THE ROYAL SOCIETY

*Published online* 31 October 2002

# No accident: genetic codes freeze in error-correcting patterns of the standard genetic code

**David H. Ardell*** **and Guy Sella†**

*Department of Biological Sciences, Stanford University, Stanford, CA 94305-5020, USA*

The standard genetic code poses a challenge in understanding the evolution of information processing at a fundamental level of biological organization. Genetic codes are generally coadapted with, or 'frozen' by, the protein-coding genes that they translate, and so cannot easily change by natural selection. Yet the standard code has a significantly non-random pattern that corrects common errors in the transmission of information in protein-coding genes. Because of the freezing effect and for other reasons, this pattern has been proposed not to be due to selection but rather to be incidental to other evolutionary forces or even entirely accidental.

We present results from a deterministic population genetic model of code-message coevolution. We explicitly represent the freezing effect of genes on genetic codes and the perturbative effect of changes in genetic codes on genes. We incorporate characteristic patterns of mutation and translational error, namely, transition bias and positional asymmetry, respectively. Repeated selection over small successive changes produces genetic codes that are substantially, but not optimally, error correcting. In particular, our model reproduces the error-correcting patterns of the standard genetic code. Aspects of our model and results may be applicable to the general problem of adaptation to error in other natural information-processing systems.

**Keywords:** frozen accident; load minimization; error minimization; quasistatic approximation; codon reassignment; quasispecies

## 1. INTRODUCTION

A communicative act requires a *signal*, the transmission of information in some medium through time and space. The transmitted information may concern the state of the source of the signal. Usually the signal *represents* the state of the source through an abstract association of signals and states that must be known at both the source and receiver of the signal to operate effectively. We call this conventional association of states and signals a *code*; a code defines which signals are associated with which states. The states may loosely be called the 'meanings' of the signals as defined by the code. Signals are generally transmitted through media that have intrinsic rates and patterns of error. A source of mismatch between signal and state (as defined by the code) may come from loss of information or error in the transmission of signals. Errors may also occur in the *encoding* of states and the *decoding* of signals once they have been received.

The code that ascribes meaning to signals plays a vital role in communication. Errors in signal transmission or decoding may cause unintended semantic interactions, when a signal meaning one thing is 'crossed with' or received as another valid signal that means something entirely different. In this case, a code that reduces the semantic differences frequently crossed with one another might be advantageous. In the extreme, a certain code may assign the exact same meaning to different signals. We say that these signals are *synonymous* and the code that contains them is partially or totally *redundant*. Signals that are not redundant may still encode very similar meanings. Short of giving a formal definition, we can say that if, in a given code, the geometry of similarity and difference over the set of meanings is preserved in the geometry of error rates among signals, the code is *structure preserving* (Sella & Ardell 2002). A structure-preserving code assigns similar meanings to frequently crossed signals.

Two questions arise: how does one design a structure-preserving code, and how would such a code originate naturally? One answer to the first question is given as an example in § 1a. In this paper, we address the second question of how structure-preserving codes originate. We focus our attention on a system at the subcellular level whose evolution has been the subject of great controversy: namely, the *genetic codes* that operate upon and express the information in *protein-coding genes* through the process of *biological translation*. In this system, the signals are units of translation in protein-coding genes called *codons*. Translation maps a sequence of codons into a sequence of amino acids in proteins. Variability in genetic codes occurs in both organellar and nuclear genomes (Osawa *et al.* 1992). However, a *standard genetic code* (or just 'standard code') is common to all major domains of life on Earth, which is perhaps among the best evidence for their

*Author and address for correspondence: Department of Molecular Evolution, Uppsala University, Norbyvägen 18C, SE-752 36 Uppsala, Sweden (dave.ardell@ebc.uu.se)

† Present address: Department of Applied Mathematics and Computer Science, The Weizmann Institute, Rehovot 76100, Israel.

One contribution of 12 to a Theme Issue 'Information and adaptive behaviour'.


DOI 10.1098/rstb.2002.1071

common ancestry. Alternative genetic codes are never different from the standard code in more than a few codons, and they are most parsimoniously explained as derived from the standard genetic code (Knight *et al.* 2001). According to one estimate the standard genetic code is almost as old as life on Earth itself (Eigen *et al.* 1989).

Many researchers have argued that the standard genetic code is structure preserving (evidence reviewed in § 1a) with respect to translational errors or mutation. They proposed that the standard code was selected to reduce the deleterious consequences of errors in the translation (e.g. Woese 1965*a*; Alff-Steinberger 1969; Swanson 1984; Haig & Hurst 1991; Szathmary & Zintzaras 1992; Goldman 1993; Di Giulio *et al*. 1994; Ardell 1998; Freeland & Hurst 1998; Freeland *et al*. 2000, and others) and hereditary transmission (e.g. Sonneborn 1965; Zuckerkandl & Pauling 1965; Epstein 1966; Goldberg & Wittes 1966; Sitaramam 1989; Joshi *et al*. 1993; Ardell 1998; Sella & Ardell 2002; Ardell & Sella 2001, and others) of protein-coding genes. There are subtle differences in detail among the various hypotheses. We gloss over this detail here and refer to what they have in common, namely that the pattern of assignments in the genetic code predominantly originated by natural selection, as the *selection hypothesis*.

The selection hypothesis is important for considerations of adaptation in information-processing systems. If it is true, we can say that there is evidence for adaptation to error in information processing at the most fundamental level of life. But there are objections to the hypothesis that stem in part from the absence of a plausible mechanism of how such selection can occur. We believe that we have found such a plausible mechanism. To describe it we must first introduce more vocabulary and concepts that are specific to our model and the system we have studied.

As genes are transmitted from generation to generation, and because the same genetic source is always indirectly reused for translation, the signals that transmit protein-coding information are persistent. We say here that a persistent encoded signal, the correct decoding of which a biological entity depends upon for fitness, is a *message*. Thus, we call the totality of protein-coding genes in a genome a 'message', not to be confused with messenger RNA. With this vocabulary in hand, we can talk naturally and more generally of 'codes' and 'messages' that are interdependent for the transmission of meaning.

In earlier work we showed that genetic codes that are structure preserving are also *error correcting* through studying their net effect on individual fitness (Sella & Ardell 2002). The distinction is necessary because messages can adapt to a code so as to avoid using unreliable signals. We found this to be the case: fewer mutations persist in messages at mutation–selection equilibrium with a less structure-preserving genetic code. Despite this, the less structure-preserving genetic code has a fitness disadvantage. We use the term 'error correction' somewhat loosely here, as it usually refers to the reduction of source uncertainty (Ash 1965) rather than to the amelioration of the consequences of uncorrected errors. Thus, we refer specifically to effects of the organizational *pattern* of coding assignments rather than to error-correcting (or detecting) *processes* such as proof reading or repair. In this respect, a structure-preserving code generalizes the concept of a 'sphere-packing' or Hamming code (Ash 1965).

The coding assignments in the standard genetic code were probably not predetermined. The initial evidence for this comes from the physical basis of translation. Physically, biological translation is mediated by a diverse assembly of macromolecular components called the *translational apparatus*. The translational apparatus may, to a first approximation, be decomposed into mutually independent *reading* and *decoding* components, with 'adapter' molecules connecting them (Crick 1966). The reading component (ribosomal initiation and elongation steps on mRNA, aminoacyl-tRNA binding to EF-Tu, etc.) enables genetic information to be transmitted in units of codons all through a common medium and format. The decoding component (aminoacylation of tRNA and codon–anticodon pairing at the ribosome) translates or decodes that information from codons to amino acids. The decoding and reading components of translation are biochemically distinct and independent; for instance, reading and polymerization steps of protein synthesis are not thought to correct or prevent errors in the decoding step (Ibba & Söll (1999), but see LaRiviere *et al.* (2001)). Consequently, genetic codes may be experimentally modified, also for the incorporation of nonstandard amino acids (e.g. Wong 1983; Bain *et al.* 1992; Döring & Marliere 1998; Wang *et al.* 2001). This modularity of the decoding component of translation implies a relative lack of constraint on genetic-code evolution by the translational apparatus.

Given the great age of the translational apparatus and the modularity of its decoding component, why do we not see greater variety in genetic codes today? The most widely accepted answer was perhaps most famously stated by Crick (1968), who called it 'freezing': an alteration in the genetic code is deleterious because it disrupts the protein 'sense' or 'meaning' of all protein-coding genes. Just as amino acid substitutions in proteins are rarely advantageous, the combined effect of many such substitutions is also rarely advantageous. Thus, all protein-coding genes act in concert to constrain change in genetic codes because their protein sense is expressed in that genetic code. One may say that genes are coadapted to the code with which they are co-inherited.

This freezing constraint of protein-coding genes on the genetic code is difficult to reconcile with evidence (reviewed in § 1a) that the genetic code was selected upon to correct errors in the transmission of protein-coding information. As a result, many alternative hypotheses have been advanced to explain both the patterns and the ubiquity of the standard genetic code. The two main alternatives are *stereochemical predetermination* of the genetic code through intrinsic affinities of the various amino acids with distinct components of the translation apparatus (e.g. Woese 1965*b*, 1967, 1973; Pelc & Welton 1966; Woese *et al*. 1966; Juncgk 1978; Shimizu 1982; Lacey & Mullins 1983; Knight & Landweber 2000; Yarus 2000, and references therein) and *coevolution with amino acid biosynthesis* wherein the code vocabulary of amino acids is thought to have expanded through the reassignment of codons from amino acids to their biosynthetic products (e.g. Wong 1975, 1980; Taylor & Coates 1989; Di Giulio 1995; Di Giulio & Medugno 1999, and references therein). A subtle relative to the metabolic coevolution hypothesis is the hypothesis that mutationally similar codons (codons that

share bases in common), and adapters between codons and amino acids (such as tRNAs and aminoacyl tRNA synthetases) are gradually donated to physicochemically similar amino acids through successive duplication and divergence of components of the translational apparatus (Cavalcanti *et al.* 2000, for example).

These hypotheses are not mutually exclusive and many have advocated that the code was shaped by a combination of these and possibly other factors (e.g. Woese 1967; Crick 1968; Szathmary & Zintzaras 1992; Di Giulio 1997; Knight *et al.* 1999). The different hypotheses make similar or consistent predictions about either codon–amino-acid orders or inter-codon orders in the genetic code. It should be noted that all mechanisms enabling and shaping genetic-code change must answer to the freezing constraint of messages.

Many arguments have been made against the selection hypothesis. Woese (1967) and Juncgk (1978) suggested that the number of different possible genetic codes is too great to have been acted upon by natural selection. This objection may be answered by supposing that selection on codes accumulated through intermediate stages (Sonneborn 1965; Fitch 1966; Crick 1968; Ardell 1998). Although gradual code change is subject to the message-constraint problem, one may suppose that the code was initially ambiguous and successively refined in parallel with the lengthening and differentiation of messages (Woese 1965*a*; Fitch 1966; Fitch & Upper 1987). Further, Woese (1967), Crick (1968) and others have argued that while translational error may have selected on the genetic code, mutation could not have. Mutations were supposed to be rare, so that ameliorating them through the amino acid configuration of the code was thought to be of relatively little benefit when compared with other forces acting on the code. Sella & Ardell (2002) studied this problem explicitly and showed that genetic codes that ameliorate mutations can have relatively large benefits over those that do not, because non-synonymous mutations accumulate in mutation–selection equilibrium.

Currently, there is a debate in the literature about how well the genetic code is optimized to ameliorate errors compared with random codes, and how this optimization is best measured (e.g. Di Giulio 2000, and references therein). Of course, freezing and other constraints will limit how optimal the code can evolve to become by natural selection (Crick 1968). It is clear that the genetic code is not optimal (e.g. Goldman 1993; Judson & Haydon 1999). It therefore seems that the argument of whether or not natural selection was a major driving force in code evolution should not rest entirely on the degree to which the genetic code is or is not optimal. Instead, we regard it as inevitable that the constitution of genes influenced code evolution, and that code changes—however they may have occurred—certainly influenced the pattern of codons in genes. We wish to explore how this dynamic may have shaped code evolution.

In an earlier work (Ardell & Sella 2001), we presented a dynamic model of *code-message coevolution*, in which small, sequential modifications of established genetic codes may only occur when they increase the fitness of existing messages. Messages then adapt to these code changes, thereby influencing the fitness of subsequent code changes. Code-message coevolution proceeds until no mutant code has better fitness with the messages of the old code than the old code has, at which point we say that the old code has *frozen*. Code-message coevolution should be distinguished both from the previously introduced metabolic coevolution hypothesis, and also from the work of Bedian (1982), who treated coevolution of the genetic code with the genetically encoded protein components of the translational apparatus. We found by simulation that codes that coevolve with messages that mutate according to a simple mutation model always freeze with some redundancy, that the diversity in encoded amino acids is always less than its potential, and that these results vary quantitatively in the mutation rate and strength of selection on proteins. Surprisingly, many codon reassignments occur *en route* to freezing, also in a parameter-dependent way. Both assignments and reassignments are selected to correct mutations in messages, producing structure-preserving genetic codes (Ardell & Sella 2001).

The current paper extends this model to show that by adding simple biologically motivated assumptions about error in the hereditary transmission and translation of protein-coding genes, we can reproduce two well known structure-preserving patterns found in the standard genetic code. Whereas previous studies have established the statistical reality of these patterns, this paper proposes a plausible mechanism of how they could have evolved through the sole agency of natural selection. In this way we demonstrate the feasibility of adaptation to errors in information processing at one of the most fundamental levels of biological organization. Our results indicate the general evolutionary principle that constraint by messages on code change may not prevent but actually cause the evolution of error-correcting codes.

### (a) *The regularities in the standard genetic code*

We briefly review the evidence of structure-preserving patterns in the standard genetic code.

Figure 1 applies a scale of physicochemical properties of amino acids called the polar requirement of Woese (1973) to assign colours to codons in the standard genetic code, with darker colours corresponding to smaller values on the scale.

We note the following patterns in the standard code and their corresponding systematic errors.

#### (i) *Pattern I*

Codons that differ in the first position are assigned to much more physicochemically similar amino acids than codons that differ in the second codon position (Woese 1965*a*; Alff-Steinberger 1969; Swanson 1984; Haig & Hurst 1991). This is apparent in the column-like pattern of the code in figure 1, and corresponds to observations that translational error is more frequent in the first codon position than in the second position (Davies *et al.* 1964, 1966; Parker 1989).

#### (ii) *Pattern II*

In all three codon positions, the amino acids associated with codons that differ by a single pyrimidine base $Y = \{U, C\}$ or purine base $R = \{A, G\}$ are more similar than other pairs of codons (Ardell 1998; Freeland & Hurst 1998). Figure 1 shows that amino acids are more similar within blocks of adjacent columns, corresponding to transitions

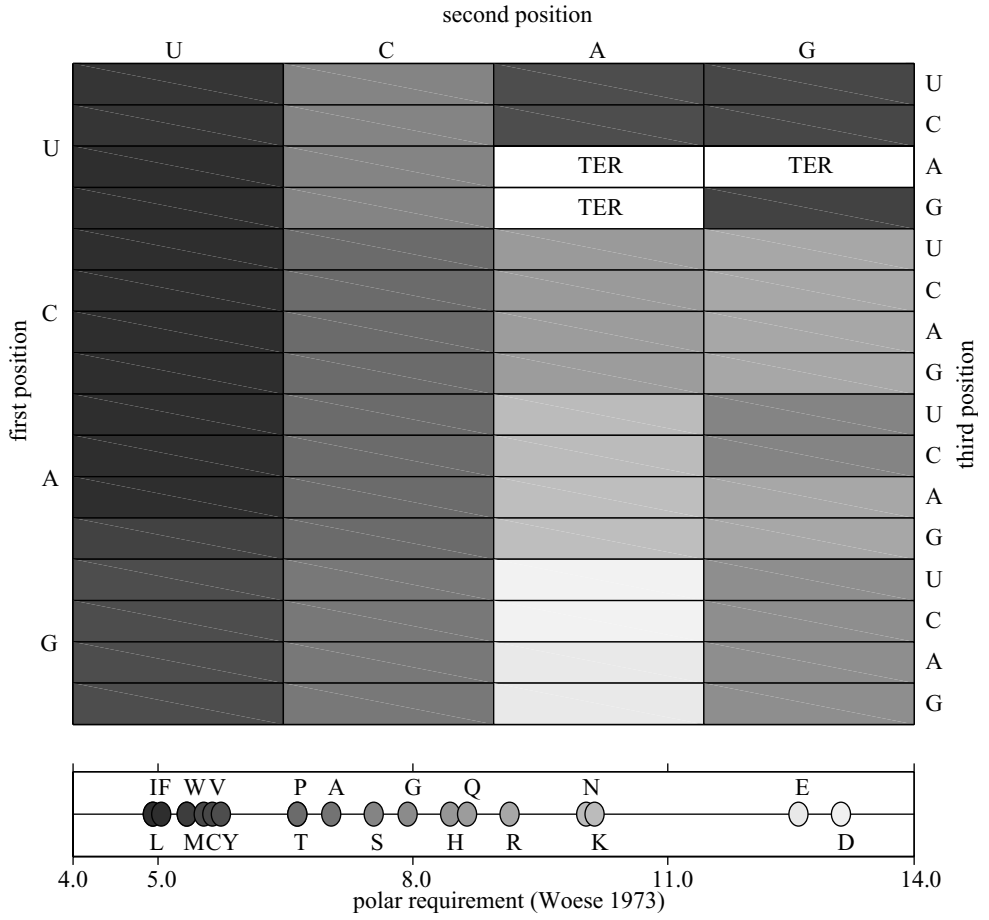


Figure 1. The standard genetic code with codons shaded according to the polar requirement (PR73) of encoded amino acids of Woese (1973). PR73 is shown in the scale below the code, labelled with one-letter amino acid codes. Darker shades correspond to smaller values of PR73. Columns of four-codon blocks vary in the first-position base, rows in the second-position base, and columns within a four-codon block in the third-position base. 'TER' denotes stop codons.

in the second codon position; the trend is more noticeable in the A, C and U first-position contexts (rows) than it is in the position I G context (Ardell 1998). The block-like pattern also occurs in the other codon positions. The pattern is associated with mutational biases in replication, damage and repair, in which transitions (mutations within these base sets) occur more frequently than transversions (mutations of a base within one set to a base in the other set) (Freese 1959; Topal & Fresco 1976; Echols & Goodman 1991). There is also some evidence for transition bias in translational error (Friedman & Weinstein (1964), but see p. 642 in Sella & Ardell (2002) for discussion).

(iii) *Pattern III*

Within the first codon position, there is greater similarity among amino acids in the context of second-position pyrimidine bases than in the context of second-position purine bases (Ardell 1998). Figure 1 shows that pattern I is more pronounced on the left-hand side of the figure corresponding to U and C in the second codon position. Pattern II, the transition-biased pattern, is more pronounced on the right-hand side of figure 1 corresponding to the second-position A and G contexts. Translational error at the first position is higher in a second-position pyrimidine (and especially U) context than it is in a second-position purine context (Davies *et al.* 1966; Negre *et al.* 1988).

(iv) *Pattern IV*

The amino acids along the third codon position are very similar and often redundant. This is associated with both wobble pairing and error rates that are among the highest in the third codon position (e.g. El'skaya & Soldatkin (1985)).

The patterns described in § 1a(i)–(iv) are robust to different methods of defining amino acid distance. Even when amino acid substitution rates are used to define amino acid distances, the same patterns are found (note that substitution matrices may be chosen to reduce the influence of the genetic code itself (Ardell 1998)). A clarifying comment about pattern II is in order. First-position transitions and first-position transversions are found to have approximately equal *P*-values, where a *P*-value is the number of permutant codes with an average physicochemical distance between amino acids smaller than that of the standard code (Ardell 1998). As there are twice as many transversions as transitions, and therefore twice as many ways to increase the average physicochemical distance of encoded amino acids in transversions, transversions should have half the *P*-value of transitions by the permutation method if transitions and transversions are equally physicochemically conservative on average. Therefore, the result implies a tendency for transitions to be more conservative than transversions in the first codon position.

In this paper, we implement three models to demonstrate that code-message coevolution can reproduce patterns I and II of the standard genetic code. With one model we study the effect of transition bias on code-message coevolution without translational misreading. With the second we examine the effect of positionally asymmetric translational misreading on code-message coevolution without transition bias. The third model brings the first two models together to show how code-message coevolution with transition bias and positionally asymmetric translational misreading reproduces patterns I and II of the standard genetic code.

## 2. MODELS AND METHODS

Our models build on the framework, assumptions and results in Sella & Ardell (2002) and Ardell & Sella (2001) to which the reader is referred for more detail. To specify our model we must define the set of codons and how they mutate; the form and characteristics of translation and translational error; the sets of amino acids and types of site in the proteins, and the elementary fitness matrix that defines the fitness of amino acids in any type of site; the target, a vector that associates site-types to a vector of codons (called the message); the initial genetic code in the population; the scheme by which genetic codes change; and the coevolutionary dynamic of codes and messages.

### (a) *The codon set and the codon-mutation scheme*

We model a codon set with two positions over the set of four bases $B = \{\mathrm{U,C,A,G}\}$. The codon set $C_{\mathrm{II}}^{B}$ then consists of 16 codons

$$C_{\mathrm{II}}^{B} = B \times B = \{\mathrm{UU,UC,UA,\ldots,GG}\}. \tag{2.1}$$

The codon mutation matrix $\mu_C$ is defined in terms of the base mutation matrix $\mu_B = \{\mu_B(y\,|\,x)\}_{x,y\,\in\,B}$, where $\mu_B(y\,|\,x)$ is the probability of base $x$ mutating into base $y$ in a single generation. Assuming sites mutate independently, the corresponding codon mutation is

$$\mu_C(zw\,|\,x\,y) = \mu_B(z\,|\,x)\mu_B(w\,|\,y) \qquad x,y,z,w \in B, \quad xy,zw \in C_{\mathrm{II}}^{B}. \tag{2.2}$$

We say that the codon set $C_{\mathrm{II}}^{B}$ together with the codon mutation matrix is a *codon space* where the geometry on codons is defined by the inverse of the mutation rate between them $d_C(i,j) \equiv \mu_C(i|j)^{-1}$ so that a codon that is a frequent (rare) mutant of another codon is 'close to' ('far from') that codon.

The base mutation matrix is structured to reflect our knowledge about the systematic errors that probably occurred in replication during the evolution of the code. We assume that mutation was always inherently transition biased, even if the code evolved in an RNA world. Reports of transition bias in natural RNA-dependent RNA replication in polio virus (Kuge *et al.* 1989), in *in vitro* evolved RNA-dependent RNA replicases (Ekland & Bartel 1996), and recent evidence that transition bias in substitution rates in mitochondria is due to mutation and not selection given the code (Denver *et al.* 2000) substantiate this assumption.

A base mutation matrix that incorporates transition bias in mutation and uniform fidelity of replication is:

$$\mu_B = \begin{array}{c} \\ \mathrm{U} \\ \mathrm{C} \\ \mathrm{A} \\ \mathrm{G} \end{array} \begin{array}{c} \begin{array}{cccc} \mathrm{U} & \mathrm{C} & \mathrm{A} & \mathrm{G} \end{array} \\ \begin{pmatrix} 1-\mu & \frac{\kappa\mu}{\kappa+2} & \frac{\mu}{\kappa+2} & \frac{\mu}{\kappa+2} \\ \frac{\kappa\mu}{\kappa+2} & 1-\mu & \frac{\mu}{\kappa+2} & \frac{\mu}{\kappa+2} \\ \frac{\mu}{\kappa+2} & \frac{\mu}{\kappa+2} & 1-\mu & \frac{\kappa\mu}{\kappa+2} \\ \frac{\mu}{\kappa+2} & \frac{\mu}{\kappa+2} & \frac{\kappa\mu}{\kappa+2} & 1-\mu \end{pmatrix} \end{array}. \tag{2.3}$$

The parameter $\mu$ is the *message mutation rate parameter*, or the probability of a base being replicated correctly after one generation. The *message* of an individual is a vector of $L$ codons $\boldsymbol{m} = \langle m_1,\ldots m_L\rangle$, $m_i \in C_{\mathrm{II}}^{B}$, representing the concatenation of all protein-coding genes. The parameter $\kappa$ is the *transition-bias parameter*, or the ratio of the probabilities of mutation between transitions and transversions

$$\frac{\mu_B(\mathrm{ts}(x)\,|\,x)}{\mu_B(\mathrm{tv}(x)\,|\,x)} = \frac{\kappa\mu/(\kappa+2)}{\mu/(\kappa+2)} = \kappa \quad \text{for any } x \in B, \tag{2.4}$$

where $\mathrm{ts}(x)$ is the base that is a transition of $x$, and $\mathrm{tv}(x)$ is one of the two bases that are transversions of $x$. This parameter corresponds to $TS:TV$ in Wakeley (1996). Note that changing $\kappa$ does not alter the rate of mutation, and that the $\kappa = 1$ case corresponds to the one-parameter mutation model studied in Ardell & Sella (2001). We study a range of transition bias ($1 \leqslant \kappa \leqslant 7$).

### (b) *Translational error*

For the models we present here, we hold translational misreading and mischarging to be evolutionarily invariant. We only consider decoding error, which we separate into *misreading* (codon–anticodon mispairing) and *mischarging* (tRNA misaminoacylation) components. We assume that misreading may occur at different rates in the different codon positions, is context independent and unbiased across bases.

The misreading filter matrix $F_r$ defines the probability of a codon from the codon set $C_{\mathrm{II}}^{B}$ being translationally misread, so that the $i,j$th element of $F_r$ describes the probability $F_r(j\,|\,i)$ of misreading codon $i$ as codon $j$ within one round of translation.

Denoting by $e_i$ the uniform misreading on the $i$th codon position, the misreading filter $F_r$ is given by

$$F_r(zw\,|\,xy) = \left[(1-e_1)I(x=z) + \frac{e_1}{3}I(x\neq z)\right] \left[(1-e_2)I(y=w) + \frac{e_2}{3}I(y\neq w)\right] \qquad x,y,z,w \in B, \quad xy,zw \in C_{\mathrm{II}}^{B}, \tag{2.5}$$

where $I(\text{condition})$ is an indicator function equal to 1 when the condition is met, and equal to 0 otherwise.

For the first of the three models we present, we assume no misreading. For the remainder, we analyse conditions where misreading in the first position ranges between $1.2 \times 10^{-5} \leqslant e_1 \leqslant 0.06$. We neglect misreading in the second position by assumption ($e_2 = 0$). We do not treat mischarging in the current study.

### (c) *The amino acid and site-type sets, genetic code, elementary fitness matrix and target*

We assume an immutable set $A$ of $M = 20$ amino acids that are always metabolically available to be translated into proteins by all individuals throughout the course of code evolution. All amino acids are assigned a physicochemical coordinate between 0 and 1 from a uniform distribution $U(0, 1)$. The coordinate of an amino acid corresponds to a normalized physicochemical index associated with functionality in proteins, such as polar requirement. Woese and his co-workers developed the measure to explore the stereochemical, rather than the selection hypothesis, but results with the measure are consistent with the latter (Woese *et al.* 1966; Woese 1973; Haig & Hurst 1991). An amino acid $\alpha \in A$ may be denoted directly by its coordinate, so that $0 < \alpha < 1$. Given two amino acids $\alpha$, $\beta \in A$, their coordinates allow us to define a physicochemical distance $d_A(\beta\,|\,\alpha) \equiv |\beta - \alpha|$ between them. We say that the amino acid set $A$ together with the distance $d_A$ is an *amino acid space* which is a metric space on amino acids, so that physicochemically similar amino acids are 'close' and so on.

A *genetic code* $c(\alpha\,|\,i)$ determines the probability of producing amino acid $\alpha \in A$ from codon $i \in C_{\Pi}^{B}$ in a message through translation, after it has passed through the non-evolvable filter $F_r$ of translational misreading. The combined translation probability $cF_r(\alpha\,|\,i)$ of translating codon $i$ as amino acid $\alpha$ may then be written

$$cF_r(\alpha\,|\,i) = \sum_{j \in C_{\Pi}^{B}} c(\alpha\,|\,j)\, F_r(j\,|\,i). \tag{2.6}$$

Translation is assumed independent by codon, so that the probability $P(\boldsymbol{p}\,|\,\boldsymbol{m})$ of producing any *protein*, which is a vector of $L$ amino acids $\boldsymbol{p} = \langle p_1,\ldots,p_L\rangle$, $p_l \in A$, from a given message $\boldsymbol{m}$ through translation using code $c$ is

$$P(\boldsymbol{p}\,|\,\boldsymbol{m}) = \prod_{l=1}^{L} cF_r(p_l\,|\,m_l). \tag{2.7}$$

A *site* refers both to the specific locus of a codon in a message and its corresponding residue location in a protein. We assume that the contribution to fitness of an amino acid in any site in a protein is completely determined by the *type* of that site, its *site-type*. We choose the set $S$ of $T = 20$ site-types to be in one-to-one correspondence to the set $A$ of amino acids, such that each site-type is associated with a distinct *target amino acid*, which is the unique amino acid conferring maximal fitness in sites of that type. The one-to-one correspondence of site-types with amino acids allows us to define a *site-type space* from the distance $d_A$ on the amino acids. Site-types are 'close' that have similar physicochemical requirements, and thus correspond to amino acids that are 'close' in the previously defined distance $d_A$.

Note that because we assume 20 amino acids (and hence 20 site-types each with a different target amino acid), no individual can attain the theoretical maximum fitness, as each individual may only encode a maximum of 16 different amino acids to satisfy the requirements of 20 different site-types.

The *target* is the vector of $L$ site-types, $\boldsymbol{s} = \langle s_1,\ldots s_L\rangle$, $s_l \in S$, which determines the type of the $l$th site of any message $\boldsymbol{m}$ in the population or any protein $\boldsymbol{p}$ in any individual, and thereby the fitness of all proteins. The frequencies $l_s$, $l_t > 0$ of site-types $s$,$t \in S$ in the target, $\Sigma_{s \in S} l_s = L$ are assumed to be all equal ($l_s = l_t$ for all $s$, $t \in S$). The target is fixed throughout evolutionary time.

The choice of the elementary fitness matrix $\omega(\cdot\,|\,\cdot)$ completely defines the fitness of any amino acid in a site of any type. We call our choices of $A$, $S$ and $\omega$ here the *physicochemical accuracy scheme*. Denote by $s_\alpha \in S$ the unique and distinct site-type associated with the target amino acid $\alpha$. The physicochemical accuracy scheme makes the fitness of an amino acid $\beta$ in a site of a given type $s_\alpha$ reflect the accuracy with which its physicochemical properties matches that of the target amino acid $\alpha$. We thus choose the elementary fitness matrix to reflect the physicochemical distances between amino acids $\beta$ and $\alpha$. In mathematical terms, we require that

$$\omega(\beta\,|\,s_\alpha) = f(d(\beta\,|\,\alpha)), \tag{2.8}$$

where the function $f$ remains to be defined.

The fitnesses $w(\boldsymbol{p}\,|\,\boldsymbol{s})$ of a protein $\boldsymbol{p}$ given the target $\boldsymbol{s}$ is assumed to be the product of the fitnesses of its amino acids, so that

$$w(\boldsymbol{p}\,|\,\boldsymbol{s}) = \prod_{l=1}^{L} \omega(p_l\,|\,s_l). \tag{2.9}$$

Denote by $\alpha_{s_l}$ the target amino acid of the $l$th site-type as determined by target $\boldsymbol{s}$. In the multiplicative scheme for protein fitness we require that

$$w(\boldsymbol{p}\,|\,\boldsymbol{s}) = \prod_{l=1}^{L} f(d(p_l\,|\,\alpha_{s_l})) = f\left(\sum_{l=1}^{L} d(p_l\,|\,\alpha_{s_l})\right). \tag{2.10}$$

The only functional form that meets this requirement is

$$\omega(\beta\,|\,s_\alpha) = \phi^{d(\beta\,|\,\alpha)}, \tag{2.11}$$

where for the fitness to decrease with chemical distance we also require that $0 < \phi < 1$. See Appendix A for additional properties of this physicochemical accuracy scheme important for our ability to analyse our results.

We refer to $\phi$ as the *missense tolerance parameter* or *missense selection parameter* because it determines the overall strength of selection against missense in proteins. Greater $\phi$ corresponds to more tolerance of selection to missense over all sites in all proteins. *Missense* is defined as the incorporation, by mutation or mistranslation, of the 'incorrect' amino acid at a given site, where the 'correct' amino acid at a site is usually defined by biochemical or fitness assays, or by multiple comparisons of natural sequence variation. Here, with the assignment of a unique amino acid to every type of site, the 'correct' amino acid of a site is defined by its type.

The number of proteins produced from a message for a given individual is assumed to be large. The fitnesses of different proteins from the same message, as created, for example, through translational ambiguity or error, are arithmetically averaged to determine overall individual fitness. Let the *codon usage*, $u(i\,|\,s)$, be the frequency of codon $i \in C_{\Pi}^{B}$ in sites of type $s \in S$ in a given message $\boldsymbol{m}$, satisfying $u(i\,|\,s) \geqslant 0$ for all $i \in C_{\Pi}^{B}$ and $s \in S$ and $\Sigma_{i \in C_{\Pi}^{B}} u(i\,|\,s) = l_s$ for all $s \in S$. The fitness $w(c,\boldsymbol{m})$ of an individual with message $\boldsymbol{m}$ and code $c$, with a corresponding target $\boldsymbol{s}$ and combined translation probability $cF_r$, may then be written as

$$w(c,\boldsymbol{m}) = \prod_{s \in S} \prod_{i \in C_{\Pi}^{B}} \left( \sum_{\alpha \in A} cF_r(\alpha \mid i)\omega(\alpha \mid s) \right)^{u(i \mid s)}. \qquad (2.12)$$

Our use of the term 'codon usage' is slightly unconventional. We do not incorporate selection on synonymous codon usage.

### (d) *The initial code*

We model the evolution of coding assignments after the biochemistry is in place to carry out translation of any codon.

On the grounds that a high frequency of mutation to stop codons is deleterious in any coding system (Sonneborn 1965), we assume an initial state where all codons have sense. Extending from the logics of Woese (1965*a*) and Fitch (1966), who argued for an early sloppiness and ambiguity of the primitive translational machinery, we assume a *uniformly ambiguous initial code*. The initial code $c_0$ is such that every codon in $C_{\Pi}^{B}$ encodes all of the 20 amino acids in $A$ with equal probability, i.e.

$$c_0(\alpha \mid i) = \frac{1}{M} \text{ for all } i \in C \text{ and } \alpha \in A, \qquad (2.13)$$

where $M$ is the total number of amino acids in $A$.

Sonneborn's argument that most codons evolved some kind of amino acid meaning very early in code evolution, and the arguments of Woese (1965*a*,*b*) and Fitch (1966) for early ambiguity certainly need not imply the uniform ambiguity that we assume as the initial coding state of all codons here. The uniformly ambiguous initial condition initializes evolution in an unbiased manner with respect to idiosyncratic factors that may have had a particularly strong influence at the very beginning of code evolution, such as stereochemical affinities, stochastic factors in code-message coevolution, and historical constraints, and coevolution of the genetic code with amino acid biosynthesis.

### (e) *Code mutation*

We assume *uniform discrete code mutation* in which:

(i) The meaning of a codon $i$ may only change to encode a single amino acid with unit probability, i.e. $c(\alpha \mid i) = 1$ for some amino acid $\alpha$, and $c(\beta \mid i) = 0$ for any other amino acid $\beta \neq \alpha$.
(ii) Codon meanings change independently of one another within and across generations.
(iii) The probabilities of changes in meaning are identical over all codons and all amino acids.
(iv) The probability for a mutation in the meaning of one codon $\mu_{cm}$ is very small, so that changes in the meaning of two or more codons at once can be ignored.

The probability $\mu_{\text{code}}(c' \mid c)$ of mutating to a code $c'$ from a code $c$ so that $c'$ differs from $c$ by one allowed change in codon meaning, is then

$$\mu_{\text{code}}(c' \mid c) = \begin{cases} 1 - N\mu_{cm} & c' = c \\ \mu_{cm} & c' \neq c \\ 0 & \text{otherwise} \end{cases}, \qquad (2.14)$$

where $N = 16$ is the number of codons in $C_{\Pi}^{B}$. Thus, exactly $N \times M$ mutant codes appear from a given genetic code at any one time, according to our mutation scheme.

Note that codons may not mutate back to the initially ambiguous state. Once such a change has occurred to a codon so that it is no longer in its initial ambiguous state, we say that the codon has come to have an *explicit* meaning.

Our initial assumptions against double code-mutation and for discrete code-mutation are for computational and analytical tractability. Our framework becomes a good description for code evolution when codon usage across messages stabilizes (unpublished data). This codon usage stabilization occurs after at least a few codons encode specific amino acids. In our system this occurs a few coevolutionary steps after the uniformly ambiguous initial condition. Thus, the uniformly ambiguous initial condition avoids bias in any specific direction of code evolution.

### (f) *The code-message coevolutionary dynamic*

With the assumptions outlined in § 2a–c, an infinite-sized asexual population with a unique genetic code $c$ will converge to a unique codon usage distribution $U(c) = \{u_c(i \mid s)\}_{s \in S}^{i \in C_{\Pi}^{B}}$ in messages, and a unique growth rate $\lambda(c)$ in mutation-selection equilibrium (Sella & Ardell 2002). Also, assuming that messages are long, the variance in codon usage among individual messages in the population is small (Sella & Ardell 2002). We call the unique code with which an equilibrium population translates its messages the *established code*.

We use these results and equation (2.12) to calculate the *invasion fitness* of a mutant code, which is the fitness of an individual using that mutant code to translate a message that has the expected equilibrium codon usage of the established code. The invasion fitness of an individual with the altered genetic code $c'$ (combined translation probability $c'$ $F_r$) and the *typical message* $\boldsymbol{m}_c$ with the expected equilibrium codon usage $U(c)$ of the established code $c$ is

$$w(c',\boldsymbol{m}_c) = \prod_{s \in S} \prod_{i \in C_{\Pi}^{B}} \left( \sum_{\alpha \in A} c'F_r(\alpha \mid i)\omega(\alpha \mid s) \right)^{u_c(i \mid s)}. \qquad (2.15)$$

The coevolutionary process on codes and messages proceeds in a series of *steps*, in which first a small number of mutant codes compete to invade and takeover a population of messages equilibrated to an established code (starting with the initial code), assuming that the message distribution does not change; then, if one such mutant code is successful, messages equilibrate in mutation and selection to it as the new established code. The assumption that messages have time to equilibrate before codes change, and that codes change slowly enough to allow a small number of variants to accumulate, but that messages do not change during the sweep of the new mutant code to fixation, is called the *quasistatic approximation*.

More specifically, if at any step $w(c',\boldsymbol{m}_c) > \lambda(c)$ for some mutant code $c'$, the mutant code with the greatest invasion fitness is assumed to takeover the population and become the new established code (in our implementation, the first code seen that ties for maximal invasion fitnesses is non-randomly chosen). If $\lambda(c) \geqslant w(c',\boldsymbol{m}_c)$ for all

mutant codes $c'$, then no mutant code invades and the established code is said to have *frozen*.

### (g) *Methods*

The CMC program flexibly implements code-message coevolutionary dynamics in C++. Class inheritance is used to optimize run-time efficiency in runs initiated without translational misreading. Eigensystem solutions for determining the growth rate and codon usage associated with a genetic code are approximated using the iterative method (Press *et al.* 1988). In iterating the evolutionary process, eigenvalues and eigenvectors are found to a fixed precision, which is $10^{-16}$. A convergence theorem due to Varga (1962) was used to accelerate the comparison of growth rates of two genetic codes. The CMC algorithm is as follows

```
initialize  genetic code c ⇐ c_0.
do:
    solve  equilibrium growth rate λ(c) and codon usage U(c) of code c.
    initialize  w_max = 0.
    foreach  allowed mutant genetic code c':
        calculate  fitness w(c', m_c) with typical message m_c of code c.
        save  mutant code fitness w_max ⇐ w(c', m_c) and
            mutant code c* ⇐ c' if w(c', m_c) > w_max.
    update  genetic code c ⇐ c*  if w_max > λ(c).
until  w_max ≤ λ(c).
halt
```

The results of a run of code-message coevolutionary dynamics with CMC depends on the specific amino acid coordinates chosen for that run. Our approach to controlling for the idiosyncratic effects of specific amino acid–site-type spaces was to run the coevolution on many different randomly distributed sets of amino acid coordinates and then average the results we obtained over these different amino acid spaces.

Mutant codes were generated in a predetermined and consistent order, which was the same for any given model at every time-step in every independent iteration of the dynamics. In the event of ties in the maximal invasion fitness of mutant codes, the first mutant code to be generated which had that fitness was arbitrarily selected to invade and takeover the population at the next time-step.

We use two methods to quantitate the degree of structure preservation in codes: a partitioning of variance method and a previously published permutation method (Alff-Steinberger 1969; Haig & Hurst 1991; Ardell 1998; Freeland & Hurst 1998). The partitioning of variance method measures different patterns of amino acid similarities in the code. Codons are partitioned into sets, called 'groups', and the average variance of amino acid physicochemical coordinates within groups is calculated along with the variance of means among groups. The inverse of the ratio of these variances is a measure of the *structure* or *pattern* in the code corresponding to that grouping of codons.

For example, given a four-base, two-position genetic code, to measure the degree to which amino acids are more similar within groups of codons related by one or two transitions than between them, we define a partition $P_\kappa$ with four codon groups

$$\begin{aligned} P_\kappa &= \{\{\mathrm{UU,UC,CU,CC}\},\{\mathrm{AA,AG,GA,GG}\},\\ &\quad \{\mathrm{UA,CA,UG,CG}\},\{\mathrm{AU,GU,AC,GC}\}\}\\ &= \{P_{\kappa 1},P_{\kappa 2},P_{\kappa 3},P_{\kappa 4}\}. \end{aligned} \tag{2.16}$$

Then, for a code $c$, if $m_{\kappa j} = \Sigma_{i \in P_{\kappa j}} c(i)/|P_{\kappa j}|$ is the mean of amino acid coordinates encoded by codon set $P_{\kappa j}$ of size $|P_{\kappa j}|$ and $\mathrm{var}(P_{\kappa j}) = \Sigma_{i \in P_{\kappa j}}(c(i)^2/|P_{\kappa j}|) - m_{\kappa j}^2$ is the variance of those coordinates, then the structure $F$ of a code $c$ associated with the partition $P_\kappa$ is

$$F = \frac{\mathrm{var}(m_{\kappa j})}{(1/|P_\kappa|)\sum\limits_{P_{\kappa j} \in P_\kappa} \mathrm{var}(P_{\kappa j})}, \tag{2.17}$$

where the denominator is the variance of the means over all the sets in the partition.

In the permutation method, given a code with some pattern of redundancy which is evolved in CMC, the program may be configured to generate $10^4$ permuted codes with the same level and pattern of redundancy by randomly permuting the set of encoded amino acids among the redundant sets of codons in that code. The fraction of permutant codes more physicochemically conservative (as measured by the average absolute difference in physicochemical distance between the amino acids encoded) than the evolved final code along each of the four codon dimensions corresponding to transitions or transversions in the first or second codon position is then recorded. Although Ardell (1998) and Freeland *et al.* (2000) have argued for attention to be paid to the modular power to which distances are raised (using squared or absolute-values distances, for example), what is important for the data we show here is that we are consistent in comparing values that were calculated with the same modular power.

We also examined the effects of variation in transition bias ($\kappa$) and missense selection ($\phi$) on the number of encoded amino acids ($N_{\mathrm{aa}}$) and a measure of the physicochemical diversity in encoded amino acids called the *normalized encoded range* (NER). The NER of a code is the range in amino acid space of its explicitly encoded physicochemical properties normalized by the maximum range for the entire amino acid space with which it evolved. Denote by $c(C_{\Pi}^{B}) \subseteq A$ the set of amino acids explicitly encoded by a genetic code $c$ on the codon set $C_{\Pi}^{B}$, and by $d_A(\beta|\alpha) = |\beta - \alpha|$ the physicochemical distance between any two amino acids $\alpha, \beta \in A$.

The NER of a code $c$ given an amino acid space $A$ and associated physicochemical distance $d_A$ is

$$\mathrm{NER} = \frac{\max_{\alpha,\beta \in c(C_{\Pi}^{B})} d_A(\beta|\alpha)}{\max_{\alpha,\beta \in A} d_A(\beta|\alpha)}, \tag{2.18}$$

thus, a code where all codons encode the same amino acid would have an NER = 0 and a code where one or more codons encode the most extreme amino acids on the particular amino acid space with which it evolved would have NER = 1.

## 3. RESULTS

### (a) *The coevolution of genetic codes with transition-biased message mutation and no translational misreading*

We begin by studying the pattern induced in a genetic code coevolved with messages mutating with transition bias. With transition-biased mutation, the codon space consists of four blocks (see step 0 in figure 2), corresponding to first- and second-position pyrimidines and purines.

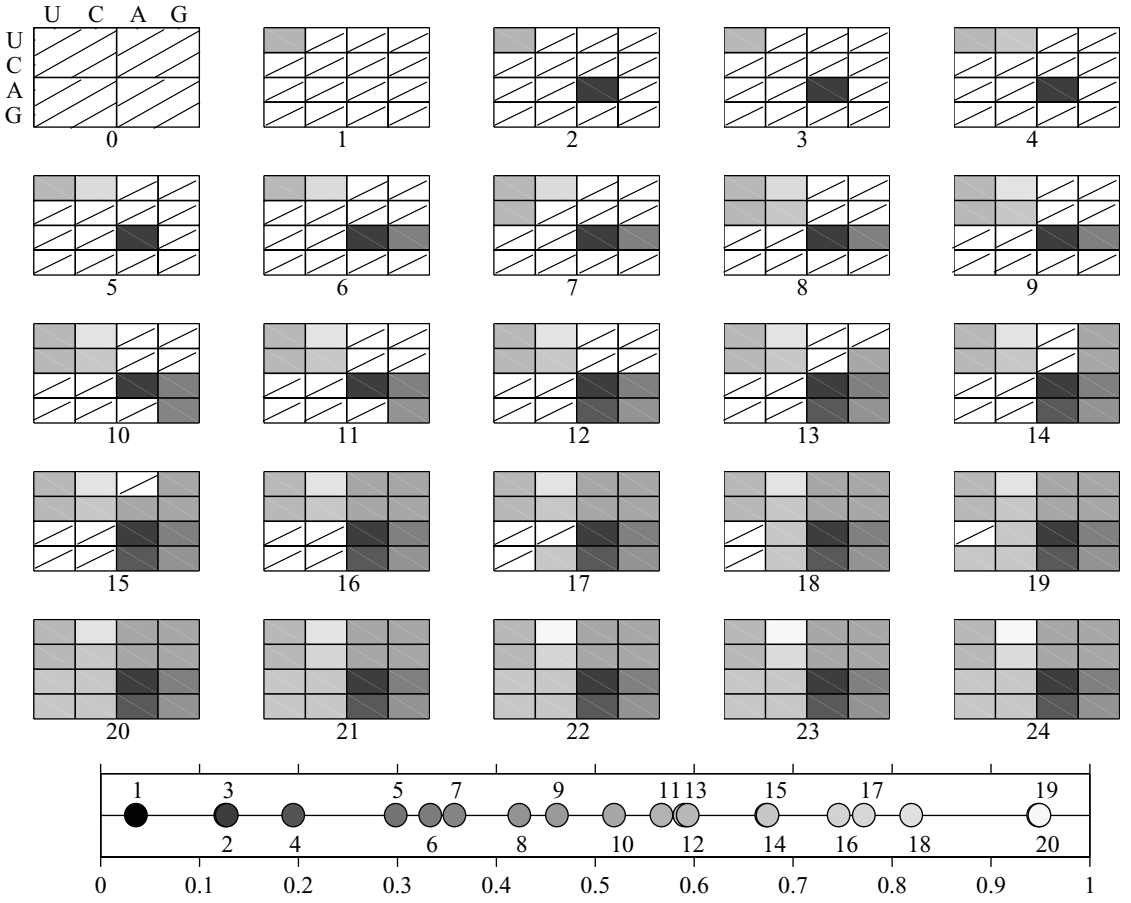


Figure 2. A typical code-message coevolutionary trajectory with transition-biased message mutation (transition bias $\kappa = 7$, mutation rate $\mu = 0.0006$ and missense tolerance $\phi = 0.92$) and the indicated randomly generated amino-acid–site-type space. The sequence of established mutant codes is shown from the initial fully ambiguous state (indicated by hatching) at time-step 0 to the final frozen code at time-step 24. The first position varies in rows and second position varies in columns. The equilibrated codon usage corresponding to the code at each time-step is not shown. The assigned and reassigned codons are shaded according to the amino acid they encode as in figure 1. See § 3a for a detailed description.

Within a block each codon has two closest neighbours that are one transition away, and one neighbour that is two transitions away. Each block has two adjacent blocks that are one transversion away, and an 'antipodal' block that is two transversions away. This structure in codon space partly determines the course of evolution by defining the regions of codon space across which 'load-minimizing' and 'diversifying' steps occur. In load-minimizing steps, codons that mutate to one another at high frequency are assigned or reassigned physicochemically similar amino acids. In diversifying steps, codons that mutate to one another infrequently are assigned or reassigned physicochemically dissimilar amino acids.

The code evolution with transition bias in mutation described in figure 2 can be explained in the following heuristic terms.

$t = 1$ The first amino acid to be encoded is number 11 with coordinate 0.563 178, approximately in the 'middle' of the one-dimensional chemical range. As all codons are used equally in all sites in the uniformly ambiguous initial code, amino acid 11 invades with highest fitness because it is the most versatile amino acid in the amino acid space, with the highest geometric mean fitness across sites. Any codon could have been assigned this amino acid with equal fitness, so the first codon examined, codon UU, broke the tie.

$t = 2$ The second mutant code to invade assigns amino acid 3, with coordinate 0.123 515, to a codon which is two transversions away from the first codon to be assigned. This is a diversifying assignment, in that the second codon to be assigned is as mutationally as far away in codon space as possible, and the amino acids are far away in amino acid space.

$t = 3$ Reassignments occur very early in this coevolutionary trajectory; this is an example of *early diversification*. In this step, the first codon to be assigned, codon UU, is reassigned from amino acid 11 with coordinate 0.563 178 to amino acid 13 with coordinate 0.589 559. This *diversifying reassignment* increased the encoded range. It came about because the assignment of an amino acid with a low coordinate to codon AA in the last step caused codon UU to be out-competed in sites selecting for low-coordinate amino acids; this narrowed the usage of the codon UU to a more specialized set of sites. Within this

now more restricted set of sites, amino acid 13 had the highest fitness of any amino acid including the old meaning of codon UU, and this reassignment was the most fit change that could occur in the code at this stage.

$t = 4,5$ The meaning of codon UC undergoes two diversifying changes, extending the encoded range at the high end of amino acid space. In step 4, it is originally assigned amino acid 15, with coordinate 0.670 356, then immediately afterwards, in step 5, it is reassigned the even more diverse amino acid, namely 17, with coordinate 0.767 591. The second change is facilitated by the redistribution of codons in sites after the first change.

$t = 6,7,8$ Load-minimizing assignments occur within the encoded range, so that the range of physicochemical diversity does not expand.

$t = 9$ The previous assignments after step 5 allow a second diversifying reassignment in codon UC, now to amino acid 18 with coordinate 0.815 486. This codon will be reassigned yet again to amino acid 19 in step 22, with coordinate 0.943 856, just before the code freezes. This third reassignment at step 22 yields the most diverse amino acid to be encoded at the high range of amino acid space.

$t = 12$ Most of the diversity in amino acids that will become encoded has become encoded by two blocks of codons separated by two transversions.

$t = 13$–$20$ All subsequent assignments are in the middle of the range of site-type space, and result in redundant (or nearly redundant) amino acid assignments within each of the other two 'blocks'. The block in the upper-right corner is completely redundant for amino acid 10 (coordinate 0.515 184), while the block in the lower-left encodes amino acids 14 and 15 (0.668 897 and 0.670 356, respectively). Amino acid 14 is already encoded by codon CC.

$t = 21$ The encoding of amino acid 14 by codons AC and GC reduced the usage of codon CC in sites where amino acid 14 had high fitness, because of redundancy (not shown). This released a diversity constraint on codon CC, causing a *load-minimizing reassignment* of CC in this step from amino acid 14 to amino acid 15. This code change is favoured because it reduces the load of CU codons mutating to CC in the extreme sites where the amino acid CU has high fitness. However, the reassignment moves the meaning of CU away from that of its other transition neighbour UC. At equilibrium with the new code the usage of CC is much greater in sites where UC has high fitness than in sites where UC has high fitness (not shown). In all, we see here an example where prior code changes redistributes codon usage, facilitating subsequent reassignments of meaning in the code.

$t = 22$ The load-minimizing reassignment of the last step reduces the usage of codon UC in sites in which the amino acid it encodes—18—had low fitness, and increases its usage where it has high fitness, at the high end of site-type space. This facilitates the diversifying reassignment of codon CU to an even more extreme amino acid, namely amino acid 19.

$t = 23$ The diversifying reassignment of UC in the last step also causes the reassignment of its mutant neighbour CC.

$t = 24$ The code freezes with block structure in its coding assignments, so that most amino acid variation occurs among blocks of codons rather than within.

The pattern required for the correction of systematic errors of transitional mutational bias is precisely the pattern that these errors induce through code-message coevolution. It is the pattern of blocks shown at the end of the evolutionary trajectory shown in figure 2.

Figure 3 shows quantitatively how this block structure of final frozen codes evolved with transition bias in mutation increases or decreases with the amount of missense selection ($\phi$) and transition bias in mutation ($\kappa$), using the partitioning of variance method (see § 2g). The graphs include the $\kappa = 1$ case of no transition bias. The evolution of the error-correcting block structure in final frozen codes shown in the example of figure 2 is robust to variation in these parameters. Both the variance among blocks and block structure (the ratio of variance among blocks to variance within blocks) increase with transition bias and increase with missense selection.

As missense selection becomes very weak (right column) in figure 3, the ratio of variances increases dramatically, because the variance within blocks drops to zero. Thus, there is a part of parameter space where practically all amino acid diversity is encoded among, but not within, the four blocks of codon space. At the weakest selection, even this small pattern of diversity disappears and the code is almost completely redundant.

In Ardell & Sella (2001), we showed that the average encoded amino acid number and diversity (as expressed in NER) in frozen genetic codes decreases as selection weakens or mutation increases, with the result of near-total redundancy at extremely weak missense selection. We call this phenomenon the *encoding catastrophe* (Ardell & Sella 2001). This is also at work in the case with transition-biased mutation: at one level of weak selection, amino acid diversity is lost within each block, like mini-encoding catastrophes within blocks, within which codons mutate at a higher rate. Then, at even weaker selection, diversity disappears among blocks, among which codons mutate at a lower rate.

Figure 4 shows the average encoded amino acid number and diversity of amino acids in final frozen codes for different levels of transition bias ($\kappa$) as a function of missense selection ($\phi$). We see that, relative to the case without transition bias ($\kappa = 1$), for relatively strong missense selection at a given mutation rate, transition bias actually *decreases* both the number and NER of encoded amino acids in final frozen codes. With weaker selection (higher $\phi$), the behaviour inverts and transition bias maintains amino acid diversity relative to the $\kappa = 1$ case. The encoded range NER is more stable to this behaviour than the number of encoded amino acids $N_{aa}$. There is also the suggestion of an inversion in the effect of transition bias

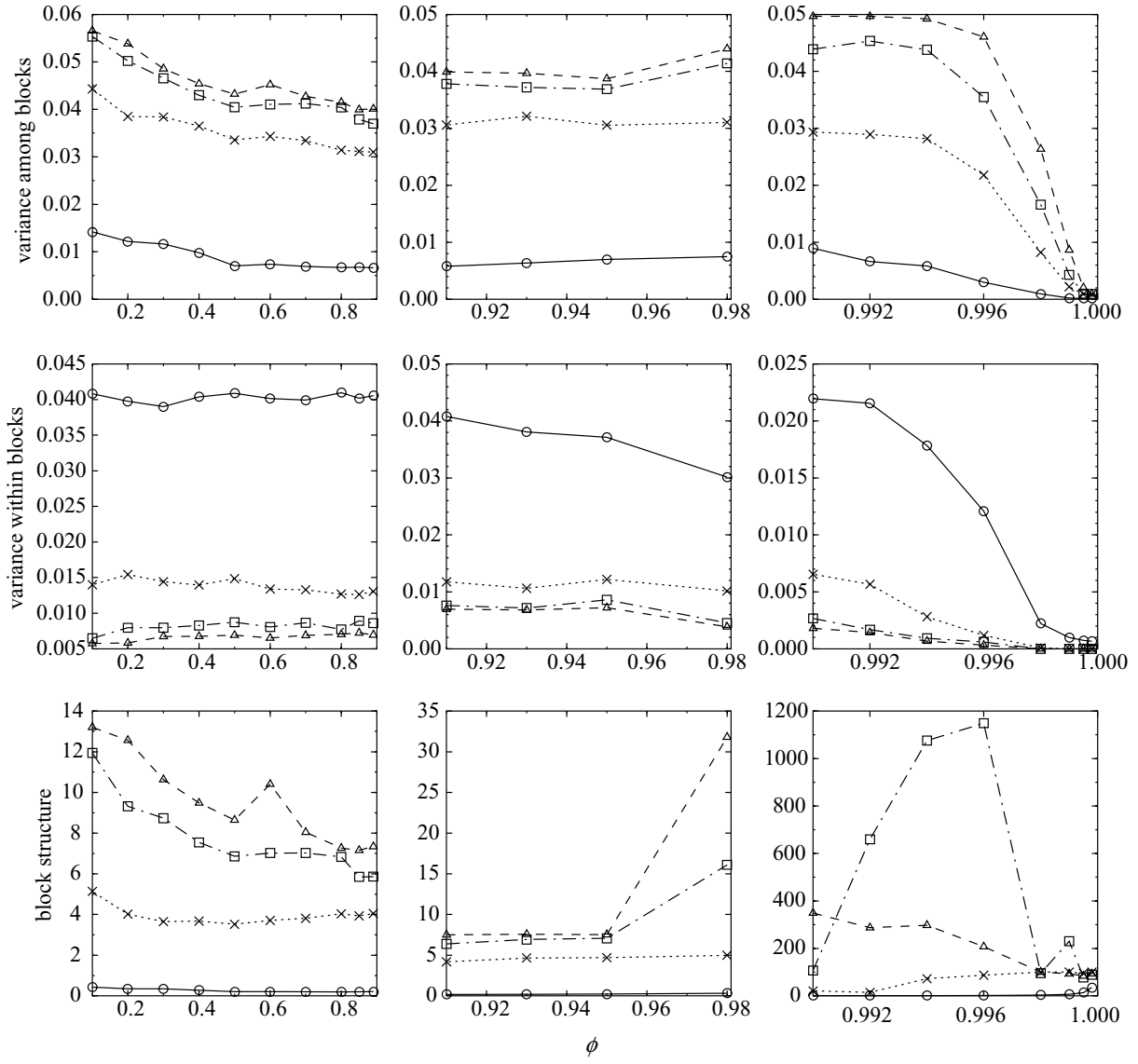


Figure 3. Average final variance of encoded amino acid properties among and within blocks of codons separated by one or two transition mutations, and their average ratio ('block structure') among 60 frozen genetic codes, as a function of missense tolerance ($\phi$). The 60 codes corresponding to a given point on the graphs coevolved with messages mutating at the same values of the transition bias, $\kappa$, and missense tolerance, $\phi$, at the fixed mutation rate $\mu = 0.0006$, but with a different random amino-acid–site-type space. The subgraph columns show three non-overlapping ranges of $\phi$. Triangles, $K = 6$; squares, $K = 4$; crosses, $K = 2$; circles, $K = 1$.

at very strong missense selection (left-hand side of left column of subgraphs). These results indicate that the effects of more complicated mutation and selection schemes on code-message coevolutionary dynamics may be able to be approximated by decomposing these schemes into combinations of simpler schemes.

### (b) *The coevolution of genetic codes with translational misreading and no transition bias in mutation*

A typical evolution with uniform positional misreading in the first codon position, with the same site-type–amino-acid space and mutation and selection parameters as in the previous example, is shown in figure 5. The evolutionary generation of the column pattern, which corresponds to the error-correcting requirement for this type of systematic error, can also be understood in the heuristic terms of load-minimizing and diversifying steps. But in contrast to how they are used in the analysis of the typical evolution in figure 2, the terms are in relation to both misreading and mutation. For instance, the assignment of similar codons to codons likely to be misread or mutate to one another is 'load-minimizing'.

### (c) *The coevolution of genetic codes with transition-biased message mutation and translational misreading*

Next, we combined the last two models into one. We found that the pattern generated by evolution varies in accordance to the relative magnitudes of misreading and mutation. An array of final codes corresponding to different combinations of mutation and misreading parameters for moderate transition bias $\kappa = 5$ is presented in figure 6. On the top left corner, where mutation dominates over misreading, the four-block structure corresponding to transition bias in mutation is clearly pronounced. On the

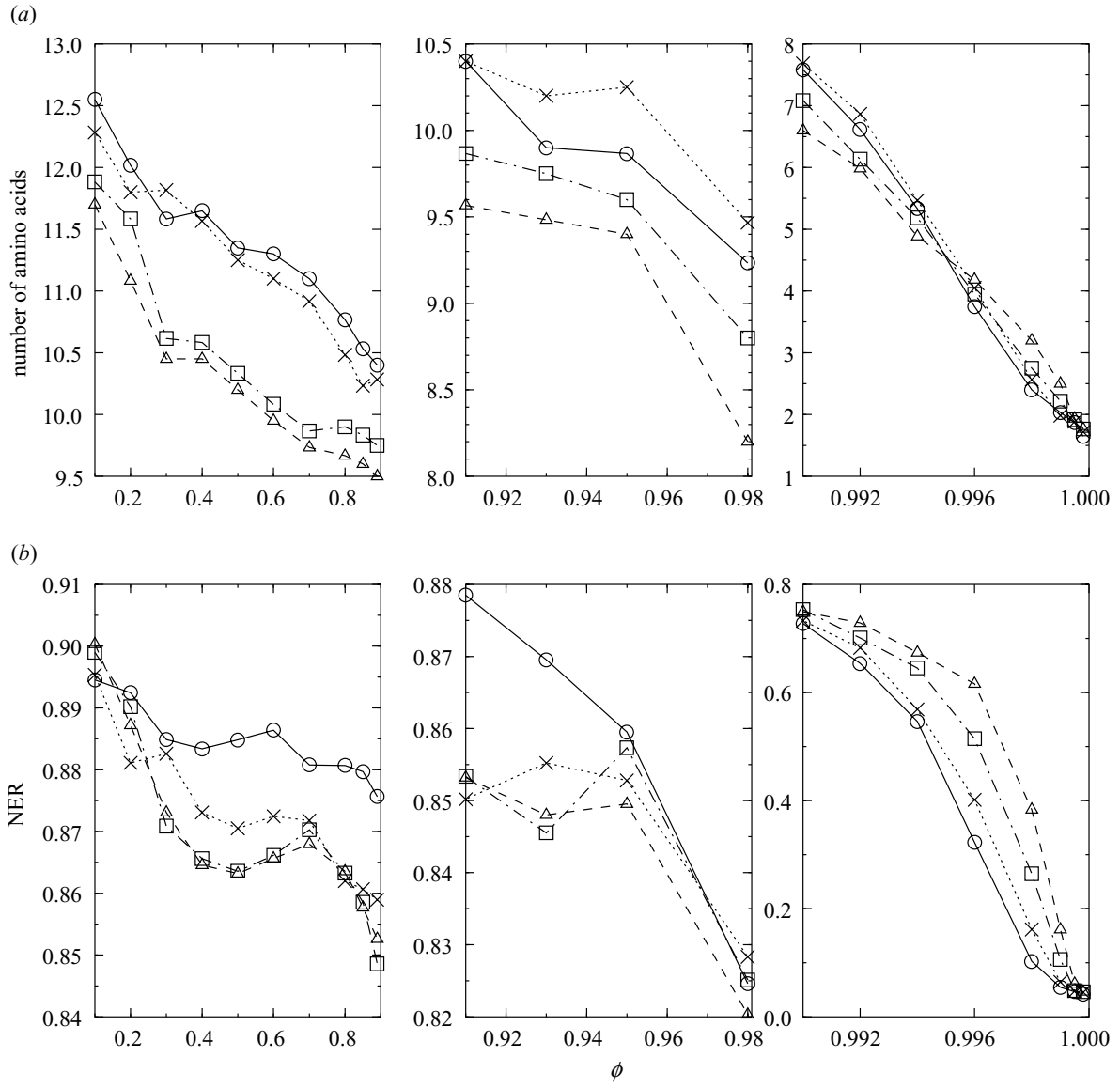


Figure 4. Average final number ($N_{aa}$) (*a*) and NER (*b*) of encoded amino acids among 60 frozen genetic codes as a function of missense tolerance ($\phi$). Each point is the average outcome among 60 frozen codes that had coevolved with messages mutating with transition bias $\kappa$ shown, fixed mutation rate $\mu = 0.0006$, and a distinct randomly generated amino-acid–site-type space. Subgraph columns show three non-overlapping ranges of $\phi$. Triangles, $K = 6$; squares, $K = 4$; crosses, $K = 2$; circles, $K = 1$.

bottom right corner, where misreading dominates over mutation in the first codon position, the model generates the 'two-column' structure corresponding to properties I and II of the standard genetic code.

The array of final codes shown in figure 6 has a striking regularity in the orientation of amino acids. Almost all codes have 'dark' amino acids in the upper-left corner and 'light' amino acids in the lower-right corner. This is an artefact of the arbitrary breaking of ties in invasion fitness that we used, which always favoured the upper-leftmost codons. Finally, we reversed the (arbitrarily oriented) colour scale to call attention to the similarity to the structure of the standard genetic code in figure 1.

The outcome of a systematic study of code organization as a function of mutation and misreading rates is presented in figure 7, where the patterns in final codes evolved with a stronger transition bias ($\kappa = 7$) are characterized by the permutation method (see § 2). Figure 7 shows in detail that the shift that occurs with increasing misreading, from a 'four-block' to a 'two-column' pattern, is quite general at all values of mutation. Before the shift, at low rates of misreading in the first position, evolved codes are specifically conservative with respect to transitions in both the first and second codon positions. Codes are more conservative in all dimensions than the median of randomized codes (corresponding to 0.5 on the respective *y*-axes) because code-message coevolution with transition-biased mutation separates most of the amino acid diversity across two transversions. As there is no clear effect from misreading in these final codes, we say that the coevolution with these low values of positional misreading is 'mutation dominated'.

After the shift, at high rates of misreading, a two-column pattern emerges, where the first position becomes highly conservative in both transitions and transversions. The largest extent of physicochemical diversity is allocated

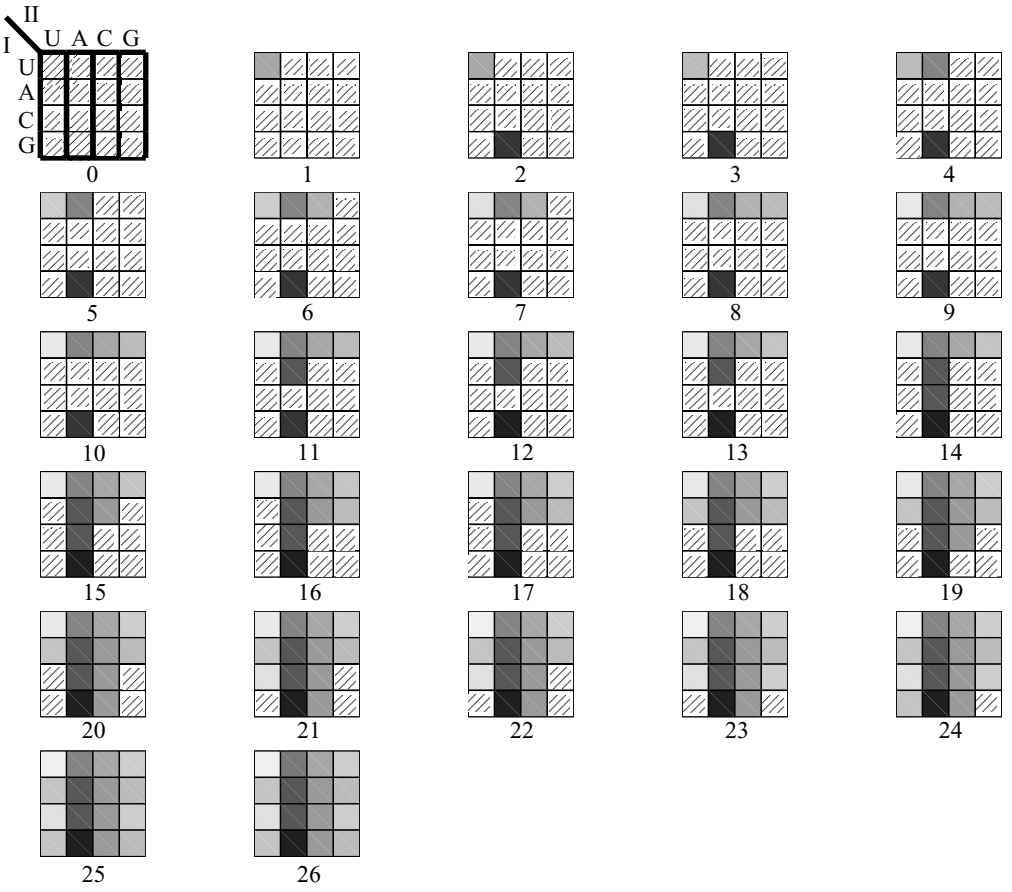


Figure 5. A typical code-message coevolutionary trajectory with uniform misreading in the first codon position (first-position misreading rate $e_1 = 0.006$, missense tolerance $\phi = 0.92$, and mutation rate $\mu = 0.0006$) and no transition bias ($\kappa = 1$). The first position varies in the columns and the second position in the rows. Hatching indicates the initial uniformly ambiguous coding state of codons. The first step is similar to the previous example for reasons described in § 3a in relation to figure 2. Steps 2, 6 and 8 are diversifying. Steps 7, 11, 14, 16 and 19–25 are the misreading analogue of load-minimizing steps. Steps 3, 5, 7, 9, 10, 12, 13, 22 and 26 are all reassignments, which are also described further in relation to figure 2.

along second-position transversions. The transition-biased structure actually becomes amplified in the second position. These two differences result from the high rate of uniform misreading in the first codon position, which causes a high apparent usage of first-position codon neighbours within sites ('apparent' with respect to code mutant invasion fitnesses). The reason for the amplification in the second codon dimension is that the high apparent usage of first-position neighbours causes a high frequency of load-minimizing code changes in this dimension. With the loss in reliability along the first position that is uniform in both transitions and transversions, the second codon position must incorporate the entire extent of final encoded physicochemical diversity. Just like in the standard code, second-position transversions become the most reliable dimension in the code within which to incorporate diversity. As this change in code-organization is due to the effect of the positional asymmetry in translational misreading, we say that the coevolution is 'misreading dominated'.

As one might expect, the shift between the two behaviours occurs with higher misreading parameters for higher mutation rates. The shift for a series of evolutions using the same amino-acid–site-type space is sharply discrete with increase in the misreading parameter, rather than smooth, as in figure 7, owing to the averaging procedure (data not shown). Note that at high levels of misreading, transitions are half as conservative as transversions according to the expectations discussed in § 1.

We see from figure 8 that the behaviour indicated in figure 7 semi-quantitatively reproduces patterns I and II of the standard genetic code listed in § 1. As in figure 6, figure 7 shows how the permutation criterion behaves as a function of misreading in the first position, except here we show the behaviour for different values of transition bias. We see that our evolved codes, just like the standard code, are not optimal by the permutation criterion. The average behaviour of final codes fits the estimated values of the standard code rather well for high rates of first-position misreading and a range of values of transition bias. At high misreading rates, transversions become about twice as conservative as transitions at all values of transition bias in the first codon position, which when compared with the behaviour at no misreading ($e_1 = 0$, left most part of subgraphs) and no transition bias ($\kappa = 1$, circles and solid lines), is as if there were no transition bias in mutation at all. Thus, at high levels of misreading, the transition-biased pattern of codon usage is filtered out with respect to the invasion fitness of code alterations, along the lines indicated by Ardell (1998).

It is worth emphasizing again that the genetic codes that we evolved were not optimal. This was true not only by

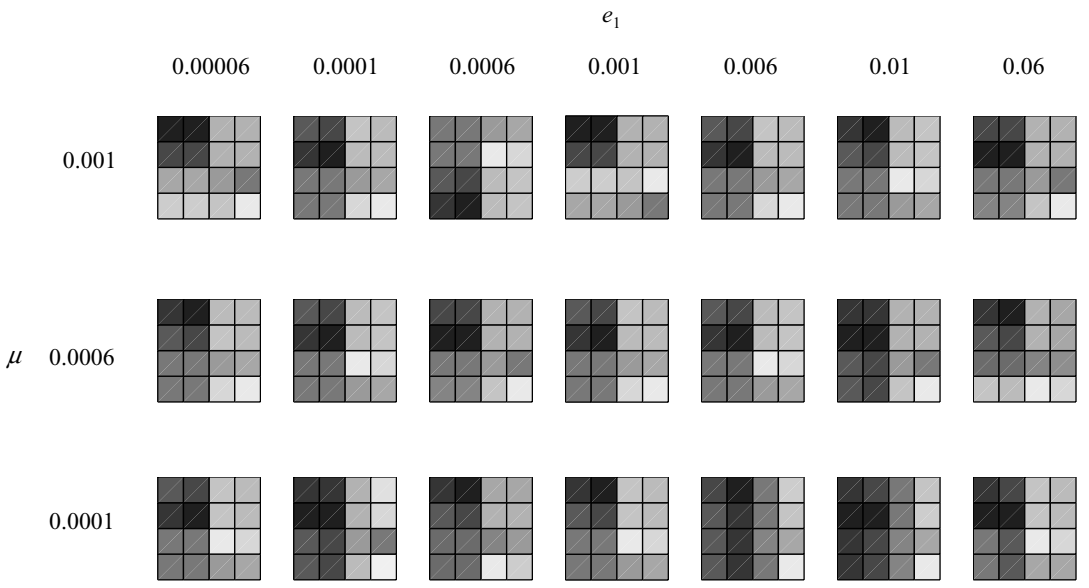


Figure 6. An array of final frozen genetic codes each coevolved with the same amino-acid–site-type space but a different parameter combination of the message mutation rate $\mu$ and the first-position misreading rate $e_1$. The transition bias in mutation $\kappa$ was taken to be $\kappa = 5$ for all runs. The colour scaling has been (arbitrarily) reversed relative to the figures and to reinforce the similarity in the evolved pattern of codes in the lower-right hand corner to that of figure 1.

the permutation criterion, but also in terms of relative fitness, in part because our final evolved genetic codes always evolved with redundancy (Ardell & Sella 2001). As another demonstration of this, we compared the growth rates of a frozen final code evolved with our method and a hand-designed genetic code with no redundancy. We used the following parameters: $\mu = 0.0006$, $\phi = 0.92$, $e_1 = 0.01$. The evolved code froze with 10 different encoded amino acids and a NER of about 0.81. We then calculated the growth rate of a designed code with the same parameters and amino acid space but where every codon was explicitly assigned a different amino acid and the encoded range was 0.91. The designed code was frozen, no code change could invade its messages. The calculated ratio $\lambda_d\lambda_e^{-1}$ of the growth rates of the designed ($\lambda_d$) and evolved ($\lambda_e$) codes was $\lambda_d\lambda_e^{-1} = 1.001\,0554$. We went further and counted the number of randomly permuted codes of the evolved code that had a higher growth rate with the same number and encoded range of amino acids, and the same parameters. There were four out of 10 000 randomly permuted codes with a higher growth rate than the typical evolved code. These results prove the non-optimality of the typical evolved code.

## 4. DISCUSSION

We have demonstrated that code-message coevolution tends to produce structure-preserving codes, and in particular can reproduce some of the structure-preserving patterns of the standard genetic code. Like the standard code (Wong 1980; Di Giulio 1989; Goldman 1993; Di Giulio *et al.* 1994), the codes that we evolved through code-message coevolution are substantially, but not optimally, error correcting. More generally, we have described a plausible mechanism by which natural selection can act on codes without disrupting the meaning of persistent messages. Our modelling framework provides a platform to model the coevolutionary dynamics of codes and messages.

Many important factors for genetic-code evolution have been simplified in our model. We wanted to provide a basis for the understanding of more complex models to follow. However, even with the relatively simple model we have described and studied here, we may have obtained insights to open questions regarding the origin of the genetic code. For instance, figure 7 may shed insight to a puzzle raised by Woese and others (Woese 1965a; Woese *et al.* 1966; Alff-Steinberger 1969; Swanson 1984; Haig & Hurst 1991; Goldman 1993), who argued that if the code had evolved to correct for mutational biases, because mutation is identical along different codon positions, one would expect the three codon positions to be equally conservative. Order-of-magnitude differences in rates of misreading in the first and second codon positions could have left the first and second codon positions in entirely different dynamic regimes, a possibility suggested by the sharpness of the transition between mutation- and misreading-dominated dynamics seen in figure 7. The sharpness of the transition was even greater when results for different amino-acid–site-type spaces were not averaged (not shown), and indicates that low misreading rates might be negligible. However, this sharpness may be an effect of always choosing the maximally fit invading code to invade, as we do here. Checking this result will require further study.

We simulated evolution for the data in figures 7 and 8 with no misreading in the second codon position as an approximation. Figure 7 indicates that, at least within our current assumptions, low rates of misreading in a position are effectively like no misreading in that position with respect to their effect on code organization. This may also be an artefact of our assumption that the maximally fit mutant always invades. Further study could confirm this. We also saw in figures 2 and 5 that, as in Ardell & Sella (2001), typical code-message coevolutionary dynamics involved a number of reassignments which, while typically physicochemically conservative, are much larger in num-

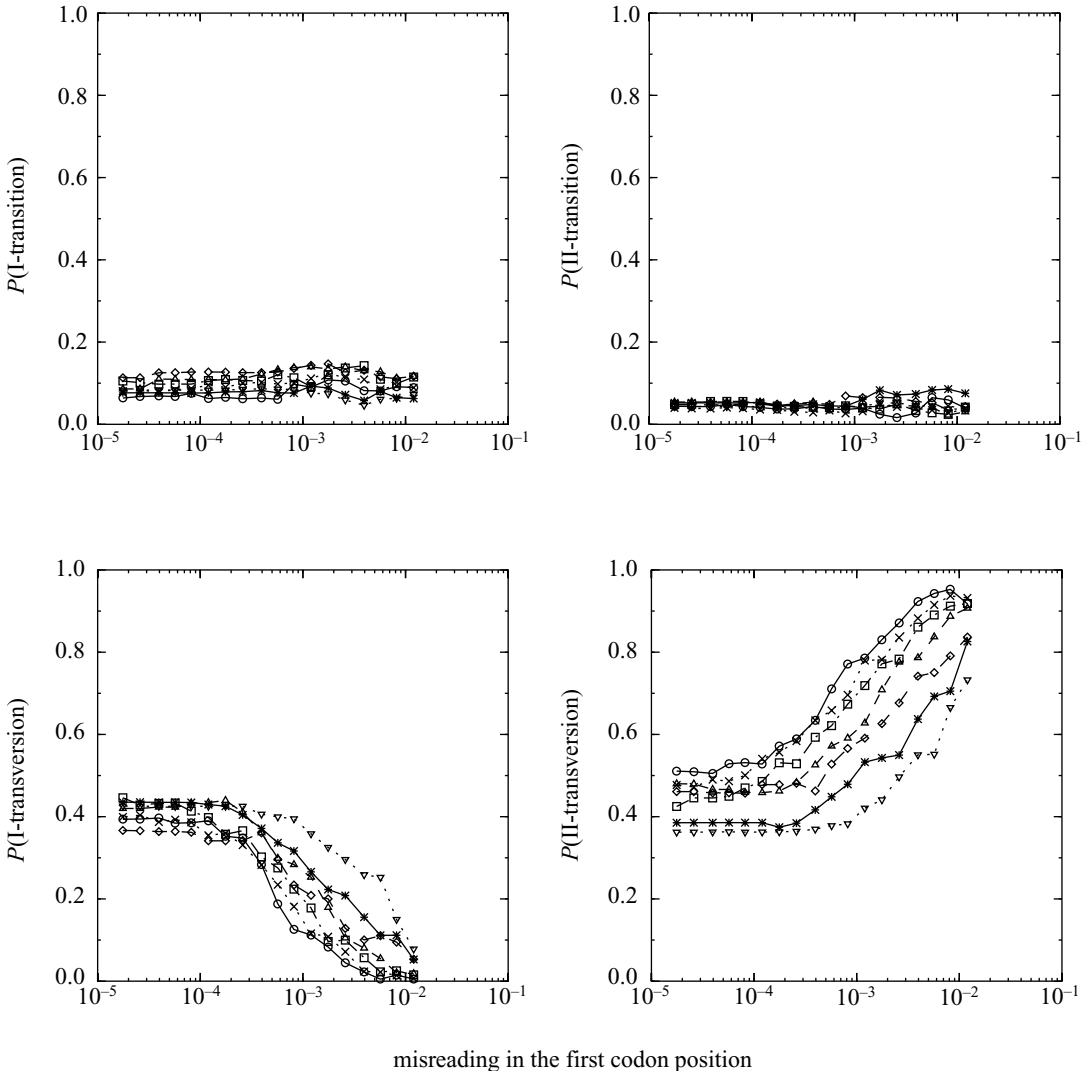


Figure 7. The transition of frozen genetic codes from the four-block pattern of codes coevolved with transition-biased message mutation but no misreading to the two-column pattern of codes coevolved with both transition-biased mutation and misreading in the first codon position, as a function of misreading rate. Each point is the average of 42 genetic codes evolved with a distinct random amino acid space, missense tolerance $\phi = 0.92$ and transition bias $\kappa = 7$. Values of mutation increase geometrically from $\mu = 1.2 \times 10^{-4}$ (circles) to $\mu = 1.2 \times 10^{-3}$ (inverted triangles) with: asterisks, $\mu = 8.18 \times 10^{-4}$; diamonds, $\mu = 5.7 \times 10^{-4}$; triangles, $\mu = 3.95 \times 10^{-4}$; squares, $\mu = 2.59 \times 10^{-4}$; crosses, $\mu = 1.76 \times 10^{-4}$. $P(X)$ denotes the fraction of randomly permuted codes that are more physicochemically conservative than the evolved codes along the $X$ codon dimension ('I' and 'II' indicate the codon position).

ber than might be expected from the arguments of Crick (1968) or Osawa *et al.* (1992). The large number of reassignments that we see may be an artefact of how we assume messages to completely equilibrate before codes change, as well as the assumed initially ambiguous state of codons (Ardell & Sella 2001).

We achieved our results without incorporating either stereochemical predetermination of the genetic code, its historical coevolution with amino acid biosynthesis, or evolution of the reading elements of the translational apparatus. By this omission we do not mean to imply that these factors did not influence the evolution of the genetic code. For instance, it is reasonable that stereochemical interactions may have played an even stronger part in early stages of evolution of the Standard Code (Knight *et al.* 1999; Di Giulio & Medugno 1999). It would be easy to incorporate these other factors into the model to study their effects. Further study is also required in the effect of variation of initial conditions, code-mutation operators (e.g. allowing codons to change through ambiguous states as in Schultz & Yarus (1994)), stochastic components in the model, and different forms of amino acid spaces, site-type spaces and targets. In whatever way that our model should be extended, we feel that a complete theory of genetic-code evolution will have to incorporate code-message coevolution. For instance, no matter how changes to a genetic code arise, such changes will reshape genotype frequencies, and these reshaped genotype frequencies will influence the fitness of future genetic-code changes.

Crick (1968) argued that selection on genetic codes for amino acid diversity, and to preserve message meaning, were much greater than selection could ever be for error correction. In the extreme, the evolution of the genetic code could have proceeded through a series of historical accidents, so that wherever a novel amino acid appeared in the code, if that code change could be tolerated by existing

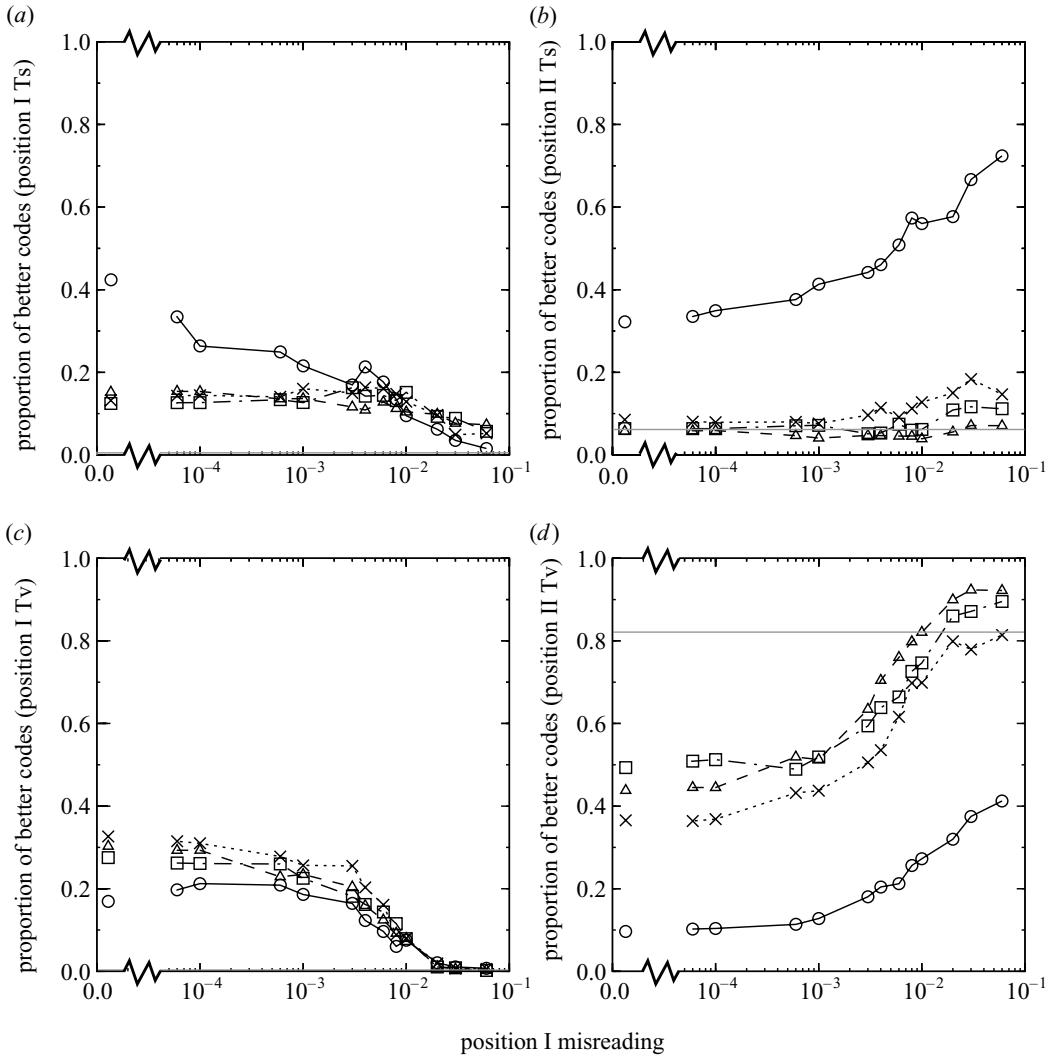


Figure 8. A comparison of final evolved genetic codes with the standard code (SGC). Each line shows the average proportion of randomly permuted codes that are more conservative than a frozen evolved code for different values of transition bias in mutation as a function of the misreading rate in the first codon position. Each point is the average among 40 independent evolutions over random amino acid spaces, $\mu = 0.0006$, and $\phi = 0.92$. Corresponding values for the SGC are indicated by stippled horizontal lines. Misreading is on a log-scale, with the values obtained with no misreading ($e_1 = 0$) included for comparison. (*a*) $P_{\mathrm{SGC}} = 6.1 \times 10^{-3}$; (*b*) $P_{\mathrm{SGC}} = 5.9 \times 10^{-2}$; (*c*) $P_{\mathrm{SGC}} = 1.1 \times 10^{-3}$; (*d*) $P_{\mathrm{SGC}} = 8.2 \times 10^{-1}$. 'Ts' means 'transition' and 'Tv' means 'transversion'. Triangles, $K = 7$; squares, $K = 5$; crosses, $K = 3$; circles, $K = 1$.

messages, it would be accepted. In the extreme then, the code could evolve no error-correcting structure at all, but would be a 'frozen accident' (it should be noted that Crick, like Woese, considered a wide spectrum of hypotheses). By formalizing and extending Crick's argument we have arrived at the opposite conclusion. Mutation, selection and translational error all bias the distributions of codons in messages in specific ways. These biased codon distributions tend to favour changes in genetic codes that result in error-correcting patterns. Not infrequently, the evolving patterns in these codes exacerbate the biased codon distributions that produced them. So it is the errors themselves, acting through the same freezing force described by Crick, that coax evolution towards an error-correcting frozen code. Selection for error correction is not incompatible with selection to preserve message meaning; rather, contingent on the parameters of evolution and other assumptions, they are one and the same.

Our work has been motivated by the belief that the patterns of the standard genetic code may be explicable as adaptations of a system of information processing. If this turns out to be plausible and correct, we may say that adaptations have reduced the deleterious consequences of genetic and physiological error at a very fundamental level of biological organization. We believe we have provided a plausible mechanism through which this could have occurred. Although we may find many information processing systems in biology that are robust to error, not all will rely on structure-preserving codes. For example, not all systems have as high a *density* of codewords that genetic codes have: in genetic codes, all possible codewords have meaning. A code that is not as dense can afford to 'pad' each of its codewords with invalid words that have no meaning, making the detection and even correction of transmission errors much easier before decoding even occurs. As another example, systems other than translation may not have a trade-off between rate efficiency and error efficiency like translation does; if they do, they may not favour rate at the expense of accuracy as much

as translation does (under certain circumstances, see Kurland *et al.* (1996)). It is especially when errors among valid signals cannot be efficiently corrected or detected during signal transmission, and when a certain tolerance to error already exists, that correction at the decoding level is called upon.

We hope that considerations such as these may shed light on adaptation in the evolution of information processing at other levels of biological organization.

The research of D.H.A. and G.S. was supported by NIH grants nos GM28016 and GM 28438 to M. W. Feldman. The research of G.S. was also supported by a Koshland Scholar award. The authors thank J. Swire and S. R. X. Dall for helpful comments on the manuscript.

## APPENDIX A: PROPERTIES OF THE PHYSICOCHEMICAL ACCURACY SCHEME

The physicochemical accuracy scheme we defined has the following properties.

(i) $\omega(\alpha\,|\,s_\alpha) = 1$ for any amino acid $\alpha$ and its corresponding site-type $s_\alpha$. In addition, if $\alpha \neq \beta$ then $\omega(\beta\,|\,s_\alpha) < 1$ for any amino acid $\beta \in A$ and site-type $s_\alpha \in S$.

(ii) Monotonicity at any site-type: if $d_A(\gamma\,|\,\alpha) > d_A(\beta\,|\,\alpha)$ then $\omega(\gamma\,|\,s_\alpha) < \omega(\beta\,|\,s_\alpha)$ for any amino acids $\beta,\gamma \in A$ and site-type $s_\alpha \in S$.

(iii) Uniformity of selection across site-types: if $d_A(\beta\,|\,\alpha) = d_A(\delta\,|\,\gamma)$ then $\omega(\beta\,|\,s_\alpha) = \omega(\delta\,|\,s_\gamma)$ for any amino acids $\beta,\delta \in A$ and site-types $s_\alpha$, $s_\gamma \in S$.

(iv) Scaling symmetry between chemical distance and the strength of selection: taking a selection coefficient $\phi' = \phi^\Lambda$ is equivalent to scaling the chemical distance $d' = \Lambda d$, since

$$\omega_{\phi',d}(\beta\,|\,s_\alpha) = (\phi^\Lambda)^{d(\beta|\alpha)} = \phi^{\Lambda d(\beta|\alpha)} = \omega_{\phi,d'}(\beta|s_\alpha). \qquad \text{(A 1)}$$

This follows from requirement (2.10).

All these properties may become useful in analysing and understanding evolution in this model.

## REFERENCES


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

Ardell, D. H. 1998 On error-minimization in a sequential origin of the standard genetic code. *J. Mol. Evol.* **47**, 1–13.

Ardell, D. H. & Sella, G. 2001 On the evolution of redundancy in genetic codes. *J. Mol. Evol.* **53**, 269–281.

Ash, R. B. 1965 *Information theory*. New York: Dover Publications.

Bain, J. D., Switzer, C., Chamberlin, A. R. & Benner, S. A. 1992 Ribosome-mediated incorporation of a non-standard amino acid into a peptide through expansion of the genetic code. *Nature* **356**, 537–539.

Bedian, V. 1982 The possible role of assignment catalysts in the origin of the genetic code. *Orig. Life* **12**, 181–204.

Cavalcanti, A. R., Neto, B. D. & Ferreira, R. 2000 On the classes of aminoacyl-tRNA synthetases and the error minimization in the genetic code. *J. Theor. Biol.* **204**, 15–20.

Crick, F. H. C. 1966 On protein synthesis. *Symp. Soc. Exp. Biol.* **12**, 138–163.

Crick, F. H. C. 1968 The origin of the genetic code. *J. Mol. Biol.* **38**, 367–379.

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

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

Denver, D. R., Morris, K., Lynch, M., Vassilieva, L. L. & Thomas, W. K. 2000 High direct estimate of the mutation rate in the mitochondrial genome of *Caenorhabditis elegans*. *Science* **289**, 2342–2344.

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. 1995 The phylogeny of tRNAs seems to confirm the predictions of the coevolution theory of the origin of the genetic code. *Orig. Life Evol. Biosph.* **25**, 549–564.

Di Giulio, M. 1997 The origin of the genetic code. *Trends Biochem. Sci.* **22**, 49.

Di Giulio, M. 2000 Genetic code origin and the strength of natural selection. *J. Theor. Biol.* **205**, 659–661.

Di Giulio, M. & Medugno, M. 1999 Physicochemical optimization in genetic code origin as the number of codified amino acids increases. *J. Mol. Evol.* **49**, 1–10.

Di Giulio, M., Capobianco, M. & Medugno, M. 1994 On the optimization of the physicochemical distances between amino acids in the evolution of the genetic code. *J. Theor. Biol.* **168**, 43–51.

Döring, V. & Marlière, P. 1998 Reassigning cysteine in the genetic code of *Escherichia coli*. *Genetics* **150**, 543–551.

Echols, H. & Goodman, M. F. 1991 Fidelity mechanisms in DNA replication. *A. Rev. Biochem.* **60**, 477–511.

Eigen, M., Lindemann, B. F., Tietze, M., Winkler-Oswatitsch, R., Dress, A. & Von Haeseler, A. 1989 How old is the genetic code? Statistical geometry of tRNA provides an answer. *Science* **224**, 673–679.

Ekland, E. H. & Bartel, D. P. 1996 RNA-catalysed RNA polymerization using nucleoside triphosphates. *Nature* **382**, 373–376.

El'skaya, A. V. & Soldatkin, A. P. 1985 The bases of translational fidelity (review). *Molekulyarna Biologiya* **18**, 1163–1180.

Epstein, C. J. 1966 Role of the amino-acid 'code' and of selection for conformation in the evolution of proteins. *Nature* **210**, 25–28.

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

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

Freeland, S. D. & Hurst, L. D. 1998 The genetic code is one in a million. *J. Mol. Evol.* **47**, 238–248.

Freeland, S. J., Knight, R. D., Landweber, L. F. & Hurst, L. D. 2000 Early fixation of an optimal genetic code. *Mol. Biol. Evol.* **17**, 511–518.

Freese, E. 1959 The difference between spontaneous and base-analogue induced mutations of phage T4. *Proc. Natl Acad. Sci. USA* **45**, 622–633.

Friedman, S. M. & Weinstein, I. B. 1964 Lack of fidelity in the translation of synthetic polyribonucleotides. *Proc. Natl Acad. Sci. USA* **52**, 988–996.

Goldberg, A. L. & Wittes, R. E. 1966 Genetic code, aspects of organization. *Science* **153**, 420–424.

Goldman, N. 1993 Further results on error minimization in the genetic code. *J. Mol. Evol.* **37**, 662–664.
Haig, D. & Hurst, L. D. 1991 A quantitative measure of error minimization in the genetic code. *J. Mol. Evol.* **33**, 412–417.
Ibba, M. & Söll, D. 1999 Quality control mechanisms during translation. *Science* **286**, 1893–1897.
Joshi, N. V., Korde, V. V. & Sitaramam, V. 1993 Logic of the genetic code, conservation of long-range interactions among amino acids as a prime factor. *J. Genet.* **72**, 47–58.
Judson, O. P. & 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.
Juncgk, J. R. 1978 The genetic code as a periodic table. *J. Mol. Evol.* **11**, 211–224.
Knight, R. D. & Landweber, L. F. 2000 Guilt by association, the arginine case revisited. *RNA* **6**, 499–510.
Knight, R. D., Freeland, S. J. & 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. & Landweber, L. F. 2001 Rewiring the keyboard, evolvability of the genetic code. *Nat. Rev. Genet.* **2**, 49–58.
Kuge, S., Kawamura, N. & Nomoto, A. 1989 Strong inclination toward transitions in nucleotide substitutions by poliovirus replicase. *J. Mol. Biol.* **207**, 175–182.
Kurland, C. G., Hughes, D. & Ehrenberg, M. 1996 Limitations of translational accuracy. In Escherichia coli *and* Salmonella typhimurium*: cellular and molecular biology*, vol. 2 (ed. F. Neidhardt & J. Ingram), pp. 979–1004. Washington, DC: ASM Press.
Lacey, J. C. & Mullins, D. W. 1983 Experimental studies related to the origin of the genetic code and the process of protein synthesis—a review. *Orig. Life* **13**, 3–42.
LaRiviere, F. J., Wolfson, A. D. & Uhlenbeck, O. C. 2001 Uniform binding of aminoacyl-tRNAs to elongation factor Tu by thermodynamic compensation. *Science* **294**, 165–168.
Negre, D., Cenatiempo, Y. & Cozzone, A. J. 1988 Differential pattern of misreading induced by streptomycin *in vitro*. *J. Mol. Biol.* **204**, 213–216.
Osawa, S., Jukes, T. H., Watanabe, K. & Muto, A. 1992 Recent evidence for evolution of the genetic code. *Microbiol. Rev.* **56**, 229–264.
Parker, J. 1989 Errors and alternatives in reading the universal genetic code. *Microbiol. Rev.* **53**, 273–298.
Pelc, S. R. & Welton, M. G. E. 1966 Stereochemical relationship between coding triplets and amino acids. *Nature* **209**, 868–872.
Press, W. H., Flannery, B. P., Teukolsky, S. A. & Vetterling, W. T. 1988 *Numerical recipes in C: the art of scientific computing*. Cambridge University Press.
Schultz, D. W. & Yarus, M. 1994 Transfer-RNA mutation and the malleability of the genetic-code. *J. Mol. Biol.* **235**, 1377–1380.
Sella, G. & Ardell, D. H. 2002 The impact of message mutation on the fitness of a genetic code. *J. Mol. Evol.* **54**, 638–651.
Shimizu, M. 1982 Molecular basis for the genetic code. *J. Mol. Evol.* **18**, 297–303.
Sitaramam, V. 1989 Genetic code preferentially conserves long-range interactions among the amino acids. *FEBS Lett.* **247**, 46–50.
Sonneborn, T. M. 1965 Degeneracy of the genetic code, extent, nature and genetic implications. In *Evolving genes and proteins* (ed. V. Bryson & H. Vogel), pp. 377–397. New York: Academic.
Swanson, R. 1984 A unifying concept for the amino acid code. *Bull. Math. Biol.* **46**, 187–203.
Szathmary, E. & Zintzaras, E. 1992 A statistical test of hypotheses on the organization and origin of the genetic code. *J. Mol. Evol.* **35**, 185–189.
Taylor, F. J. R. & Coates, D. 1989 The code within the codons. *BioSystems* **22**, 177–187.
Topal, M. D. & Fresco, J. R. 1976 Complementary base pairing and the origin of substitution matricies. *Nature* **263**, 285–293.
Varga, R. S. 1962 *Matrix iterative analysis*. Englewood Cliffs, NJ: Prentice-Hall.
Wakeley, J. 1996 The excess of transitions among nucleotide substitutions, new methods of estimating transition bias underscore its significance. *Trends Ecol. Evol.* **11**, 158–163.
Wang, L., Brock, A., Herberich, B. & Schultz, P. G. 2001 Expanding the genetic code of *Escherichia coli*. *Science* **292**, 498–500.
Woese, C. R. 1965*a* On the evolution of the genetic code. *Proc. Natl Acad. Sci. USA* **54**, 1546–1552.
Woese, C. R. 1965*b* Order in the genetic code. *Proc. Natl Acad. Sci. USA* **54**, 71–75.
Woese, C. R. 1967 *The genetic code: the molecular basis for genetic expression*. New York: Harper & Row.
Woese, C. R. 1973 Evolution of the genetic code. *Naturwissenschaften* **60**, 447–459.
Woese, C. R., Dugre, D. H., Dugre, S. A., Kondo, M. & Saxinger, W. C. 1966 On the fundamental nature and evolution of the genetic code. *Cold Spring Harbor 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. 1980 Role of minimization of chemical distances between amino acids in the evolution of the genetic code. *Proc. Natl Acad. Sci. USA* **77**, 1083–1086.
Wong, J. T.-F. 1983 Membership mutation of the genetic code, loss of fitness by tryptophan. *Proc. Natl Acad. Sci. USA* **80**, 6303–6306.
Yarus, M. 2000 RNA-ligand chemistry, a testable source for the genetic code. *RNA* **6**, 475–484.
Zuckerkandl, E. & Pauling, L. 1965 Evolutionary divergence and convergence in proteins. In *Evolving genes and proteins* (ed. V. Bryson & H. Vogel), pp. 97–166. New York: Academic.

## GLOSSARY

NER: normalized encoded range