the result is remarkable, it would be valuable to provide further analysis incorporating better mistranslational bias data.

The results, then, indicate that the code appears to be very well structured to minimize the effects of mistranslation and point mutation and that biases in these processes are reflected in the structure of the code. This could be because the code evolved in a world in which these biases were found. Alternatively, once the code had evolved, selection might have favored the biases in the processes (a midground of coevolution between the biases and the codes is also possible). We consider the first explanation the most likely, as both biases are most probably simple chemical consequences of the processes involved. If natural selection has indeed acted to shape the

natural genetic code, did it act on mutational effects, mistranslational effects, or both?

The 50-fold improvement in estimated relative efficiency of the code from 1 in 20,000 better variants under WMS0 at a transition bias of 3 to 1 in 1,000,000 better variants under tMS calculations seems rather too high to attribute to chance. Hence a role for translation seems likely. Comparably, the fact that at the third site all codons pairs are transition pairs suggests a role for mutational biases. The natural genetic code contains six "family boxes" which comprise two transition pairs of codon meanings (e.g., UUU, UUC: Gly; UUA, UUG: Leu) and one family box which contains one transition pair and two individual codons (UGU, UGC: Cys; UGA: Ter; UGG: Trp). It contains no family boxes which comprise transversion pairs. For each family box, the probability of assigning 4 codon meanings into 2 transition pairs is $1/3$ (there are $4!$ ways of assigning 4 meanings and 8 ways of forming 2 transition pairs or 1 transition pair and 2 individual meanings). The probability of obtaining the distribution of transition pairs observed in the natural genetic code is therefore $(1/3)^7 = 0.00046$.

In spite of the above we have one reason to suspect that translational biases rather than mutational biases might have constituted the more important selective force. The single better code (in terms of tMS) shows behavior very similar to that of the natural code when tested under general transition bias (Fig. 9), while the "best" code under general transition bias (code 09, Fig. 6) gives a tMS value of 3.85 for which our sample of 1

million random variants provides 872 better codes (i.e., code 09 is two orders of magnitude less relatively efficient than the natural code in terms of mistranslation). In other words, our evidence suggests that codes as good as (or better than) the natural code in terms of minimizing the effects of mistranslation may automatically behave as well as the natural genetic code in terms of minimizing the effects of point mutation but that the reverse is not necessarily true. The observed behavior of each base under general transition bias may then be a side effect of approaching optimality for minimizing the effects of translational errors, as Woese (1965, 1973; Woese et al. 1966) originally suggested.

Acknowledgments. We should like to thank Gilean McVean, Nick Goldman, John Barrett, and Anne Oakenfull for helpful discussion of aspects of this work. The manuscript was improved by comments from two anonymous referees.

References

- Alf-Steinberger C (1969) The genetic code and error transmission. *Proc Natl Acad Sci USA* 64:584–591
- Amirnovin R (1997) An analysis of the metabolic theory of the origin of the genetic code. *J Mol Evol* 44:473–476
- Collins DW (1994) Rates of transition and transversion in coding sequences since the human-rodent divergence. *Genomics* 20:386–396
- Di Giulio M (1989) The extension reached by the minimization of the polarity distances during the evolution of the genetic code. *J Mol Evol* 29:288–293
- Di Giulio M (1997) On the origin of the genetic code. *J Theor Biol* 187:573–581
- Epstein CJ (1966) Role of the amino acid “code” and of selection for conformation in the evolution of proteins. *Nature* 210:25–28
- Fitch WM (1967) Evidence suggesting a non-random character to nucleotide replacements in naturally occurring mutations. *J Mol Biol* 26:499–507
- Friedman SM, Weinstein IB (1964) Lack of fidelity in the translation of ribopolynucleotides. *Proc Natl Acad Sci USA* 52:988–996
- Goldberg AL, Wittes RE (1966) Genetic code: aspects of organisation. *Science* 153:420–424
- Goldman N (1993) Further results on error minimization in the genetic code. *J Mol Evol* 37:662–664
- Haig D, Hurst LD (1991) A quantitative measure of error minimization in the genetic code. *J Mol Evol* 33:412–417
- Kimura M (1983) *The neutral theory of evolution*. Cambridge University Press, Cambridge.
- Kumar S (1996) Patterns of nucleotide substitution in mitochondrial protein-coding genes of vertebrates. *Genetics* 143:537–548
- Moriyama EN, Powell JR (1997) Synonymous substitution rates in *Drosophila*: Mitochondrial versus nuclear genes. *J Mol Evol* 45:378–391
- Morton BR (1995) Neighbouring base composition and transversion transition bias in a comparison of rice and maize chloroplast non-coding regions. *Proc Natl Acad Sci USA* 92:9717–9721.
- Parker J (1989) Errors and alternatives in reading the universal genetic code. *Microbiol Rev* 55:273–298
- Sonneborn TM (1965) *Degeneracy of the genetic code: extent, nature and genetic implications*. Academic Press, New York.
- 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 FJR, Coates D (1989) The code within the codons. *Bio Systems* 22:177–187
- 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 Harbour Symp Quant Biol* 31:723–736
- Wong JT-F (1975) A co-evolution theory of the genetic code. *Proc Natl Acad Sci USA* 72:1909–1912