THE CASE FOR AN ERROR MINIMIZING STANDARD GENETIC CODE

STEPHEN J. FREELAND*, TAO WU and NICK KEULMANN
Department of Biology, University of Maryland, Baltimore County, Catonsville, MD, U.S.A.
(author for correspondence, e-mail: freeland@umbc.edu)*

(Received 10 June 2002; accepted in revised form 31 October 2002)

Abstract. Since discovering the pattern by which amino acids are assigned to codons within the standard genetic code, investigators have explored the idea that natural selection placed biochemically similar amino acids near to one another in coding space so as to minimize the impact of mutations and/or mistranslations. The analytical evidence to support this theory has grown in sophistication and strength over the years, and counterclaims questioning its plausibility and quantitative support have yet to transcend some significant weaknesses in their approach. These weaknesses are illustrated here by means of a simple simulation model for adaptive genetic code evolution. There remain ill explored facets of the 'error minimizing' code hypothesis, however, including the mechanism and pathway by which an adaptive pattern of codon assignments emerged, the extent to which natural selection created synonym redundancy, its role in shaping the amino acid and nucleotide languages, and even the correct interpretation of the adaptive codon assignment pattern: these represent fertile areas for future research.

Keywords: adaptation, error minimization, evolution, Genetic Code, natural selection

1. Introduction

The evolution of the genetic code ranks among the most significant transitions in the history of life on earth (Szathmary *et al.*, 1995). Before biological coding emerged, the hypothesized RNA world, in which a single biopolymer performed both as genetic information and metabolically functional phenotype, forms one of the best supported models for the early evolution of life (Gesteland *et al.*, 1993; Gesteland *et al.*, 1999). Since that time, not only have all organisms split into a dichotomy of nucleic acid genotype and protein phenotype, but the code that bridges this dichotomy has influenced the general process of molecular evolution. For example, the pattern of codon assignments within the code defines the relative frequencies with which amino acids interconvert as the result of random nucleotide substitutions in protein coding genes (Fitch, 1966a, b) and the redundancy with which each is encoded correlates well with the composition of the proteome (King *et al.*, 1969; Knight *et al.*, 2001c). Though natural selection may distort these underlying patterns, it operates upon translated protein phenotypes. Thus the genetic code determines what natural selection 'sees' and the precise coding rules of a genome influence both the neutral and adaptive evolution that occur within it.

Origins of Life and Evolution of the Biosphere **33:** 457–477, 2003.

Understandably then, the topic of how and why the genetic code itself evolved has received considerable attention from the evolutionary research community. However, the general topic of 'genetic code evolution' comprises at least 3 potentially independent lines of enquiry. First, we may enquire how and why did genetic coding originate (e.g. Szathmáry, 1999; Knight *et al*., 2000a)? Second, how and why did the standard genetic code emerge prior to the last universal ancestor of modern life (e.g. Di Giulio 1997; Knight *et al*., 1999)? Third, how and why has this standard code subsequently diverged in numerous distinct lineages (e.g. Osawa, 1995; Knight *et al*., 2001a)? Although suggested syntheses of answers to all of these questions periodically surface (e.g. Davis, 1999) they must, at present, rely on anecdotal evidence and speculation to navigate through more gray areas than is comfortable for scientific enquiry, given that each sub-topic comprises very different theories that continue to maneuver relative to one another in the face of emerging evidence (e.g. see Hartman, 1975, 1978, 1984, 1995a, b).

In particular, the past decade has seen significant maturation of explanations as to why specific amino acids were assigned to specific codons within the standard genetic code, producing the pattern that links primordial evolutionary events to the general molecular evolutionary dynamics of the biosphere. Mainstream theory has coalesced around 3 major themes (Knight *et al*., 1999): (1) codon assignments reflect the RNA/amino acid direct stereic interactions from which it arose (Sowerby *et al*., 1998; Knight *et al*., 2000a; Yarus, 2000), (2) the amino acid (e.g. Trifonov *et al*., 1997; Di Giulio, 1998; Trifonov, 2000) and perhaps nucleotide (Crick, 1968; Eigen, 1971; Shepherd, 1981; Baumann *et al*., 1993; Jimenez-Sanchez, 1995; Trifonov *et al*., 1997) alphabet have grown in complexity since the origin of coding, and (3) the pattern of amino acid assignments within the standard code exhibits adaptive features that were produced by natural selection. At a general level, growing evidence for each theory renders it increasingly unlikely that they represent competing alternatives (Woese *et al*., 1966; Crick, 1968), but rather that all contributed to the emergence of the standard genetic code (Di Giulio, 1997; Knight *et al*., 1999). At the edges of this debate, entirely novel explanations for its form continue to emerge but tend to be so entirely divorced from a recognized biological context, whether through mathematical abstraction (e.g. Bashford *et al*., 1998) or speculation (e.g. Davydov, 1996, 1998) as to contribute little to the debate.

Among these major themes, the nature and extent of natural selection as a determinant of codon assignments is critical to our overall interpretation of patterns within the standard genetic code. For example, the simplest synthetic theory assumes a straightforward succession of evolutionary factors: biological coding originated through stereochemical interactions, the primordial code then gained new amino acids as primitive metabolism biosynthesized novel, adaptive amino acids and finally random variations to coding rules were filtered by natural selection to produce an 'error minimizing' pattern of codon assignments. Under this model, selection operates independently from the previous evolutionary forces, potentially overwriting the footprints of stereochemical origins and biosynthetically mediated

code expansion. As a consequence of assuming this model, the strength of natural selection has been gauged by the presence or absence of biosynthetic patterns within the code (Di Giulio, 1998; Freeland *et al*., 1998) and by the precise level of optimization displayed by the standard code relative to theoretical alternatives (Wong, 1980; Goldman, 1993; Di Giulio, 1989, 1991, 2000; Di Giulio *et al*., 1994; Judson *et al*., 1999). Indeed, the detailed evidence for each putative evolutionary factor remains hotly debated (e.g. Ellington *et al*., 2000 vs. Knight *et al*., 2000; Di Giulio, 1999b vs. Szathmáry, 1999; Di Giulio, 1999a vs. Amirnovin, 1997; Di Giulio, 2001a vs. Freeland *et al*., 2000; Di Giulio, 2001b vs. Ronneberg *et al*., 2000), often focusing on the validity of explicit quantitative claims that have been advanced to support each theory. However, the recent reviews of the general case for stereochemical patterns (Knight and Landweber, 2000; Yarus, 2000), and biosynthetic patterns (Di Giulio, 1998), have yet to be mirrored for the error minimizing genetic code.

In this context, our purpose here is to review the history and growth of the evidence for an adaptive arrangement of codon assignments within the standard genetic code, explore the validity of key criticisms, and highlight alternative, equally parsimonious interpretations of current evidence that have been largely overlooked, in the hope of stimulating debate into new productive territory.

2. Early Evidence for an Error Minimizing Genetic Code

Even before the full details of the standard genetic code were formally presented (Frisch, 1966), researchers began to report some decidedly non-random characteristics. In particular, attention in some quarters focused on the 'block' structure of codon assignments: for every amino acid (except, to some extent, Serine) no other pattern of amino acid assignments would present a more connected set of synonymous codons. This prompted two independent proposals (Sonneborn, 1965; Zuckerkandl *et al*., 1965) that the codon assignments of the standard code reflect an adaptive outcome of natural selection for 'error minimization'. Both hypotheses limited themselves to the argument that 'better' codes are those in which a higher proportion of random nucleotide substitutions result in no change to amino acid meaning. Thus, long before formal recognition and analysis of problem of 'error catastrophe' within unsophisticated replicators (Eigen, 1971; Eigen *et al*., 1979; Kauffman, 1993), the underlying concept of error-limited primordial evolution was partially addressed by these hypotheses.

However, Crick's (1966) subsequent 'wobble hypothesis', which invoked nothing more than stereochemical constraints on a tRNA's ability to discriminate codons, offered a simpler and more intuitive explanation for the same phenomenon and, ever since, the contiguity of synonymous codons has been largely ignored except by non-biologists (e.g. Cullman *et al*., 1983, 1987; Figureau *et al*., 1984, 1987, 1989). Instead, the concept of error minimization was taken to a new level of soph-

istication by Woese's (1965) seminal observation that *different* amino acids with similar biochemical characteristics were assigned to codons connected in mutation space. In particular, he observed that the hydrophobicity of an amino acid's side chain correlates well with its position within the code, a property that would intuitively form a key role in the process by which peptide chains fold into 3-dimensional structures with sophisticated catalytic activities within the watery cytoplasm of the cell (see Pace *et al*., 1996; Tomii *et al*., 1996 for recent corroboration of this point). Woese went further, observing (qualitatively) that the spatial/biochemical correlations between coding space and amino acid hydrophobicity were consistent with existing research into general patterns of translational error (Woese, 1965, 1973). For example, mistranslation events were found to occur with highest frequency at the third codon position (Davies *et al*., 1964; Friedman *et al*., 1964), where similarities in coded amino acid hydrophobicity are most pronounced. It is noteworthy that since the 1960's, these pioneering, general studies of *in vitro* mistranslation have been almost completely replaced by context specific studies of *in vivo* mistranslation (e.g. Parker, 1989; Stahl *et al*., 2002): it would be of great significance to our understanding of genetic code evolution if modern technologies were used to replicate and extend this exploratory 1960's analysis of artificially magnified mistranslation patterns.

In several restricted forms, the general idea reported by Woese (1965) was explored quantitatively using early computational methods (Volkenstein, 1965; Epstein, 1966; Goldberg *et al*., 1966). Of particular note, Alff-Steinberger (1969) performed one of the earliest biological Monte Carlo simulations, comparing the standard code against 200 alternatives in terms of error minimization for several fundamental metrics of amino acid similarity. However, these studies ignored the subtleties of pattern inherent to translation that Woese had noted, assuming that all nucleotide interconversions are equally likely regardless of base identity or codon position. Meanwhile, the apparent universality of codon assignments across life paved the way for Crick's (1968) influential 'Frozen Accident' hypothesis. This asserted that changing any codon assignment would effectively introduce numerous, simultaneous errors throughout the genome of an organism such that the deleterious effects would far outweigh the advantage of any improvement in coding strategy. In other words the code could not have been optimized because it was incapable of variation. The beguiling simplicity of this argument complemented a history of ingenious theoretical predictions for the code that subsequently turned out to be utterly wrong, from direct templating codes (e.g. Gamow, 1954; Gamow *et al*., 1955) to reading-frame independent 'comma-less' codes (Crick *et al*., 1957: see Hayes (1998) for a thorough introduction), feeding skepticism about interpreting apparent patterns within the standard code.

Thus, during the following decade, it was the very different idea of biosynthetic relationships between amino acids reflected in the standard code that rekindled interest in a possible significance to the pattern of codon assignments. Dillon's (1973) seminal work, though incorrect in biochemical detail, paved the way for

Wong's (1975, 1976, 1980, 1981, 1988; Wong *et al*., 1979) much more thorough development of this 'coevolution' hypothesis, and adaptive analyses assumed a rather low profile.

Eventually the credibility of the 'frozen accident' argument was shaken by the discovery of a non-standard genetic code within vertebrate mitochondria (Barrell *et al*., 1979). Some initial reactions sought to downplay the significance of this find, interpreting this aberrant code as either a genetic 'fossil' of an ancestral code from which the standard code subsequently evolved (Jukes, 1981; Grivell, 1986), or as a unique consequence of the fact that vertebrate mitochondria encoded so few protein products (Hasegawa *et al*., 1980). Gradually, however, the steady discovery of new variant genetic codes (see Knight *et al*., 2001a, for a review), all secondarily derived from the standard code, has shown that even within the large, complex genomes of extant organisms, codon assignments are flexible. Though the causes, and some of the mechanisms for such changes remain imperfectly understood (see Knight *et al*., 2001a, b) the message that codon assignments can vary over evolutionary time is clear.

3. Recent Quantitative Support for an 'Error Minimizing' Standard Genetic Code

The knowledge that codon assignments are evolutionarily flexible lies a long way from demonstrating that the standard genetic code is a product of adaptive evolution. In particular, it remains to formally render the concept of an 'error minimizing' genetic code as a testable hypothesis. Several novel analyses attempted to do just this during the 1980's but all introduced unnecessary new assumptions: claims for a Baudot genetic code (Cullman and Labouygues, 1983, 1987) or a Gray genetic code (Swanson, 1984) both hinted at an adaptive pattern of codon assignments, but both relied on methodologies brought from abstract coding theory with elements that lie beyond an obvious biological relevance (e.g. that the similarity of coded objects, in this case amino acids, is best measured on a ring rather than as a simple linear metric).

The first step towards applying modern computing power to a straightforward quantitative test of Woese's (1965) qualitative claims came later (Haig *et al*., 1991). The methodology was essentially an extension of Alff-Steinberger's (1969) Monte Carlo approach*, a simple 3-step process: first quantify a given code's susceptibility to errors (i.e. mutations); second, define a set of possible code structures of which the standard genetic code is one example; third, generate a large sample of codes

* The results of Haig *et al*. (1991), however, are at odds with those of Alff Steinberger (1969). As the latter's results have proved irreproducible, both by us and others (Knight and Burch, pers. comm.), whereas those of Haig and Hurst have been verified by us and others (Burch, pers. Comm., but see Haig *et al*., 1999), it appears that Alff Steinberger contributed the idea and the method, but not the first results in this lineage of analysis.

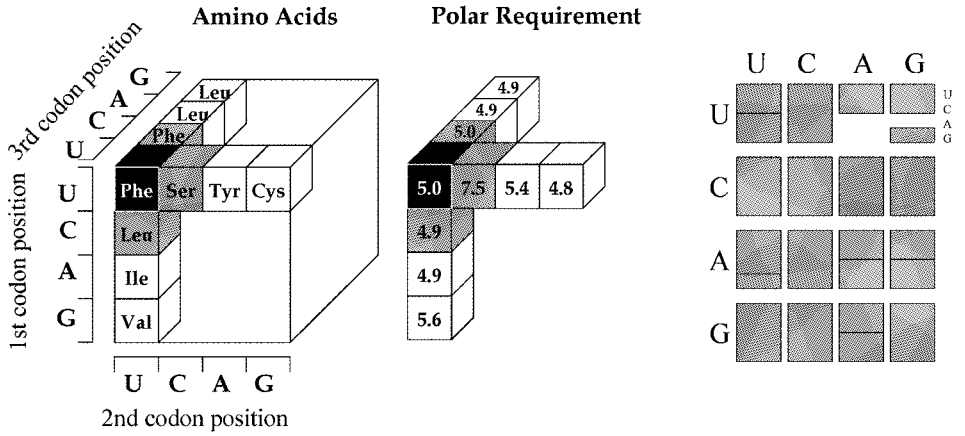

Figure 1. Calculating the error value of a code: (a) Quantifying errors: first, take a quantitative index that describes some physiochemical property of each amino acid (in this case, we use values of Polar Requirement, a measure of hydrophobicity). Then take a codon (here, UUU = Phe) and a possible alternative meaning that can be reached by mutating or misreading a single nucleotide (e.g. UCU = Ser). Hence quantify the difference between actual and intended amino acid meanings (e.g. Phe→Ser corresponds to $|5.0 - 7.5| = 2.5$) for the error in question. Repeat this process for all possible single nucleotide errors, and then for all codons of the code under scrutiny. Divide this total summed error value by the number individual pair-differences from which it is comprised to produce Δ_{code}, the code's error value. This is a quantification the *average* magnitude associated with single nucleotide errors to organisms that use this code. However, not all of the possible errors occur with equal frequency. In particular, the unequal chemical similarity of the 4 nucleotides to one another means that transition errors (U↔C, A↔G: codons that lie a transition error away from codon UUU are shown here in gray) occur more frequently than transversions (C *or* U↔A *or* G). This can be incorporated into calculautions of Δ_{code} by giving an arbitrary weighting $W(>1)$ to differences caused by transition errors (i.e. UUU:Phe→UCU:Ser would receive this weighting factor) when calculating individual errors prior to summation. A final level of sophistication may be added by giving an additional weighting to individual errors according to whether they correspond to changes in the first, second or third codon position. A strong base effect has been noted for translational errors. (b) Defining a set of possible codes. For most Monte Carlo analyses of the adaptive code, 'possible codes' are defined as those which maintain the pattern of synonymous coding (i.e. redundancy) found within the standard genetic code. This conservative restriction controls for possible biochemical restrictions on code variation, and biases analysis against a false positive result for the adaptive hypothesis (see text for detailed discussion).

that meet this definition, and find where the standard code lies relative to the rest of the sample according to this measurement of error susceptibility.

For the first step, Haig *et al.* (1991) took several quantitative measures of amino acid similarity and used each of them to calculate the arithmetic mean change of all pairs of actual and intended amino acids that can result from single nucleotide errors within a codon (Figure 1a). Extending this process to all the codons of a code produces a quantitative measure of a code's susceptibility to error (its 'error value', Δ_{code}). For the second step, assuming the 4 bases, 64 codons, 20 amino acids and the (imperfectly understood) pattern of redundancy found within the standard code

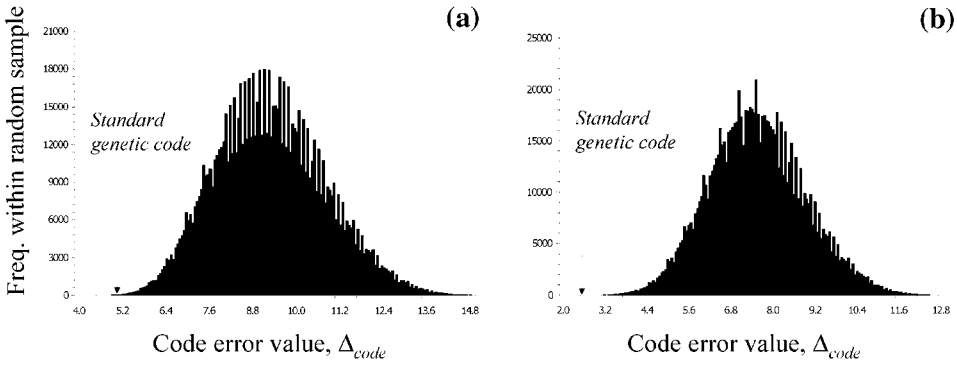

Figure 2. A comparison of Δ_{code} for the standard genetic code with equivalent values for 1 million randomly generated codes: (a) Without any transition bias or codon-position weighting, measuring amino acid similarity in terms of Polar Requirement (a measure of hydrophobicity), a proportion of 0.0001 codes within the random sample provide a lower Δ_{code} value than the standard genetic code (Haig *et al.*, 1992); (b) Mapping experimentally determined patterns of mistranslation onto Δ_{code} reduces this figure by 2 orders of magnitude (Freeland *et al.*, 1998).

as fixed, they defined a set of possible codes that comprises 20! or 2.4×10^{18} members: each is a possible 1:1 mapping of the 20 amino acids to the 20 synonymous coding blocks (Figure 1b). For the third step, they generated a sample of 10 000 random genetic codes and found that for one measure of amino acid hydrophobicity ('Polar Requirement', Woese *et al.*, 1966) only 2 random codes provided a smaller average change in amino acids (Figure 2a) than the standard code. Statistically significant (though quantitatively weaker) support for an 'error minimizing' code was reported for other measures of amino acid hydrophobicity, but not for other chemical properties of the amino acids such as charge and size; (but see Haig *et al.*, 1999). These findings complemented various multivariate statistical analyses that had sought correlations between physiochemical measures of amino acid similarity and codon assignments (e.g. Szathmary *et al.*, 1992).

Analysis was then extended to incorporate biological biases that are known to influence patterns of mutation (Ardell, 1998) and finally mistranslation (Freeland *et al.*, 1998). Using polarity to measure amino acid similarity, the perceived optimality of the code increases 1 order of magnitude when calculations are adjusted to incorporate a reasonable transition bias, as justified by both biochemical first principles (Topal *et al.*, 1976) and wherever nucleotide substitution patterns have been estimated for real sequence data (Wakeley, 1994; Petrov *et al.*, 1999), and a further order of magnitude when they are adjusted to reflect the fact that the three positions of a codon are not misread with equal frequency: the first quantification of Woese's (1965) observations revealed that the standard genetic code outperformed 999,999 out of a sample of 1 million random alternatives (Freeland *et al.*, 1998) (Figure 2b).

Complementary analyses revealed that these results are qualitatively robust to possible confounding factors of biosynthetic relatedness (Freeland *et al*., 1998; but see Di Giulio *et al*., 2001) and methodological variation (Freeland *et al*., 2000). Since then, further insights have continued to build the sophistication of the argument: for example, the amino acid composition of the proteome enhances the perceived adaptation of the code, and this is especially pronounced when multidimensional measures of amino acid similarity are derived to replace polarity (Gilis *et al*., 2001). One of the most interesting refinements is the finding that the first- and second-position nucleotides of a codon have a roughly consistent, additive effect on several amino acid properties, smoothing the connectivity of the coded protein landscape (Aita *et al*., 2000) perhaps rendering natural selection's 'search' of protein phenotype more efficient.

However, all this evidence may be considered 'top-down' approach to testing the error minimizing hypothesis (Szathmáry *et al*., 1995): it asks simply whether the pattern of codon assignments seen in the standard genetic code meets the expectations of selection for error minimization. Far less research has been carried out to assess the mechanisms and pathways that would lead to an adaptive organization of codon assignments from a random (or sterically determined) starting point. Szathmáry (1991) produced a toy model of code evolution based on the 'codon capture' (Osawa *et al*., 1989) model of reassignment. He showed that under oscillating GC/AT mutational biases, any two codon identities that are adjacent within the code can swap, lending mechanistic credibility to the general adaptive claims. More recently, Ardell and Sella (Ardell *et al*., 2001; Sella *et al*., 2002) have begun to develop a far more sophisticated model of adaptive code evolution: their unique contribution has been to model the interaction (coevolution) of codes with associated genomes, recognizing the potential differences between selection on a genetic code as opposed to other phenotypic traits. In particular, theirs is the first mechanistic model to incorporate Crick's (1968) objection that any codon reassignment event introduces potentially profound disruption on the existing genome. Though their model has yet to expand to a full, 64 codon representation, already they have shown that significant code optimization does take place, and that key features of the standard genetic code are reliably produced in all simulations. Clearly, this sort of model presents an excellent approach to further analysis of the codon assignment evolution, whether inferred (for the standard genetic code) or observed (for secondary code variation, e.g. within metazoan mitochondria).

4. Objections to the Error Minimizing Code

Against this evidence, one lineage of critical analysis has repeatedly disputed that the standard genetic code is adapted to minimize the effects of genetic errors, asserting that selection has been a weak and minor factor in steering codon assignments. The most frequent form of this criticism has been the repeated use of

analytical calculus (Wong, 1980; Di Giulio, 1989) or powerful computer search algorithms (Di Giulio, 1991, 1999, 2000; Di Giulio *et al*., 1994; Goldman, 1993; Judson *et al*., 2000) to demonstrate the existence of theoretical codes that would minimize errors to a significantly greater extent than the standard genetic code. Problematically, however, all such studies have consistently failed to address the peculiar susceptibility of powerful computer searches to the GIGO (garbage-in-garbage-out) computing principal (Freeland *et al*., 2000b): computer predictions for an optimal code (or anything else) are necessarily limited to the optimization criteria provided by the programmer, and even subtle changes to these criteria can lead to very different predictions. By analogy, a program asked to produce an optimal design for a fuel efficient airliner is unlikely to design something with space for a hundred passengers with luggage if this is not made an explicit requirement. In terms of genetic code optimization, there are two key vulnerabilities: the quantification of amino acid similarity, and the assumed model by which mutations (or mistranslations) occur.

The evolutionary similarity of amino acids (meaning their substitutability within proteins) is unlikely to be perfectly represented by a single physiochemical measure (e.g. Polar Requirement) or indeed by any simple combination of two or three such indices. Instead we have every reason to expect that amino acid similarity is a multidimensional concept that remains far from fully understood: it may well be a partly relative phenomenon, dependent on the precise sequence of amino acids within a protein, rendering the concept of an 'averaged' similarity somewhat fuzzy. Attempts to obviate this problem by measuring similarity directly from estimates of the frequencies with which amino acids substitute for one another within real proteins do suggest the code to be close to a global optimum (Ardell, 1998; Freeland *et al*., 2000a), but have been criticized as tautologous given a correlation between the code and patterns of substitution (Di Giulio, 2001b), though the flow of causality in this correlation is yet to be determined. Encouragingly, a recent attempt to derive a multidimensional measure of amino acid similarity that is truly independent from the code supports the counter criticism: the more sophisticated our representation of similarity, the better the code appears (Gilis *et al*., 2001).

On a related theme, almost all criticisms of the adaptive theory have assumed the pre-Woese (1965) idea that all nucleotides mutate/misread for one another with equal frequency despite plentiful, well documented evidence to the contrary (e.g. Wakeley, 1994; Petrov *et al*., 1999; Topal *et al*., 1976). In fact, the only study to incorporate even a straightforward transition bias (Di Giulio *et al*., 2001) also introduced a set of highly contentious (Amirnovin, 1997; Amirnovin *et al*., 1999; Di Giulio, 1999; Di Giulio *et al*., 2000, 2001; Ronneberg *et al*., 2000) biosynthetically derived restrictions on possible codon assignments, hindering a straightforward interpretation of results. Thus, estimates of the standard genetic code's optimization relative to engineered 'perfect' codes that are quoted as exact figures to tenths of a percentage point (Di Giulio, 1989, 1991, 2000; Di Giulio *et al*., 1994, 1999), are misleading in their apparent accuracy (Figure 3): not even the existence of biologic-

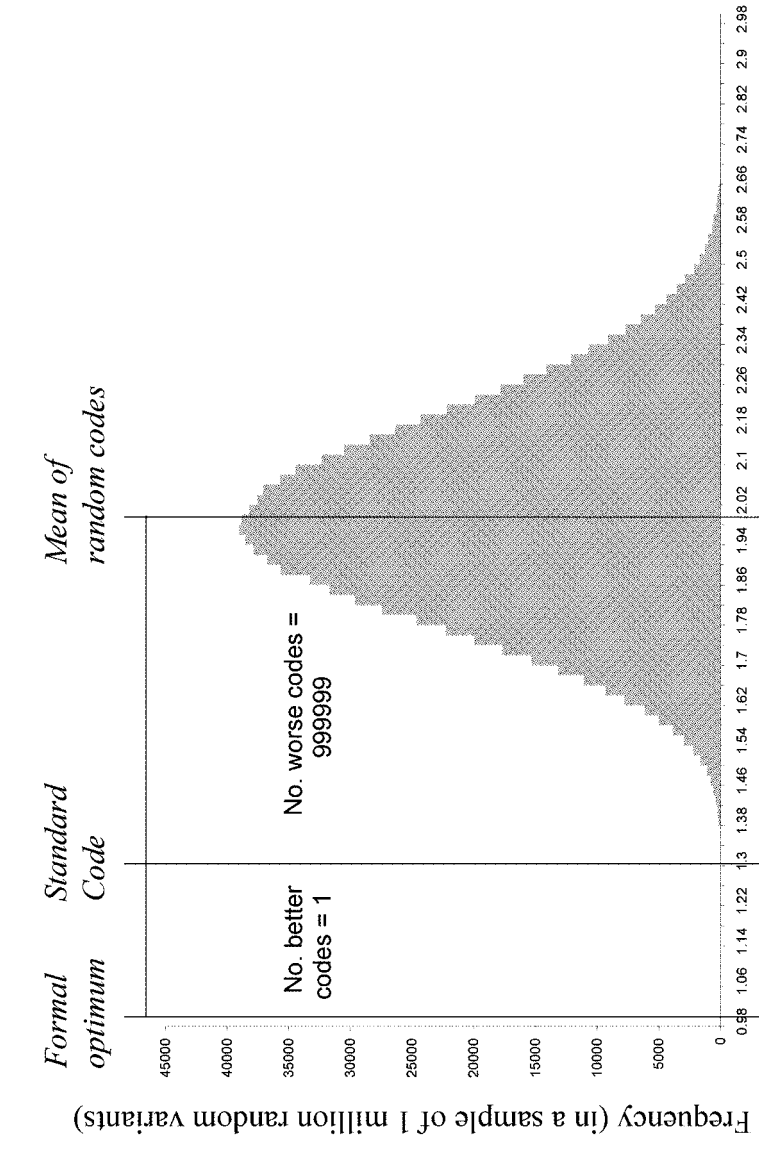

Figure 3. Misleading estimates of code optimality. Measuring the Δ_{code} for the standard genetic code on a linear scale bounded by a theoretical global optimum and the mean error value of random codes ('percentage distance minimization'), gives very low estimates of code optimization. This is misleading given that (1) we do not fully understand the criteria under which the code was optimized; (2) we know very little about the connectedness of codes with different error values; and (3) natural selection is likely to approach the optimum asymptotically: the bell-shaped distribution of possible codes' error values indicates that the better adapted a code became, the slower further improvement would be.

ally meaningful better codes has truly been demonstrated conclusively. In contrast, the results derived from Monte Carlo simulations are qualitatively robust to fine-tuning of the index of amino acid similarity, given that the standard genetic code lies at the edge of the Gaussian distribution of possible levels of error minimization (Freeland *et al*., 2000a, b).

A separate and more subtle problem with criticisms based on the existence of far better codes is that, even assuming they are what the programmer considers them to be, none have been rigorously investigated in terms of evolutionary accessibility. This is of key significance because arguments for an adaptive pattern of codon assignments is about an evolutionary, rather than 'engineering', optimum (Di Giulio, 1989; Freeland *et al*., 2000a, b). Indeed, the search algorithms used to identify optimal theoretical codes are deliberately designed to surpass the blind, stepwise (and asymptotic) process of improvement by which natural selection is thought to proceed (e.g. see Di Giulio *et al*., 1994; Judson and Haydon, 1999). The only critical study to address this issue by directly simulating code evolution (Di Giulio *et al*., 2001) also incorporated an unusual tolerance for maladaptive evolutionary steps, a different measure of amino acid similarity and a mutational mechanism of 'block reassignments' to produce a model that the authors themselves described as 'totally unrealistic' (Di Giulio *et al*., 2001, p. 379), placing something of a question mark over the results' precise relevance to general claims for an adaptive pattern of codon assignments.

To illustrate the significance of these counter criticisms, we have developed a simple computer simulation of code evolution. The simulation takes place over successive, discreet generations. Each generation begins with a 'parent' genetic code (at the start of the simulation, this parent is formed by randomly assigning amino acid meanings to codons). The program creates 10 'offspring' genetic codes from this parent, each offspring differing by 1 random mutation (a single codon reassignment) from the 'parent'. Associated Δ_{code} values are then calculated, as described in Figure 1, for each of these offspring codes. The parent code (and associated Δ_{code} value) is then added to create a set of 11 codes, from which the program selects the code with the lowest associated Δ_{code} value (or, failing the emergence of an adaptive mutation, the parent is retained). This 'fittest' offspring then becomes the parent for the next generation. The program stops when it achieves a given threshold value for Δ_{code}. The inclusion of the parent in each generation simply reflects the assumption that at no point in the ancestral lineage of the standard genetic code did no offspring possess an unmutated genetic code. Copies of this program are available on request from the corresponding author.

Results (Table I) show that when it is assumed that all nucleotides mutate/are misread for one another with equal probability, it is indeed not difficult to evolve from a random genetic code to one that displays the same degree of 'error minimization' as the standard code (specifically, in 1000 simulations, it took an average of 25 codon reassignments, and left an average of 42 out of 64 codons unchanged in their amino acid assignment). When a mild transition bias is introduced (consistent

TABLE I

Results of a simple simulation for adaptive evolution of genetic code codon assignments under a range of assumptions. Each simulation began with a random genetic code, and a target Δ_{code} value (representing the quantitative error minimization of the standard code under the assumptions of the simulation). For each set of assumptions, we record the number of generations taken to reach the target level of 'error minimization', and the number of codon assignments that remain unchanged in the course of this optimization. Assumptions reflect the calculation of Δ_{code} both for the standard genetic code (i.e. the target Δ_{code} value) and for the theoretical codes of the simulation, and are described in the text

Description	Number of generations		Number of codons unchanged	
	Mean	Std. Dev. (3sf)	Mean	Std. Dev. (3sf)
No transition bias	25	8.60	42	5.38
Transition bias of 2	33	8.37	37	7.68
Transition bias of 9	70	28.4	25	7.81
Mistranslation parameters	(most simulation runs fail to reach the target Δ_{code} value after 10 000 generations)			

with estimates from modern pseudogene sequence data, Petrov *et al*., 1999), the associated improvement in our perception of the standard code's adaptation increases this slightly to 33 reassignments and only 37 codons remain unchanged. When the transition bias is increased to meet the predictions from biochemical first principals (Topal *et al*., 1976), which might be construed as a better estimate for a primordial genome unfettered by sophisticated error checking/repair molecular machinery, it takes an average of 70 mutational steps to match the standard code's adaptation (though a surprisingly high average of 25 codon assignments remain unchanged from the start to the end of the simulation). Finally, when we introduce the crude estimates of mistranslation patterns into the code, including a base position effect for the codon/anticodon match (derived from Friedman *et al*., 1964, via Woese, 1965, as quantified by Freeland *et al*., 1998a), we find that after 10 000 generations, the program has usually failed to match the target error minimization value of the standard genetic code. We advance these results as unambiguously demonstrating that, beyond the problem of accurately measuring amino acid similarity, the mere existence of far 'better' genetic codes is of dubious relevance to arguments about the strength of selection acting on codon assignments of the standard genetic code.

5. Interpreting the Error Minimizing Genetic Code

Continued and somewhat repetitive debate over the quantitative details of an adaptive pattern of codon assignments within the standard genetic code has hindered development of a deeper debate over the interpretation of the general observation.

The results we present above, and most of this review have focused on the simplest interpretation: that amino acid assignments were shuffled according by natural selection to produce an error minimizing code, but this is by no means the only plausible explanation. Assuming that the general idea of biosynthetically mediated code expansion is correct, and only the detailed interpretation requires caution (Ronneberg *et al*., 2000), then there is no reason to preclude the possibility that novel amino acids were incorporated on an adaptive basis either in terms of their properties (for a fixed location in the code), or in terms of their location (for a fixed set of properties), or in some combination of both. This is certainly consistent with current evidence for an adaptive genetic code (Freeland *et al*., 1998b), and the recent simulation studies that assume one specific, controversial model of biosynthetic expansion (Di Giulio *et al*., 2001) could usefully be broadened to consider non-coded amino acids and a more flexible expansion pathway.

Other, subtler nuances have been offered for the error minimizing code: one is that the earliest genetic code was wholly or partially ambiguous (i.e. no one codon was assigned to any single amino acid), but that an ensuing positive feedback process saw increasingly sophisticated proteins permitting (and requiring) a code of greater accuracy (Fitch, 1966; Woese, 1973). This process would, it was argued, ultimately produce a code in which mutationally connected codons were assigned to amino acids with similar biochemical properties. Subsequent analysis of tRNA phylogeny concluded support for the theory (Fitch *et al*., 1987) but other studies of tRNA phylogeny have come to very different conclusions (e.g. Di Giulio, 1994), and it may be that tRNA's are too evolutionarily labile to use in this context (Knight *et al*., 1999). Interestingly, given this background, the recent code/genome coevolution simulations (Ardell *et al*., 2001; Sella *et al*., 2002) assume this ambiguous coding as a start point for code evolution. They find that a process of ambiguity reduction does indeed lead to an 'error minimizing' genetic code, though they also find that codon reassignments continue long after ambiguity has ended.

A stark alternative is the residual possibility that the standard genetic code could display adaptive properties without ever having been selected. In particular, recent experimental work has given new life to the old idea of direct templating between nucleic- and amino acids by demonstrating a surprising affinity between various amino acid side-chains and their associated anticodons/codons within the standard code (Yarus, 2000). Although the association is somewhat mystifying in a code with ancient adaptor tRNA's, the association appears statistically robust (Knight *et al*., 1998, 2000b; Ellington *et al*., 2000; but see Illangasekare, 2002) and the emphasis has moved to tentative theoretical explanations (see Szathmáry, 1999; Knight *et al*., 2000a). If such associations are found for all amino acids,

this could suggest that widespread adaptive codon reassignments never took place: although the genetic code exhibits adaptive error minimizing properties, they could be a byproduct of stereochemical interactions from which the code arose (though it would also remain plausible that such interactions merely acted together with selection to steer particular amino acids into the primordial code). At present, just a handful of amino acids and, while the general pattern is clear, not all show stereochemical affinity for their codons (Illangasekare *et al*., 2002): it remains parsimonious that any footprint of a stereochemical code was partially overwritten by adaptive reshuffling (or ambiguity reduction, or selection- mediated biosynthetic expansion) and that we are seeing the remnants left untouched by selective reassignment. But it is interesting to note that the original development of the Polar Requirement (Woese *et al*., 1966) was to test for a stereochemically determined code on the assumption that 2,6-dimethyl pyridine might mimic nucleotides.

To illustrate the sheer scope of possible interpretations for an error minimizing genetic code, one of us (Freeland, 2002) has recently drawn attention to Fisher's 'geometric theorem' (Fisher, 1930), a simple abstract model of evolution that predicts an inversely proportional relationship between the magnitude of effect of a random mutation and the probability that it will represent an adaptive improvement for a pleiotropic trait. Now, the pattern of amino acid assignments found within the standard genetic code is such that random nucleotide substitutions produce smaller differences in amino acid polarity than would be true for most other codes: we propose that this might maximize the probability of adaptive evolution wherever change is required (Freeland, 2002). This hypothesis invokes clade selection: organisms using such a code came to dominate primordial ecosystems because, over multiple generations, their offspring tended to win out in intraspecific struggles whenever environmental or biotic factors induced selective pressure for change. For now, the suggestion remains an intriguing possibility: evaluating its plausibility will require extensive simulations to test whether an error minimizing code can spontaneously emerge by out-competing alternatives over a broad range of conditions whenever genome change is required. However, it serves to illustrate the point that a code which maximizes the similarity of mutationally connected codons need not necessarily be a product of selection for minimizing the deleterious impact of errors.

6. The Extent of Adaptive Coding Properties

Finally, it is of course true that codon assignments represent just one aspect of how the standard code evolved: the wider context of how natural selection has influenced its form remains unclear. In part, this is because the 'genetic code' references a highly complex set of molecular machinery that creates a general interface between genetic information and all the protein products of an organism, and at a trivial level it is clear that the genetic code is adaptive in many senses. For

example, the aminoacyl tRNA synthetase enzymes, the tRNA's that they charge with amino acids and the ribosomes that coordinate translation cooperate to perform a highly sophisticated metabolic function: selection must have played a major role in coordinating this interaction. However, beyond such biochemical detail, several general properties may be abstracted by considering that every code is, in essence, a mapping function that connects a set of elements in one language to a set of elements in a second language. As such, three potential foci for selection emerge: the 'input' language (biologically, nucleic acid), the 'output' language (biologically, proteins) and the one discussed here, namely the set of rules by which one is mapped to the other (biologically, the set of 64 codon assignments). Thus, in addition to researching an adaptive pattern of codon assignments, it is legitimate to ask whether any properties of the nucleic acid and/or protein languages are themselves adaptations. Eschenmoser (1999), has made a plausible case that the fundamental chemical nature of nucleic acid (specifically the ribose backbone) represents an adaptive tradeoff in base-pairing strength, and Weber *et al*. (1981) have made an analogous claim for the amino acids (specifically that alpha amino acids were selected for their conformational rigidity). Further, the constituent elements of each alphabet might be adaptive choices: certainly we know that other base pairs can be incorporated into DNA and RNA (Piccirilli *et al*., 1990) and that other amino acids (Wong *et al*., 1979) were probably available to the primordial biosphere, though arguments for why specific chemical building blocks were accepted while others were rejected remain incomplete. Finally, the sizes of the respective chemical alphabets have been studied for adaptive properties: although numerous analyses have claimed that the original genetic code used only a subset of the 4 bases we see today, the claims are so varied (e.g. only A,U (Jimenez-Sanchez, 1995), no-A,U (Lehman *et al*., 1988; Hartman, 1995), all-purine (Hartman, 1995) or GCU-only (Trifonov *et al*., 1997)) as to prevent any clear message. Probably the most rigorously developed theoretical model asserts that a genetic alphabet of 4 bases represents a left-over optimum from an RNA world (Szathmáry, 1991b, 1992), where an increase in alphabet size would benefit catalytic potential but cost in terms of replication fidelity. Analogous models, asserting that 20 amino acids represent an adaptive trade-off between the increased catalytic potential afforded by expanding the encoded amino acid alphabet and the genome disruption caused by this change (Wong, 1976; Szathmary, 1991b), are suggestive but have been less thoroughly developed. Indeed the size and constituents of the biological alphabets have received far less attention than the topic of codon assignments and could add significantly to our understanding of the standard genetic code.

Even within the topic of the codon assignments of the standard code, one key aspect remains relatively ill explored. Recent observations from both biology (e.g. Maeshiro, *et al*.) and computer science (e.g. Kargupta, 2001) have started to re-examine the powerful influence of genetic code redundancy on the general dynamic of evolution. Though mainstream biology tends to interpret Crick's (1966) wobble hypothesis as a description of biochemical constraint, Ardell and Sella's (Ardell

et al., 2001; Sella *et al*., 2002) simulations find that patterns of codon degeneracy can spontaneously emerge from selection for an error minimizing code. This supports Szathmáry's (1991a) suggestion of code redundancy as an evolved state, though it stands contrary to the findings of a previous analysis that used a powerful computer search for 'ideal' codes, but once again ignored biases in nucleotide mutation/mistranslation (Judson *et al*., 1999). The unambiguous resolution of this issue would have significant impact on the sub-branch of codon assignment research that has included flexible coding within analyses that conclude an adaptive genetic code (e.g. Cullman and Labouygues, 1983, 1987; Figureau and Pouzet, 1984; Figureau, 1987, 1989; Luo, 1988, 1989; Luo *et al*., 2002)

Acknowledgements

We are grateful to the John Templeton Foundation for Research Grant 938-COS27 and to UMBC's DRIF funding for financial support of this research.

References

Aita, T., Urata, S. and Husimi, Y.: 2000, From Amino Acid Landscape to Protein Landscape: Analysis of Genetic Codes in Terms of Fitness Landscape, *J. Mol. Evol.* **50**, 313−323.

Alff-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.

Amirnovin, R. and Miller, S. L.: 1999, Response, *J. Mol. Evol.* **48**, 253−255.

Ardell, D. H.: 1998, On error Minimization in a Sequential Origin of the Standard Genetic Code, *J. Mol. Evol.* **47**, 1−13.

Ardell, D. and Sella, G.: 2001, On the Evolution of Redundancy in Genetic Codes, *J. Mol. Evol.* **53**, 269−281.

Barrell, B. G., Bankier, A. T. and Drouin, J.: 1979, A Different Genetic Code in Human Mitochondria, *Nature* **282**, 189−194.

Bashford, J. D., Tsohantjis, I. and Jarvis, P. D.: 1998, A Supersymmetric Model for the Evolution of the Genetic Code, *Proc. Natl. Acad. Sci. USA* **95**, 987−992.

Baumann, U. and Oro, J.: 1993, Three Stages in the Evolution of the Genetic Code, *Biosystems* **29**, 133−141.

Crick, F. H. C.: 1966, Codon-Anticodon Pairing: The Wobble Hypothesis, *J. Mol. Biol.* **19**, 548−555.

Crick, F. H. C.: 1968, The Origin of the Genetic Code, *J. Mol. Biol.* **38**, 367−379.

Crick, F. H. C., Griffith, J. S. and Orgel, L. E.: 1957, Codes Without Commas, *Proc. Natl. Acad. Sci. USA* **43**, 416−421.

Cullman, G. and Labouygues, J.: 1983, Noise Immunity of the Genetic Code, *Bio Systems* **16**, 9−29.

Cullman, G. and Labouygues, J.: 1987, The Logic of the Genetic Code, *Math. Model* **8**, 643−646.

Davies, J., Gilbert, W. and Gorini, L.: 1964, Streptomycin, Suppression and the Code, *Proc. Natl. Acad. Sci. USA* **51**, 883−890.

Davis, B. K.: 1999, Evolution of the Genetic Code, *Progr. Biophys. Molec. Biol.* **72**, 157−243.

Davydov, O.: 1996, Internal Logic of the Genetic Encoding: End-atom Rules of Doublet Composition, *ISSOL Newsletter* **23**, 12.

Davydov, O. V.: 1998, Amino Acid Contribution to the Genetic Code Structure: End-atom Chemical Rules of Doublet Composition, *J. Theor. Biol.* **193**, 679–690.

Di Giulio, M.: 1989, The Extension Reached by the Minimisation of Polarity Distances During the Evolution of the Genetic Code, *J. Mol. Evol.* **29**, 288–293.

Di Giulio, M.: 1991, On the Relationships Between the Genetic Code Co-evolution Hypothesis and the Physiochemical Hypothesis, *Z. Naturforsch* **46c**, 305–312.

Di Giulio, M.: 1994, The Phylogeny of tRNAs Seems to Confirm the Coevolution of the Origin of the Genetic Code, *Orig. Life Evol. Biosph.* **25**, 549–564.

Di Giulio, M.: 1997, On the Origin of the Genetic Code, *J. Theor. Biol.* **187**, 573–581.

Di Giulio, M.: 1998, The Historical Factor: The Biosynthetic Relationships Between Amino Acids and their Physiochemical Properties in the Origin of the Genetic Code, *J. Mol. Evol.* **46**, 615–621.

Di Giulio, M.: 1999a, The Coevolution Theory of the Origin of the Genetic Code, *J. Molec. Evol.* **48**, 253–254.

Di Giulio, M.: 1999b, The RNA World, the Genetic Code and the tRNA Molecule, *Trends Genet.* **15**, 223–229.

Di Giulio, M.: 2000, Genetic Code Origin and the Strength of Natural Selection, *J. Theor. Biol.* **205**, 659–661.

Di Giulio, M.: 2001a, A Blind Empiricism Against the Coevolution Theory of the Origin of the Genetic Code, *J. Mol. Evol.* **53**, 724–732.

Di Giulio, M.: 2001b, The Origin of the Genetic Code cannot be Studied using Measurements Based on the PAM Matrix Because this Matrix Reflects the Code itself, Making any such Analyses Tautologous, *J. Theor. Biol.* **208**, 141–144.

Di Giulio, M., Capobianco, M. R. and Medugno M.: 1994, On the Optimisation of the Physiochemcial Distances Between Amino Acids in the Evolution of the Genetic Code, *J. Theor. Biol.* **168**, 43–51.

Di Giulio, M. and Medugno, M.: 1999, Physicochemical Optimization in the Genetic Code Origin as the Number of Codified Amino Acids Increases, *J. Molec. Evol.* **49**, 1–10.

Di Giulio, M. and Medugno, M.: 2000, The Robust Statistical Bases of the Coevolution Theory of Genetic Code Origin, *J. Molec. Evol.* **50**, 258–263.

Di Giulio, M. and Medugno, M.: 2001, The Level and Landscape of Optimization in the Origin of the Genetic Code, *J. Molec. Evol.* **52**, 372–382.

Dillon, L. S.: 1973, The Origins of the Genetic Code, *The Botan. Rev.* **39**, 301–345.

Eigen, M.: 1971, Self-organization of Matter and the Evolution of Biological Macromolecules, *Naturwissenschaften* **58**, 465–522.

Eigen, M. and Schuster, P.: 1979, *The Hypercycle: A Principle of Natural Self-organisation*, Springer, New York.

Ellington, A. D., Khrapov, M. and Shaw, C. A: 2000, The Scene of a Frozen Accident, *RNA* **6**, 485–498.

Epstein, C. J.: 1966, Role of the Amino-acid 'Code' and of Selection for Conformation in the Evolution of Proteins, *Nature* **210**, 25–28.

Eschenmoser, A.: 1999, Chemical Etiology of Nucleic Acid Structure, *Science* **284**, 2118–2124.

Figureau, A.: 1987, Information Theory and the Genetic Code, *Orig Life* **17**, 439–449.

Figureau, A.: 1989, Optimization and the Genetic Code, *Orig. Life Evol. Biosph.* **19**, 57–67.

Figureau, A. and Pouzet, M.: 1984, Genetic Code and Optimal Resistance to the Effect of Mutations, *Orig. Life Evol. Biosph.* **14**, 579–588.

Fisher, R. A.: 1930, *The Genetical Theory of Natural Selection*, Oxford University Press, Oxford.

Fitch, W. M.: 1966a, An Improved Method for Testing for Evolutionary Homology, *J. Mol. Biol.* **16**, 9–16.

Fitch, W. M.: 1966b, The Relation Between Frequencies of Amino Acids and Ordered Trinucleotides, *J. Mol. Biol.* **16**, 1–8.

Fitch, W. M. and Upper, K.: 1987, The Phylogeny of tRNA Sequences Provides Evidence for Ambiguity Reduction in the Origin of the Genetic Code, *Cold Spring Harbour Symp. Quant. Biol.* **52**, 759–767.

Freeland, S. J.: 2002, The Darwinian Code: An Adaptation for Adapting, *J. Gen. Progr. Evolv. Machines* **3**, 113–127.

Freeland, S. J. and Hurst, L. D.: 1998a, The Genetic Code is One in a Million, *J. Mol. Evol.* **47**, 238–248.

Freeland, S. J. and Hurst, L. D.: 1998b, Load Minimisation of the Code: History does not Explain the Pattern, *Proc. Roy. Soc. Lond. B* **265**, 2111–2119.

Freeland, S. J., Knight, R. D. and Landweber, L. F.: 2000a, Measuring Adaptation within the Genetic Code, *Trends Biochem. Sci.* **25**, 44–45.

Freeland, S. J., Knight, R. D., Landweber L. F. and Hurst, L. D.: 2000b, Early Fixation of an Optimal Genetic Code, *Mol. Biol. Evol.* **17**, 511–518.

Friedman, S. M. and Weintstein, I. B.: 1964, Lack of Fidelity in the Translation of Ribopolynucleotides, *Proc. Natl. Acad. Sci. USA* **52**, 988–996.

Frisch, L. (ed.): 1966, 'The Genetic Code', *Cold Spring Harbor Symposia on Quantitative Biology*, pp. 1–747.

Gamow, G.: 1954, Possible Relation Between Deoxyribonucleic Acid and Protein Structures, *Nature* **173**, 318.

Gamow, G. and Ycas, M.: 1955, Statistical Correlation of Protein and Ribonucleic Acid Composition, *Proc. Natl. Acad. Sci. USA* **41**, 1011–1019.

Gesteland, R. F. and Atkins, J. F.: 1993, *The RNA World*, Cold Spring Harbour, Cold Spring Harbour Laboratory Press, New York.

Gesteland, R. F., Cech, T. and Atkins J. F. (eds): 1999, *The RNA World*, Cold Spring Harbor Monograph Series, Cold Spring Harbor Laboratory, New York.

Gilis, D., Massar, S. and Rooman M.: 2001, Optimality of the Genetic Code with Respect to Protein Stability and Amino-acid Frequencies, *Genome Biol.* **2**, RESEARCH0049.

Goldberg, A. L. and R. E. Wittes: 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.

Grivell, L. A.: 1986, Deciphering Divergent Codes, *Nature* **324**, 109–110.

Haig, D. and Hurst, L. D.: 1991, A Quantitative Measure of Error Minimisation within the Genetic Code, *J. Mol. Evol.* **33**, 412–417.

Haig, D. and Hurst, L. D.: 1999, A Quantitative Measure of Error Minimization in the Genetic Code, *J. Mol. Evol.* **49**, 708.

Hartman, H.: 1975, Speculations on the Evolution of the Genetic Code, *Orig. Life.* **6**(3), 423–427.

Hartman, H.: 1978, Speculations on the Evolution of the Genetic Code. II, *Orig. Life* **9**, 133–136.

Hartman, H.: 1984, Speculations on the Evolution of the Genetic Code III: The Evolution of t-RNA, *Orig. Life* **14**, 643–648.

Hartman, H.: 1995a, Speculations on the Evolution of the Genetic Code IV. The Evolution of the Aminoacyl-tRNA Synthetases, *Orig. Life Evol. Biosph.* **25**, 265–269.

Hartman, H.: 1995b, Speculations on the Origin of the Genetic Code, *J. Mol. Evol.* **40**, 541–544.

Hasegawa, M. and Miyata, T.: 1980, On the Asymmetry of the Amino Acid Code Table, *Orig. Life* **10**, 265–270.

Hayes, B.: 1998, The Invention of the Genetic Code, *Amer. Scientist* **86**, 8–14.

Illangasekare, M. and Yarus, M.: 2002, Phenylalanine-binding RNAs and Genetic Code Evolution, *J. Mol. Evol.* **54**, 298–311.

Jimenez-Sanchez: 1995, On the Origin and Evolution of the Genetic Code, *J. Mol. Evol.* **41**, 712–716.

Judson, O. and Haydon, D.: 1999, The Genetic Code: What is it Good for? An Analysis of the Effects of Selection Pressures on Genetic Codes, *J. Mol. Evol.* **49**, 539–550.

Jukes, T. H.: 1981, Amino Acids Codes in Mitochondria as Possible Clues to Primitive Codes, *J. Mol. Evol.* **18**, 15–17.

Kargupta, H.: 2001, A Striking Property of Genetic Code-like Transformations, *Compl. Syst.* **13**, 1–32.

Kauffman, S. A.: 1993, The Origins of Order: Self Organisation and Selection in Evolution, Oxford University Press, New York.

King, J. L. and Jukes, T. H.: 1969, Non-Darwinian Evolution, *Science* **164**, 788–798.

Knight, R. D., Freeland, S. J. and Landweber, L. F.: 1999, Selection, History and Chemistry: The Three Faces of the Genetic Code, *Trends Biochem. Sci.* **24**, 241–247.

Knight, R. D., Freeland, S. J. and Landweber L. F.: 2001a, Rewiring the Keyboard: Evolvability of the Genetic Code, *Nat. Rev. Genet.* **2**, 49–58.

Knight, R. D., Landweber, L. F. and Yarus, M.: 2001b, How Mitochondria Redefine the Code, *J. Mol. Evol.* **53**, 299–313.

Knight, R. D., Freeland, S. J. and Landweber L. F.: 2001c, A Simple Model Based on Mutation and Selection Explains Trends in Codon and Amino-acid Usage and GC Composition within and Across Genomes. *Genome Biol. 2001* **2**(4), RESEARCH0010.

Knight, R. D. and Landweber, L. F: 1998, Rhyme or Reason: RNA-arginine Interactions and the Genetic Code, *Chem. Biol.* **5**, R215–R220.

Knight, R. D. and Landweber, L. F: 2000a, The Early Evolution of the Genetic Code, *Cell* **101**, 569–572.

Knight, R. D. and Landweber, L. F: 2000b, Guilt by Association: The Arginine Case Revisited, *RNA* **6**, 499–510.

Lehman, N. and Jukes, T. H.: 1988, Genetic Code Development by Stop Codon Takeover, *J. Theor. Biol.* **135**, 203–214.

Luo, L. F.: 1988, The Degeneracy Rule of Genetic Code, *Orig. Life* **18**, 65–70.

Luo, L. F.: 1989, The Distribution of Amino Acids in Genetic Code, *Orig. Life* **19**, 621–631.

Luo, L. F. and Li, X.: 2002, Coding Rules for Amino Acids in the Genetic Code: The Genetic Code is a Minimal Code of Mutational Deterioration, *Orig. Life Evol. Biosph.* **32**, 621–631.

Maeshiro, T. and Kimura, M.: 1998, The Role of Robustness and Changeability on the Origin and Evolution of Genetic Codes, *Proc. Natl. Acad. Sci. USA* **95**, 5088–5093.

Osawa, S.: 1995, *Evolution of the Genetic Code*, Oxford University Press, Oxford.

Osawa, S. and Jukes, T. H.: 1989, Codon Reassignment (Codon Capture) in Evolution, *J. Mol. Evol.* **21**, 271–278.

Pace, C., Shirley, B., McNutt, M., and Gajiwala, K.: 1996, Forces Contributing to the Conformational Stability of Proteins, *FASEB J.* **10**, 75–83.

Parker, J.: 1989, Errors and Alternatives in Reading the Universal Genetic Code, *Microbiol. Rev.* **55**, 273–298.

Petrov, D. and Hartl, D.: 1999, Patterns of Substitution in Drosophila and Mammalian Genomes, *Proc. Natl. Acad. Sci. USA* **96**, 1475–1479.

Piccirilli, J. A., Krauch, T., Moroney S. E. and Benner S. A.: 1990, Enzymatic Incorporation of a New Base Pair into DNA and RNA Extends the Genetic Alphabet, *Nature* **343**, 33–37.

Ronneberg, T. A., Landweber, L. F. and Freeland, S. J.: 2000, Testing a Biosynthetic Theory of the Genetic Code: Fact or Artifact? *Proc. Natl. Acad. Sci. USA* **97**, 13690–13695.

Sella, G. and Ardell, D.: 2002, The Impact of Message Mutation on the Fitness of a Genetic Code, *J. Mol. Evol.* **54**, 638–651.

Shepherd, J. C.: 1981, Periodic Correlations in DNA Sequences and Evidence Suggesting their Evolutionary Origin in a Comma-less Genetic Code, *J. Mol. Evol.* **17**, 94–102.

Sonneborn, T. M.: 1965, *Degeneracy in the Genetic Code: Extent, Nature and Genetic Implications, Evolving Genes and Proteins*, V. Bryson and H. J. Vogel (eds), Academic Press, New York and London.

Sowerby, S. J. and Heckl, W. M.: 1998, The Role of Self-assembled Monolayers of the Purine and Pyrimidine Bases in the Emergence of Life, *Orig. Life Evol. Biosph.* **28**, 283–310.

Stahl, G., McCarty, G. and Farabaugh P. J.: 2002, Ribosome Structure: Revisiting the Connection Between Translational Accuracy and Unconventional Decoding, *Trends Biochem. Sci.* **27**, 178–183.

Swanson, R.: 1984, A Unifying Concept for the Amino Acid Code, *Bull. Math. Biol.* **46**, 187–203.

Szathmáry, E.: 1991a, Codon Swapping as a Possible Evolutionary Mechanism, *J. Mol. Evol.* **32**, 178–182.

Szathmáry, E.: 1991b, Four Letters in the Genetic Alphabet: A Frozen Evolutionary Optimum? *Proc. Roy. Soc. Lond. B* **245**, 91–99.

Szathmáry, E.: 1992, What is the Optimum Size for the Evolutionary Alphabet? *Proc. Natl. Acad. Sci. USA* **89**, 2614–2618.

Szathmáry, E.: 1999, The Origin of the Genetic Code, *Trends Genetics* **15**, 223–229.

Szathmáry, E. and Maynard Smith, J.: 1995, *The Major Transitions in Evolution*, W. H. Freeman, Oxford and New York.

Szathmáry, E. and Zintzaras, E.: 1992, A Statistical Test of Hypotheses on the Organization and Origin of the Genetic Code, *J. Mol. Evol.* **35**, 185–189.

Tomii, K. and Kanehisa, M.: 1996, Analysis of Amino Acid Indices and Mutation Matrices for Sequence Comparison and Structure Prediction of Proteins, *Protein Eng.* **9**, 27–36.

Topal, M. D. and Fresco, J. R.: 1976, Complementary Base Pairing and the Origin of Substitution Mutations, *Nature* **263**, 285–289.

Trifonov, E. and Bettecken, T.: 1997, Sequence Fossils, Triplet Expansion, and Reconstruction of Earliest Codons, *Gene* **205**, 1–6.

Trifonov, E. N.: 2000, Consensus Temporal Order of Amino Acids and Evolution of the Triplet Code, *Gene* **261**, 139–151.

Volkenstein, M. V.: 1965, Coding of Polar and Non-polar Amino Acids, *Nature* **207**, 294–295.

Wakeley, J.: 1994, Substitution-rate Variation Among Sites and the Estimation of Transition Bias, *Mol. Biol. Evol.* **11**, 436–442.

Weber, A. L. and Miller, S. L.: 1981, Reasons for the Occurrence of the Twenty Coded Protein Amino Acids, *J. Mol. Evol.* **17**, 273–284.

Woese, C. R.: 1965, On the Evolution of the Genetic Code, *Proc. Natl. Acad. Sci. USA* **54**, 1546–1552.

Woese, C. R.: 1973, Evolution of the Genetic Code, *Naturwissenschaften* **60**, 447–459.

Woese, C. R., Dugre, D. H., Saxinger W. C. and Dugre S. A.: 1966, On the Fundamental Nature and Evolution of the Genetic Code, *Cold Spring Harbour 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.

Wong, J. T.-F.: 1976, The Evolution of a Universal Genetic Code, *Proc. Natl. Acad. Sci. USA* **73**, 2336–2340.

Wong, J. T.-F.: 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.: 1981, Co-evolution of the Genetic Code and Amino Acid Biosynthesis, *Trends Biochem. Sci.* **6**, 33–35.

Wong, J. T.-F.: 1988, Evolution of the Genetic Code, *Microbiol. Sci.* **5**, 174–181.

Wong, J. T.-F. and P. M. Bronskill: 1979, Inadequacy of Pre-biotic Synthesis as the Origin of Proteinaceous Amino Acids, *J. Mol. Evol.* **13**, 115–125.

Yarus, M.: 2000, RNA-ligand Chemistry: A Testable Source for the Genetic Code, *RNA* **6**, 475–484.

Zuckerkandl, E. and Pauling, L.: 1965, *Evolutionary Divergence and Convergence in Proteins. Evolving Genes and Proteins*, V. Bryson and H. J. Vogel (eds), Academic Press, New York and London.