J Mol Evol (2001) 53:269–281
DOI: 10.1007/s002390010217

JOURNAL OF MOLECULAR EVOLUTION

On the Evolution of Redundancy in Genetic Codes

David H. Ardell,[1,2] Guy Sella[2]

[1] Department of Molecular Evolution, Evolutionary Biology Center, Uppsala University, Norbyvägen 18C, 752 36 Uppsala, Sweden
[2] Department of Biological Sciences, Stanford University, Stanford CA 94305, USA

Received: 21 December 2000 / Accepted: 12 March 2001

Abstract. We simulate a deterministic population genetic model for the coevolution of genetic codes and protein-coding genes. We use very simple assumptions about translation, mutation, and protein fitness to calculate mutation-selection equilibria of codon frequencies and fitness in a large asexual population with a given genetic code. We then compute the fitnesses of altered genetic codes that compete to invade the population by translating its genes with higher fitness. Codes and genes coevolve in a succession of stages, alternating between genetic equilibration and code invasion, from an initial wholly ambiguous coding state to a diversified frozen coding state. Our simulations almost always resulted in partially redundant frozen genetic codes. Also, the range of simulated physicochemical properties among encoded amino acids in frozen codes was always less than maximal. These results did not require the assumption of historical constraints on the number and type of amino acids available to codes nor on the complexity of proteins, stereochemical constraints on the translational apparatus, nor mechanistic constraints on genetic code change. Both the extent and timing of amino-acid diversification in genetic codes were strongly affected by the message mutation rate and strength of missense selection. Our results suggest that various omnipresent phenomena that distribute codons over sites with different selective requirements—such as the persistence of nonsynonymous mutations at equilibrium, the positive selection of the same codon in different types of sites, and translational ambiguity—predispose the evolution of redundancy and of reduced amino acid diversity in genetic codes.

Key words: Evolution — Origin — Code-message coevolution — Redundancy — Amino acid — Wobble — Mutation-selection balance — Codon usage

Introduction

In relation to the problem of the origin and evolution of the standard genetic code, Crick (1968) argued that the greatest increase in fitness should have come from encoding more diverse amino acids. Selection to preserve the meaning of protein-coding genes[1] provided the counter-balance to this advantage. Crick envisioned a primitive genetic code that was highly redundant, producing relatively simple proteins. Codons were then subsequently reassigned to novel amino acids, increasing diversity. This led to increased reliance upon larger and more complex genetic messages for individual fitness, thereby increasing the constraint to preserve message meaning. Presumably, this constraint froze the genetic code before amino-acid diversification could be fully attained and its advantages fully realized.

The message constraint hypothesis has not yet been studied quantitatively. It may be consistent with most

Correspondence to: David H. Ardell; *email:* dave.ardell@ebc.uu.se

[1] As a complement to the word "code," we call the concatenation of all protein-coding genes a "message," not to be confused with messenger RNA.

extant variation in genetic codes, at least in organellar genomes, where reductions in genome size and compositional complexity may cause codons to become infrequent (Osawa et al., 1992). Yet, against the diversity advantage hypothesis, most variant codes reassign codons to already encoded amino acids; this may be partly due to the effect of genomic reductive evolution on the translational apparatus (Andersson and Kurland 1995). Therefore, it behooves us to demonstrate that the vocabularies of extant genetic codes are limited in both number and quality. Furthermore, on the assumption of this limitation, we wish to discern the extent to which it may be explained by the hypothesis of message constraint. Because there are several different, not necessarily mutually exclusive, hypotheses for the origin of genetic codes, it is useful to examine their various necessities and sufficiencies to explain a putative restriction on the genetic code vocabulary, either in concert with Crick's message constraint hypothesis and each other or alone.

Extant genetic codes show natural and experimental evidence of at least three kinds of restrictions to their vocabularies: strict codon synonymy, restricted diversity in chemical property kinds (such as chemical reactivity under various conditions), and redundancy or near-redundancy in the values of chemical properties that are represented among encoded amino acids (for example, the extent of encoded hydrophobicity).

Even accounting for wobble coding, there is a surplus of strict redundancy in the standard genetic code. Wobble rules are taxon-dependent, but one invariant rule is that the third-position pyrimidines (C and U) are not read independently (Osawa et al. 1992). Therefore, allowing for stop codons, that leaves $45 = 64 - 16 - 3$ as an estimate of the maximum encodable number of amino acids in the standard genetic code. The strict redundancy of the standard genetic code, then, may be quantified as $1 - \frac{20}{45} \approx 0.56$ (on a scale from 0 to 1).

Furthermore, it is arguable from natural evidence that the 20 canonical amino acids do not span all dimensions of chemical variety that would be potentially advantageous in modern proteins. For instance, an alternative genetic code encodes a 21st amino acid, selenocysteine (Chambers et al. 1986), and 150 or so different known post-translational covalent modifications to amino acids occur in proteins (reviewed in Wold 1981).

There is also evidence of redundancy in the values of physicochemical properties that do vary among encoded amino acids, such as size and polarity. Amino acids encoded by codons starting with U or C—especially U—in the standard genetic code have very small differences in polarity as measured by, for example, Woese's (1966) Polar Requirement (Ardell, 1998). These codons encode most of the aliphatic amino acids, which substitute for each other more frequently than any other amino-acid pairs (Benner et al. 1994, and references op. cit.). Modern translation also admits the highest rates of translational error among codons encoding these amino acids (Davies et al. 1966; Parker 1989). These data point to a high physicochemical redundancy among the encoded aliphatic amino acids.

The amino-acid vocabulary could well have been shaped by forces independent of selection for increased diversity in, and conservation of, message meaning. That is to say, it is difficult to know whether more extreme or different kinds of physicochemical properties among amino acids were available or encodable during various stages in the evolution of genetic codes. For example, specific stereochemical affinities between certain amino acids and certain components of the translational apparatus, or of messages, could have predetermined amino-acid vocabulary to some extent. Stereochemical predetermination was proposed by Jukes (1973), who argued that ornithine was once encoded but subsequently replaced by arginine, its metabolic product, which has hypothetically greater stereochemical affinity with the translational apparatus. In support of this hypothesis, Knight and Landweber (2000) have shown convincing statistical evidence for an affinity between arginine and its codons in in vitro-evolved aptamers selected for specific amino-acid binding. A more general role for this hypothesis is supported by evidence for affinities of isoleucine and tyrosine with their codons (Yarus 2000).

However, there are problems with the stereochemical theory as the sole explanation to limits to genetic code vocabulary. The statistical evidence for specific stereochemical affinities between codons and amino acids is limited to the amino acids examined experimentally, namely, the 20 canonical amino acids. This sheds no light on possible interactions of aptamers with other amino acids. Also, because these experiments expressly select for the binding of specific amino acids, they are inconclusive as to whether the evolution of a translational system inevitably included certain amino acids and excluded others. Indeed, Wong (1983) showed that *E. coli* could be selected to completely replace tryptophan by 4-fluoro-tryptophan in such a way that cells grew slower in tryptophan-supplemented medium than with 4-F-tryptophan. This change did not come about through an intrinsically higher affinity of the translational apparatus with 4-F-tryptophan. Thus, stereochemical interactions alone cannot comprehensively explain limits to encoded amino-acid diversity.

There may also have been historical or mechanistic constraints on the diversification of amino acids in genetic codes. As an example of an historical constraint, certain amino acids were likely to be metabolically and environmentally unavailable to primitive cells. Wong (1975) and others proposed that codons were donated from metabolic precursors to metabolic products as genetic codes coevolved with metabolism (Taylor and Coates 1989, DiGiulio and Medugno 1999). Like Crick's

verbal model, this hypothesis presupposes an initially more redundant state of the genetic code, in this case encoding metabolically fundamental amino acids that also happen to be abundant in chemical models of the prebiotic earth (Miller 1987). As in Crick's hypothesis, the addition to the code of novel amino acids produced by metabolism could have been partially prevented by a message constraint.

The proximal mechanisms that generate potential variation in genetic codes could also have intrinsically restricted the diversification of encoded amino acids. A proximal mechanism for code change that has been proposed is the duplication and divergence of tRNAs (Fitch and Upper 1987; Schultz and Yarus 1994; DiGiulio 1995). Restriction in vocabulary from this mechanism could have come from specific transformation of an amino acid to its metabolic product after acylation to tRNA, as has been naturally observed in organelles by Schön et al. (1988). However, this particular case is almost certainly a derived rather than an ancestral condition. If code changes did not generally occur through the metabolic transformation of amino acids after acylation to tRNA, then duplication and divergence of tRNAs (and aminoacyl-tRNA synthetases) need not favor the donation of codons to metabolic relatives. Indeed, experimental evidence on misacylation supports that charging errors occur between physicochemically related amino acids (Fersht 1986). If misacylation is a reasonable model for a proximal mechanism of code change through duplication and divergence, then this evidence would seem to favor the encoding of novel amino acids that are stereochemically or physicochemically related, rather than metabolically related, to an ancestral ligand.

It remains to be seen whether historical and mechanistic constraints can comprehensively explain the limited amino-acid vocabularies of genetic codes. For example, analyses incorporating the metabolic coevolutionary hypothesis tend to examine only canonically encoded amino acids (Amirnovin 1997; DiGiulio and Medugno 1999; Freeland et al. 2000). Such studies cannot explain why some amino acids within a metabolic pathway are included and others not.

Weber and Miller (1981) take a different approach to explaining the exclusion of non-canonical amino acids from the standard code. They use biochemical reasoning to argue that certain classes of amino acids were excluded from the code through selection against adverse effects that they caused on protein synthesis and protein structural and catalytic chemistry. Their post hoc arguments are testable and based on considerable biochemical knowledge and experience. However, it is impossible to comprehensively explain the exclusion of amino acids that do not violate the rules they enunciate, but that might have had a positive diversity advantage.

Using an adaptationist approach to explain the twenty canonical amino acids merits caution. King and Jukes (1969) made this point for another purpose in criticizing the following passage by E. L. Smith: "One of the objectives of protein chemistry is to have a full and comprehensive understanding of all the possible roles that the 20 amino acids can play in function and conformation. Each of these amino acids must have a unique survival value in the phenotype of the organism . . ." The hypothesis that the 20 canonical amino acids form a unique and irreducible basis of life is contradicted by the aforementioned experiment by Wong (1983).

On the basis of the present study, we describe the following novel, subtle aspect of the message constraint on code evolution that may have promoted redundancy in genetic codes. Various phenomena cause the same codon to be found simultaneously in different, possibly dissimilar, "types" of sites at once in the same genome. Different types of sites in this sense correspond to distinct sets of locations in proteins with different biochemical requirements, among which amino acids have different relative fitnesses. The distribution of the same codon over different types of sites induces spatial heterogeneity of selection on its meaning. The message constraint then favors the assignment of amino acids to the code with relatively generalized biochemical properties, as a sort of functional compromise to the various different types of sites in which codons occur. We show that this promotes redundancy and restricted diversification of amino acids in genetic codes, without need of additional stereochemical, historical, or mechanistic constraints. Nor is it necessary to postulate that messages became larger and more complex in order to increase the message constraint, as Crick postulated. The phenomena that cause the distribution of codons in multiple types of sites include the persistence of nonsynonymous mutations at mutation-selection equilibrium, and the positive selection of codons in multiple types of sites. Thus, the message mutation rate and selective tolerance to missense influence how specialized and diverse the vocabulary of a genetic code may evolve to become.

Methods

The model described in the appendix has been implemented in the program CMC, written in C++, and available upon request from the authors. The appendix also defines terms we use below that may be unfamiliar or are used unconventionally (such as "codon usage"). Eigensystem solutions for determining the growth rate $\lambda(c)$ and equilibrium codon usage $U(c)$, associated with a genetic code c according to the model, were obtained using the iterative method (Press et al. 1988). All values were calculated to double precision (10^{-16}). Ties in the genetic code take-over condition were broken arbitrarily by picking the first code observed with maximal invasion fitness.

A simulation of code-message coevolution according to our model is fully determined by picking values for the message mutation rate parameter μ, the missense tolerance parameter ϕ, and a set A of 20 uniform randomly distributed values between 0 and 1. These values represent both the physicochemical requirements of the 20 site-types in

proteins and the physicochemical properties of the 20 encodable amino acids. In the following, the set A of 20 randomly distributed values is called an "amino-acid/site-type space" or simply "amino-acid space." For simplicity, the values themselves are called "amino acids."

We examined the effects of variation in the parameters on the number of explicitly encoded amino acids (N_{aa}), the number of explicit encoding codons (N_c), and a measure of physicochemical diversity called *NER*, to be defined below, both in frozen genetic codes and over simulation time. We examined parameter values of ϕ ranging from $\phi = 2.8 \times 10^{-7}$ to $\phi = 0.9999$ and values of μ ranging from $\mu = 5.0 \times 10^{-6}$ to $\mu = 1.0 \times 10^{-3}$. In order to control for any idiosyncratic effects of particular randomly generated amino-acid/site-type spaces, we examined the mean and standard deviation (and sometimes the median and interquartile range) of these observables over simulations run with 40 different uniformly distributed amino-acid spaces.

The measure of physicochemical diversity of amino acids in a code that we used was the range of physicochemical properties that it explicitly encoded divided by the maximum range for the amino acid space with which it evolved. Denote by $c(C_{II}^B) \subseteq A$ the set of amino acids explicitly encoded by a genetic code c on the codon set C_{II}^B, and by $d(\beta|\alpha) = |\beta - \alpha|$ the physicochemical distance between any two amino acids $\alpha, \beta \in A$.

The *Normalized Encoded Range (NER)* of a code c given an amino-acid space A and associated physicochemical distance d is:

$$NER = \frac{\max_{\alpha,\beta \in c(C_{II}^B)} d(\beta|\alpha)}{\max_{\alpha,\beta \in A} d(\beta|\alpha)} \qquad (1)$$

Results

Redundancy in Codes Increases Directly with Both Mutation Rate and the Tolerance of Missense in Messages

Figure 1 shows a simulated evolutionary trajectory of a 4-base, 2-position genetic code from an initial uniformly ambiguous state to a diversified and explicit frozen state. This trajectory was typical for these parameters of μ and ϕ. A complete depiction of the code-message coevolutionary trajectory would include graphs of the equilibrated usage of all 16 codons within each of the 20 site-classes at each step. However, for simplicity, the codon usage patterns are not shown.

In this example, the initial ambiguous state persisted until step 23; 7 of 9 reassignments occurred before the code evolved to become fully explicit. The code froze at step 25 with only 10 different encoded amino acids for a final redundancy of $1 - \frac{10}{16} = 0.375$. The amino acids that were encoded did not include the most physicochemically extreme amino acids, labeled as 1 and 20. Instead, most of the encoded amino acids, and all of the redundantly encoded amino acids, came from the middle of amino-acid space, despite that the target protein selected for the encoding of all 20 amino acids.

The genetic code in Fig. 1 did not freeze with any codons in the initially ambiguous state. In fact, in none of our simulations did any genetic code freeze that was not fully explicit (despite that the ambiguous state encodes

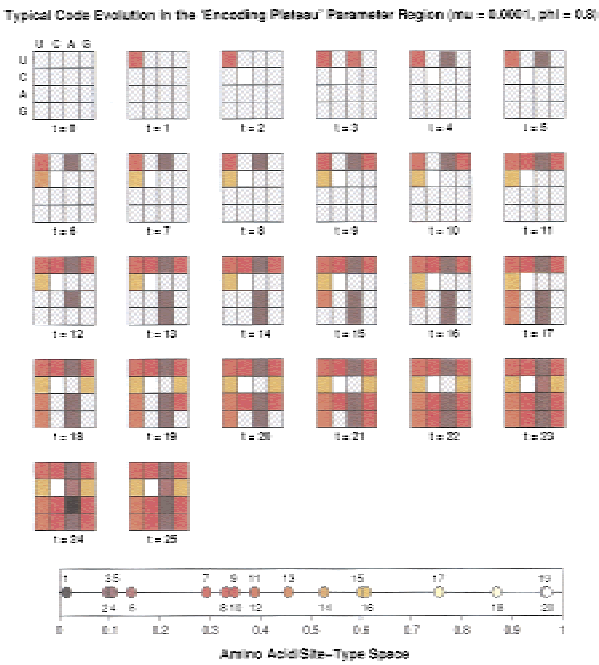

Fig. 1. A typical code evolution in the Encoding Plateau region of parameter space. Four-base, 2-position genetic codes are represented as 4-by-4 grids. The initial uniformly ambiguous coding state is indicated by the striped pattern. The one-dimensional scale at the bottom of the graph is the amino-acid/site-type space with which this code evolved. Codes are shown above the step number in which they were uniquely established in the population. The code of a subsequent step was the unique invading mutant that most increased the fitness of messages equilibrated to its predecessor. No mutant codes could invade the messages of the code frozen at step 25. The colors of codons indicate the physicochemistry of their assigned amino acids as shown in the scale.

the optimal amino acid of any site with some probability). Instead, all codons eventually became explicitly assigned to some amino acid during all runs.

In Fig. 2 we show that this result of redundancy and limited amino-acid diversity was typical of frozen codes evolved with any of 40 different random amino-acid/site-type spaces over a broad range of message mutation rates and missense tolerances. Indeed, over the entire parameter-space that we examined, averages of both N_{aa} (top of Fig. 2) and *NER* (bottom) were less than their theoretical maximum values of 16 and 1.0, respectively. Average N_{aa} remained between 10 and 13 (redundancy remained between 0.1875 and 0.375), and average *NER* stayed at about 0.9, for $10^{-4} \leq \mu \leq 10^{-3}$ and $0.1 \leq \phi \leq 0.95$, a stability in the parameter space that we call the *Encoding Plateau*.

We extended our studies to extremes of strong selection and low mutation rates to see if we could force the average behavior of the genetic codes we evolved to encode the maximum of 16 amino acids and the maximum encodable physicochemical range. These data are shown in Fig. 3. Even at the strongest selection we examined, ($\phi = 2.8 \times 10^{-7}$, $\mu = 0.0001$, $N = 40$), both the average and median number (not shown) of encoded

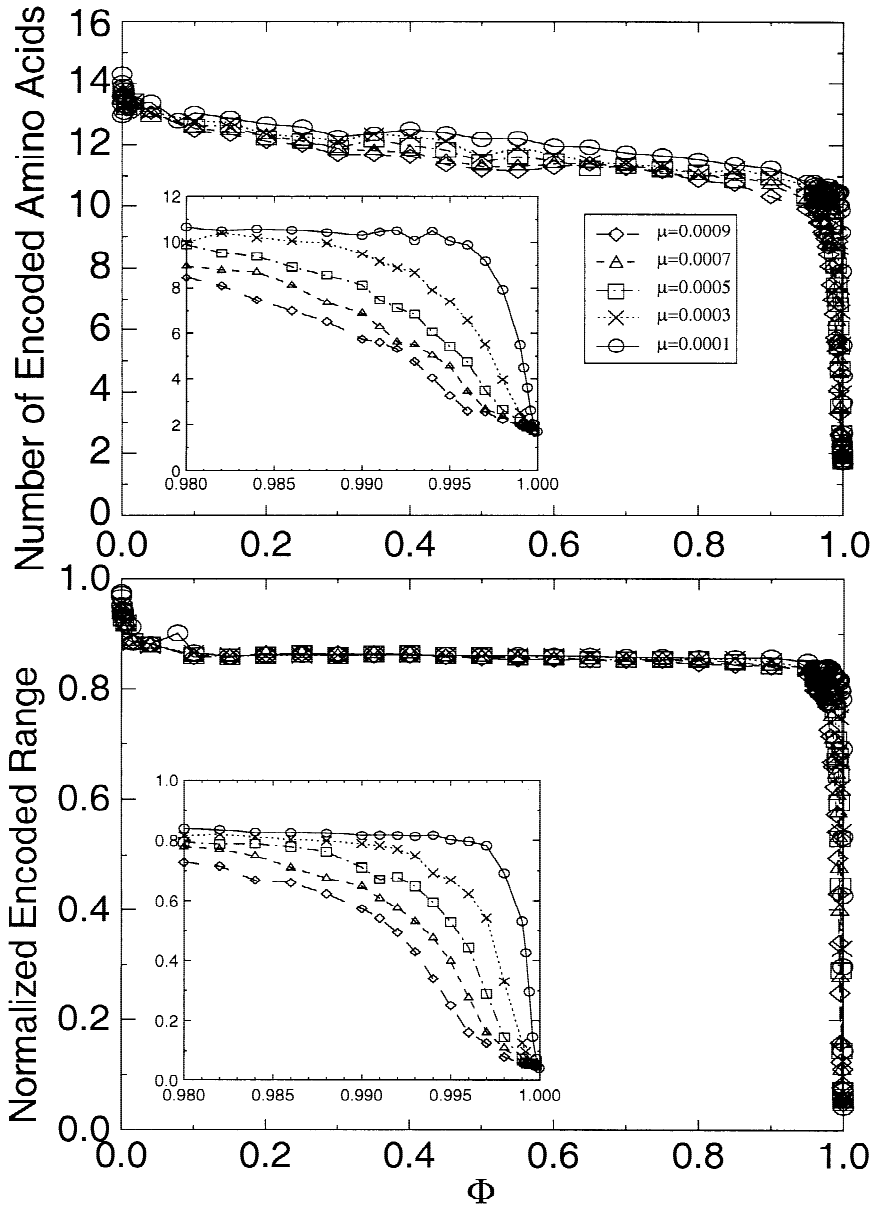

Fig. 2. The average number (N_{aa}) and Normalized Encoded Range (*NER*, see Methods) of amino acids in final frozen codes as functions of message mutation rate (μ) and missense tolerance by selection (ϕ). Each point represents an average over 40 runs with different amino-acid/site-type spaces. A smaller value of ϕ corresponds to stronger missense selection (see Appendix).

amino acids was less than 15. That is to say, the majority of frozen codes evolved encoded less than 15 amino acids under these conditions. At lower mutation rates, the rate of convergence of our calculated eigensystems was too slow to statistically average over many different runs with different site-type/amino acid spaces. However, both of the two genetic codes evolved under the most extreme conditions that we examined ($\phi = 0.0001$, $\mu = 5.0 \times 10^{-6}$, $N = 2$) remained partially redundant and did not encode the most physicochemically extreme amino acids. Both frozen genetic codes encoded 15 different amino acids, and had an *NER* of approximately 0.97. In the majority of simulations under almost all parameter conditions that we examined, some redundancy remained in our frozen genetic codes. The *NER* was below its maximum of 1.0 in all frozen codes that we evolved.

Changing the mutation and selection parameters caused profound differences in the level of redundancy. Frozen genetic codes encoded both more and increasingly diverse amino acids, on average, when they coevolved with messages mutating at lower rates or stronger missense selection (Figs. 2 and 3). With lower ϕ or μ, the reduction of redundancy in genetic codes tended to occur through the encoding of amino acids from the middle of amino-acid space rather than its extremes. This is evident, for instance, within the Encoding Plateau (Fig. 2), where the rate of increase in N_{aa} with the strength of missense selection is greater than that for *NER*.

Decreasing selection above threshold values caused sharp reductions in both N_{aa} and *NER* on average (Fig. 2). The value of the threshold depended on the mutation

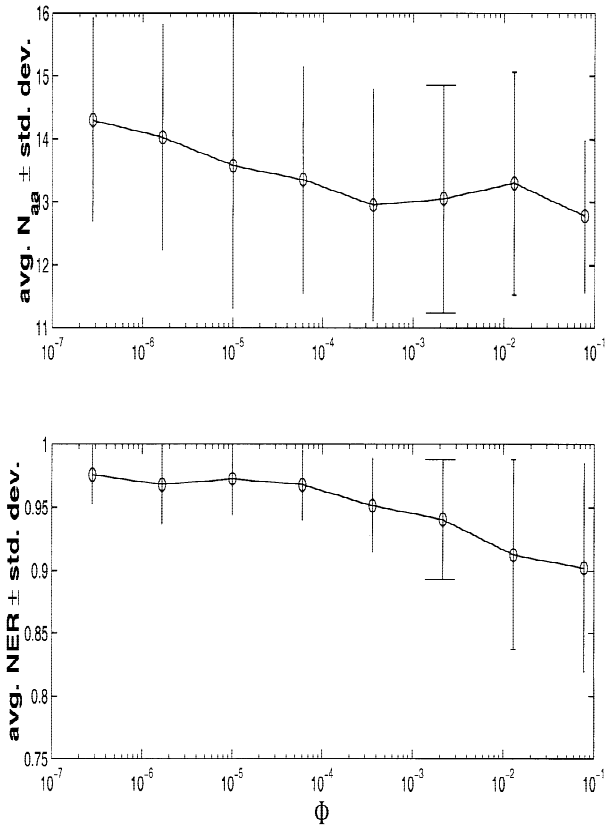

Fig. 3. Average and standard deviation (error bars) in N_{aa} and *NER* of amino acids in final frozen codes evolved under conditions of extremely low missense tolerance and $\mu = 0.0001$. As before, $N = 40$ amino-acid/site-type spaces for each point.

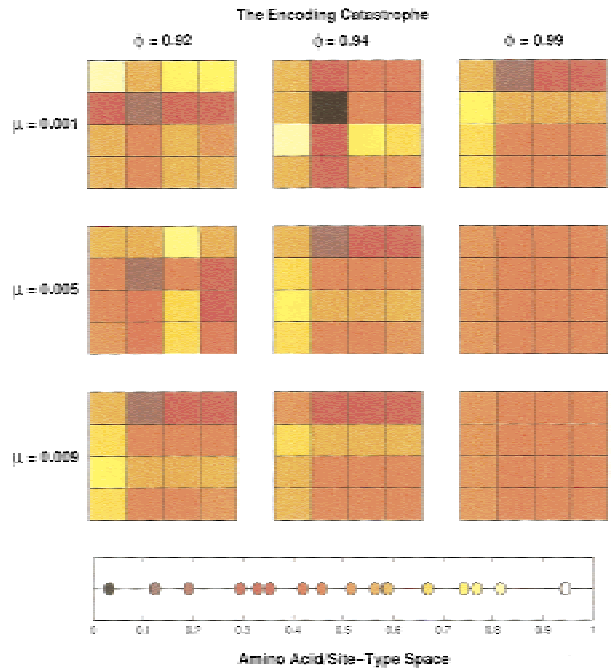

Fig. 4. An array of final frozen codes placed in a grid according to the selection (column) and mutation rate (row) parameters with which they evolved. All codes were evolved with the same amino-acid/site-type space shown.

rate and much less on the form of the amino-acid/site-type space. Under conditions of high mutation rate and extremely weak missense selection, near-total redundancy evolved, a phenomenon we call the *Encoding Catastrophe*.

Figure 4 shows the effect of the Encoding Catastrophe on the morphology of final codes. All 9 codes shown were evolved in simulations using the same single amino-acid/site-type space and different combinations of μ and ϕ parameters as indicated. It is clear that when diversity decreased and redundancy increased in the Encoding Catastrophe, they did so towards the center of amino-acid space. It may be shown that amino acids from the center of amino-acid space are preferably encoded in the various circumstances we describe because they have the highest geometric mean fitness over site-types in our model. Therefore, we call such amino acids "versatile" or "generalist" amino acids, compared to those at the extremes of amino-acid space, which are relatively "specialized" to site-types at the same extremes of site-type space.

Early Versus Late Diversification of Amino Acids in Genetic Codes

Mutation and selection affected not only the extent but also the timing and evolutionary path of amino-acid diversification. In our model, this diversification generally occurred through codon reassignments except under conditions of very low tolerance to missense. Small changes in parameter values could radically change the timing and nature of codon reassignments that occurred in code evolution, and thus, the timing of amino-acid diversification.

Recall that codon reassignments occurred rather frequently in the Encoding Plateau region of parameter space shown in Fig. 1. Seven reassignments occurred before every codon evolved explicit meaning (*reassignments before explicit*), and 2 reassignments occurred after, for a total of 9 in all. A close look shows that many of these reassignments tended to be diversifying. For example, in step 4, a codon that had just been assigned amino acid 4, at the low extreme of amino-acid space, became reassigned in the next step to the even more extreme amino acid 2. Another diversification through reassignment occurred just before freezing, at step 24.

When we looked at the average behavior over 40 different amino-acid/site-type spaces, between 8 and 11 reassignments of codon meaning occurred in genetic codes coevolving in the Encoding Plateau (depending on the mutation rate), mostly occurring before the code was explicit (Fig. 5). In the Encoding Catastrophe (as ϕ or μ were increased), codes rapidly froze without any reassignments at all. Reassignments also decreased under conditions of low missense tolerance relative to the Encoding Plateau, suggesting as expected that selection against missense inhibited changes in codon meaning.

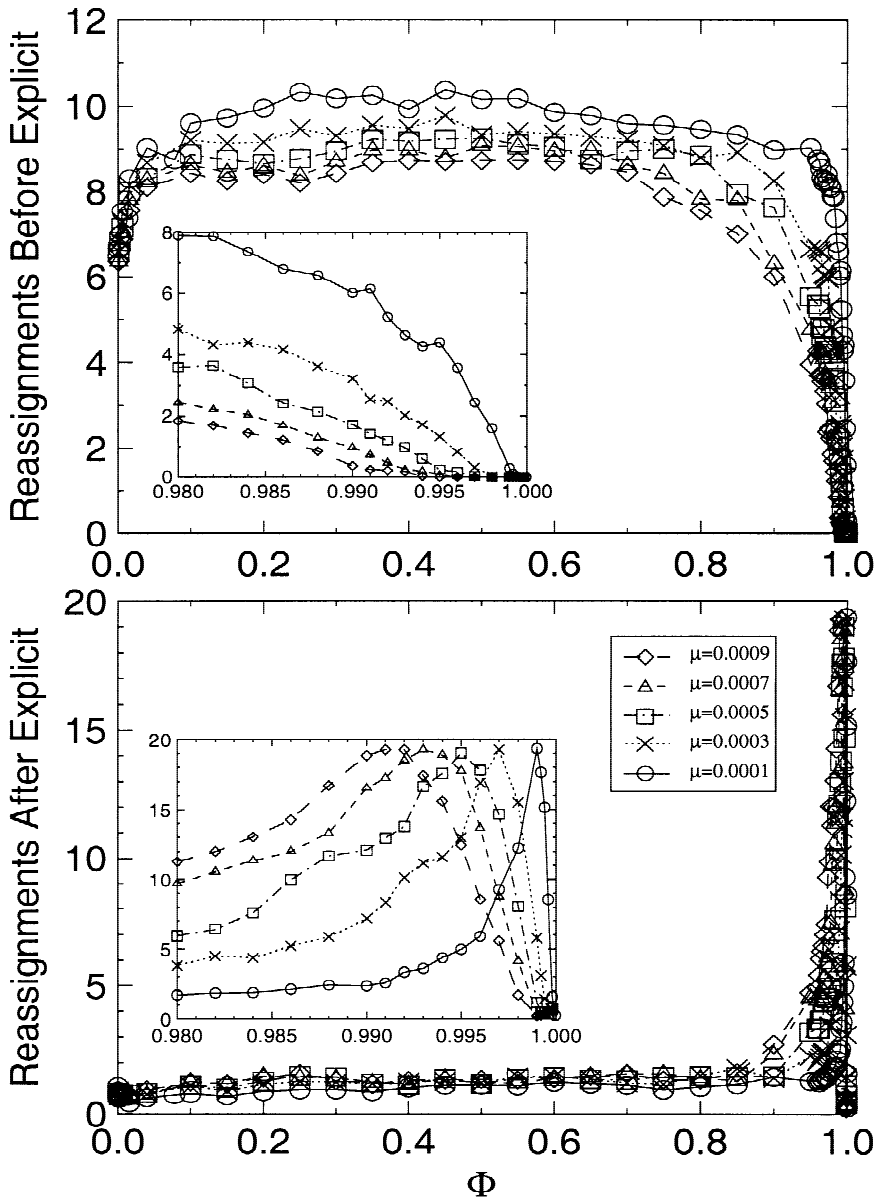

Fig. 5. The effects of message mutation and selection on the average numbers codon reassignments before and after the step in which all codons were assigned explicit amino-acid meaning. $N = 40$ runs with different amino-acid/site-type spaces.

We call this part of parameter space the *Strong Selection* region.

However, in a region of parameter space between the Encoding Plateau and the Encoding Catastrophe, there was a major transition in the timing of codon reassignments, best seen in the inset figures of Fig. 5. In this transitional region, genetic codes first evolved to be explicit without reassignments, and then went through a prolonged epoch of up to almost 20 reassignments before finally freezing. For reasons explained in the discussion, we call this region of parameter space the *Sonneborn Region*.

The transition in the timing of reassignments corresponded exactly in parameter space to the decrease in average encoded amino-acid diversity which culminated in the Encoding Catastrophe (Figs. 2 and 5). A typical evolutionary trajectory of a genetic code coevolving in the Sonneborn Region is shown in Fig. 6. Compared to the genetic code in Fig. 1, this genetic code evolved to become explicit much earlier (step 19 versus step 23), with many more reassignments of meaning occurring after explicit (18 versus 2). The initial explicit state was more redundant than that which evolved in the Encoding Plateau. The amino acids that it encoded tended to be from the middle of amino-acid space. Like in the Encoding Plateau, reassignments tended to diversify the code, but the frozen code in step 37 was more redundant and less diverse than the code that evolved in the Encoding Plateau (Fig. 1).

These differences in dynamical behavior are shown to be general in Fig. 7, which illustrates how the different regions of parameter space affected the median time-course number of explicit codons (N_c), number (N_{aa}), and Normalized Encoded Range (*NER*) of encoded amino acids, and the cumulative distribution over time of codes frozen (out of 40 runs for each value of ϕ shown).

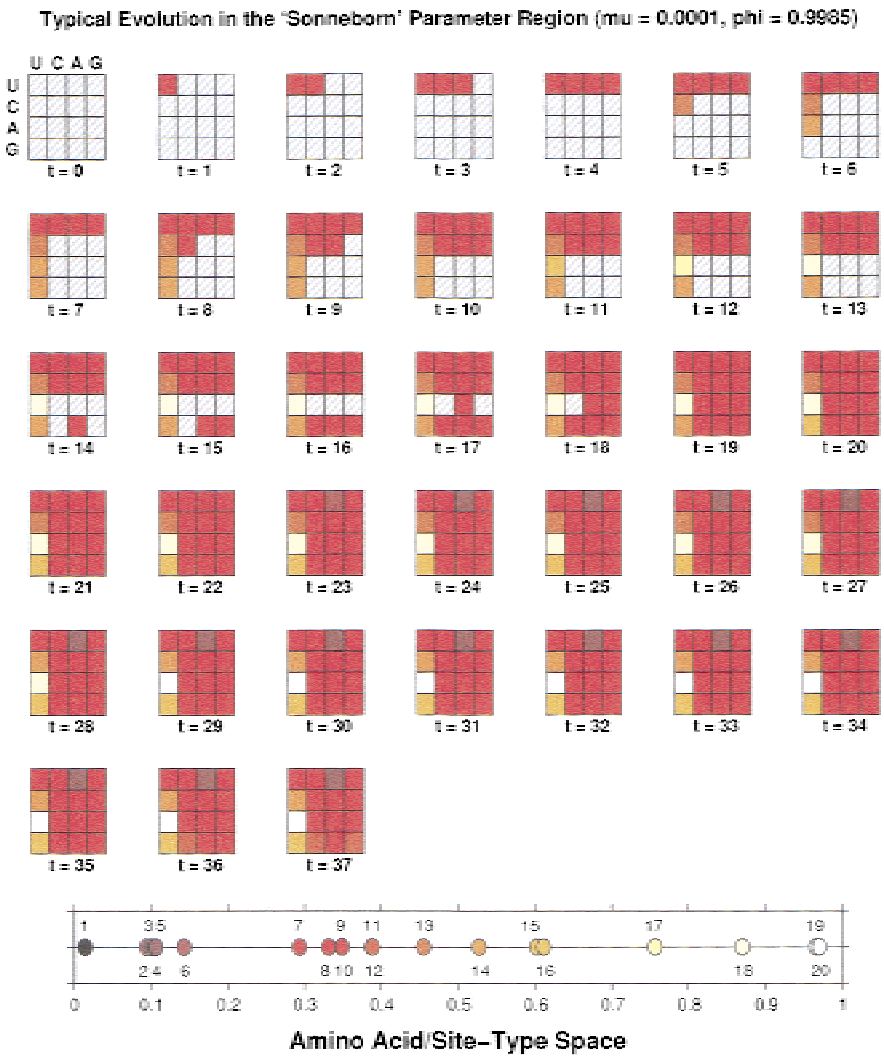

Fig. 6. A typical code evolutionary trajectory in the Transitional Region of parameter space between the Encoding Plateau and the Encoding Catastrophe that we call the Sonneborn Region for reasons given in the discussion. The amino-acid/site-type space shown is the same as in Fig. 1.

As in earlier figures, we controlled for idiosyncratic effects of individual amino-acid/site-type spaces by studying the central tendencies and dispersion of the results over multiple runs with different spaces. The interquartile range (itq) of N_c, N_{aa}, and *NER* are sporadically indicated with error bars. The five different values of ϕ illustrated survey the regions of parameter space we have so far described. From Fig. 5, the dynamical behavior of runs with $\mu = 0.0005$ is dominated by Strong Selection at $\phi = 0.001$, is in the Encoding Plateau at $\phi = 0.4$ and 0.8, is in the peak of the Sonneborn Region at $\phi = 0.995$, and is approaching the Encoding Catastrophe at $\phi = 0.999$.

Genetic codes took almost twice as many steps to freeze in the Sonneborn Region as in other parts of parameters space. On the other hand, codes froze fastest (with the least reassignments) at the extremes of parameter space, both in the Strong Selection and Encoding Catastrophe regions.

Even though N_c increased fastest in the Encoding Catastrophe and Sonneborn Region, amino-acid diversity evolved much later, if at all, compared to the Encoding Plateau or Strong Selection regions of parameter space. There, the number of encoded amino acids increased steadily throughout evolution, and codes froze soon after they became explicit. *NER*, on the other hand, increased to its maximum almost immediately. This implies that amino acids which were encoded later in the Encoding Plateau and Strong Selection regions tended to be in the middle of amino-acid space rather than at the extremes.

Because codons could not revert to the ambiguous state and we know that *NER* did not decrease over time in our runs, it might be thought that the running median curves in Fig. 7 should all increase monotonically. They don't because as more and more codes froze under the same conditions (bottom subfigure of Fig. 7) fewer and fewer runs contributed to the median. Evidently, runs which persisted longer in the Sonneborn Region had smaller N_{aa} and *NER*.

The overall similarity of the sets of runs with $\phi = 0.4$ and $\phi = 0.8$ in Fig. 7 illustrates the dynamic equivalence of runs within the Encoding Plateau.

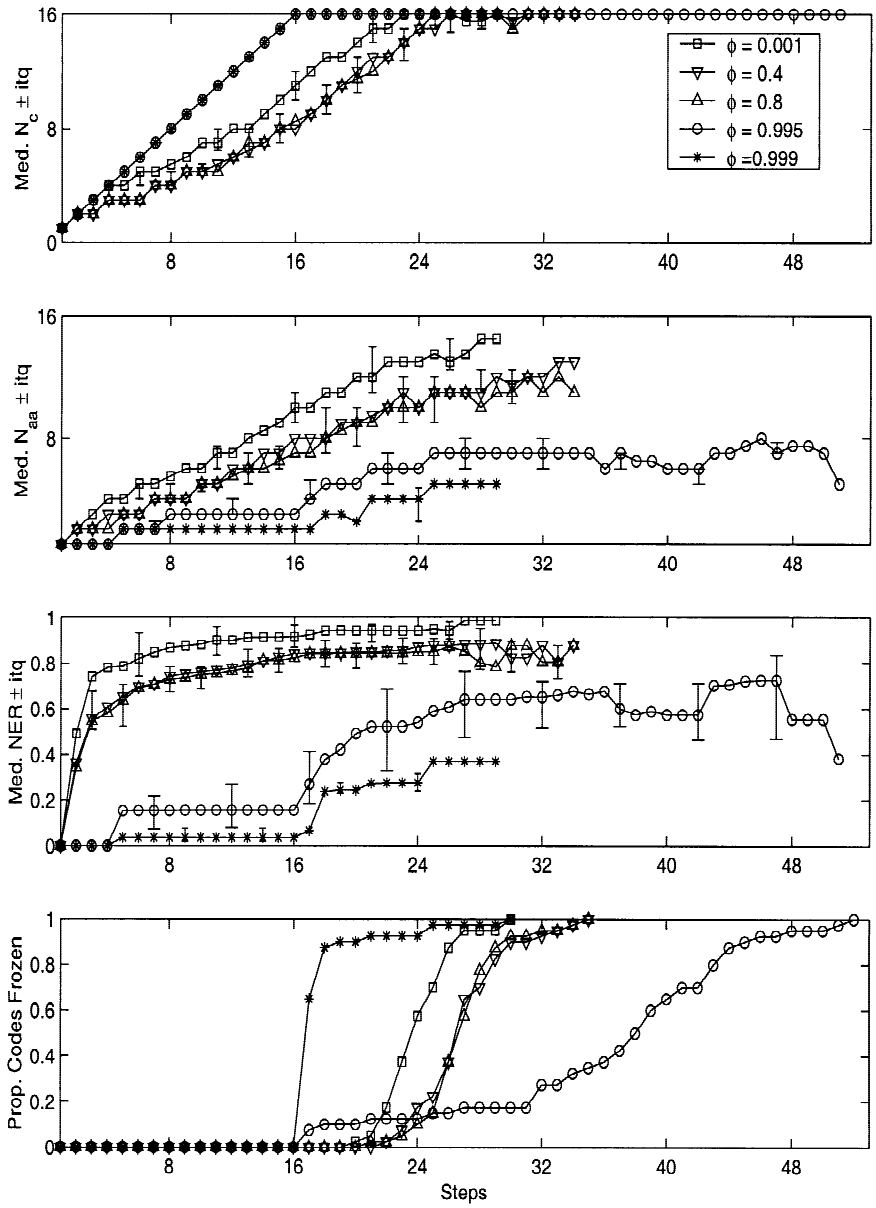

Fig. 7. Running median over simulation time of the number of explicit codons (N_c) and the number (N_{aa}) and Normalized Encoded Range (*NER*) of explicitly encoded amino acids. Each point shown is the median these observables in runs at a given step over 40 simulations with different amino-acid/site-type spaces and message mutation rate $\mu = 0.0006$. Interquartile ranges (error bars) are shown only sporadically for clarity. The bottom panel shows the proportion of codes frozen as a function of step-time.

Discussion

It is widely understood that the usage patterns of codons, in a general sense, depend on interactions among evolutionary forces such as mutation, selection on nonsynonymous and synonymous changes, genetic drift, and upon the genetic code itself. For instance, King and Jukes (1969) demonstrated that amino-acid composition in proteins is largely predictable from redundancy patterns in the standard genetic code, mutation, and genetic drift. Yet few studies have treated how these factors will influence the constraint of messages on genetic code evolution. An exception is a theory of derived genetic code change due to Osawa et al. (1992), who consider the effects of neutral changes in equilibrium base composition on codes. By assuming that codons must become very rare to be reassigned, they assumed that selection against multiple simultaneous amino-acid replacements will be very strong.

We have introduced a quantitative model for the evolutionary interactions of genetic codes and messages. Analysis of codon usage patterns in our model, not presented here, illuminated several factors that influenced how codons were distributed over sites with different selective requirements. These factors were: the equilibrium usage of nonsynonymous mutant codons, positive selection of the same codon in multiple site-types, translational ambiguity, and the genetic code itself. In various ways, these factors caused the fitness consequences of a change in meaning of a single codon to depend on the fitnesses of amino acids in different selective environments at the same time.

The magnitude of these various effects depended predictably on the magnitude of evolutionary forces acting on messages. For example, both high mutation rates and high selective tolerance to missense increased the persistence of nonsynonymous mutant codons at mutation-selection equilibrium. This favored the assignment of amino acids with relatively high average fitness in multiple site-types, which were those from the middle of amino-acid space. These amino acids, which may be called versatile or generalist amino acids, were the only ones to be encoded at sufficiently high mutation rates or missense tolerance (Fig. 4).

Analysis of codon usage patterns (not shown) illuminate that, under some conditions, the genetic code itself could wield a strong influence on codon usage patterns in messages and thereby on its own subsequent evolution. This effect was greatest in the Encoding Plateau and Sonneborn Region, where messages evolved at intermediate mutation rates and missense tolerances. If mutation rates or missense tolerances were too high, as in the Encoding Catastrophe, codons were distributed so randomly, with respect to their amino-acid meaning, that changes in the code did not alter codon usage patterns and, in turn, did not influence subsequent evolution of the code. If, on the other hand, mutation rates and missense tolerances were very low, the frequencies of mutant codons were very low. Codons were mainly used where they were positively selected. Low frequencies of mutant codons caused the outcome of code changes to be predominantly determined by a few site-types with similar fitness requirements. In this case, explicit codons could change meaning more independently of the detailed organization of amino-acid assignments in the code. This explains why reassignments were more common in the Encoding Plateau and later stages of evolution in the Sonneborn Region, where they often occurred in consecutive cascades, than in the Encoding Catastrophe or Strong Selection Regions. At intermediate parameter values, changes to the code perturbed codon usage patterns in multiple types of sites, often reducing the number of types of sites in which a codon was used. This released constraints on subsequent code changes and facilitated amino-acid diversification, often through codon reassignments.

Even at extremely low mutation rates and missense tolerance, we were unable to evolve a majority of non-redundant genetic codes. Nonsynonymous mutations should have had negligible effects on the fitness of code changes in this part of parameter space. However, the same codon could be positively selected in more than one site-type on the basis of its fitnesses relative to other codons within a given code. The resulting high usage of the same codon in multiple types of sites would then restrict further specialization of its meaning. This probably partly explains why the most extremely physicochemically diverse amino acids were never encoded in our final frozen genetic codes under any conditions. It also describes another way in which earlier forms of a code may influence its own future evolution though patterns of codon usage.

Under our assumptions, the initial ambiguous state of codons had low fitness under most patterns of codon usage. For example, after a few codons were assigned explicit meaning in the Encoding Plateau or Strong Selection regions, the frequency of ambiguous codons in messages was negligible, so that only a few explicit codons dominated the codon usage in all types of sites. In contrast, in the Encoding Catastrophe and early stages of the Sonneborn Region, ambiguous codons were used with frequencies similar to explicit codons at mutation-selection equilibrium-(data not shown). Thus, if codons were distributed over many site-types, an explicit generalist amino acid coding state was more fit than the uniformly ambiguous initial coding state. The transition in parameter space that we observed then, between early and late amino acid diversification, was actually a transition between two alternative evolutionary routes through which codes and messages reduced the use of relatively unfit ambiguous codons: the purification of them from messages through selection or the rapid assignment of explicit meaning to them through code mutation. This second evolutionary route, ambiguity avoidance through rapid assignment of explicit sense, is a missense analogue to the scenario proposed by Sonneborn (1965), in which most codons in the early code were rapidly assigned some meaning to reduce the frequency of nonsense mutations (termination). It is for this reason we call the transitional region between the Encoding Plateau and Encoding Catastrophe the Sonneborn Region.

In later stages of evolution in the Encoding Plateau, it may be shown that ambiguous codons were used in many types of sites at low frequencies, on the order of the mutation rate. Such patterns of usage favored the (re-) assignment of generalist amino acids. If we had considered restricted ambiguity over a smaller set of perhaps physicochemically similar amino acids, then such ambiguous codons could be positively selected in site-types that favored one or more of the amino acids thereby ambiguously encoded. This might happen only if none of the ambiguously encoded amino acids or highly fit substitutes for them were not explicitly encoded by other codons. In such a case, translational ambiguity would also restrict subsequent amino-acid diversification by favoring the assignment of an explicit meaning with high relative average fitness in the types of sites in which that codon was used. Thus, translational ambiguity, through both the low fitness that it induces, and its possible tendency to distribute codons over sites with different fitness requirements, may ultimately promote a less diverse genetic code.

We have described several ways in which code-message coevolution may lead to a more uniform distri-

bution of codon usage over sites with different selective requirements, and shown that this is an intrinsic factor limiting amino-acid diversification in the evolution of genetic codes. The frozen genetic codes we evolved had higher fitness with their own messages than mutant codes that increased the range of encoded amino acid physicochemical properties. Indeed, at each step in evolution, the fitness of assigning or reassigning exactly one of any of the 20 amino acids to any of the 16 codons was calculated as a potential path of evolution. Every amino acid was available throughout evolution to optimally fulfill the selective requirements of a unique site-type, yet the maximum diversity in vocabulary was not attained. Thus, historical, stereochemical, and mechanistic constraints may be unnecessary to explain redundancy in genetic codes. Further, our data show that the rate of message mutation and the strength of missense selection have probably had profound effects on the extent and timing of diversification of encoded amino acids in genetic code evolution.

It is worth asking whether the notion of a generalist amino acid makes biological sense. Such a claim was made for alanine by Zuckerkandl and Pauling (1965) in their analysis of amino-acid substitution in hemoglobin. Döring and Marlière (1998) make such a claim for cysteine on the basis of its small size, amphiphilicity and that they achieved partial replacement of isoleucine by cysteine in *E. coli* with relatively low loss in fitness. Finally, we note that generalist amino acids should tend to substitute more frequently and uniformly with other amino acids. Physicochemical generality is a property, then, that could be measured in the analysis of amino-acid substitution matrices.

Further work could examine the extent to which the size and dimensionality of codon and amino-acid spaces, the form of fitness interactions of amino acids within and among sites, and the particulars of the code-message coevolutionary dynamics affect the results and hence the generality of our interpretations. The effect of finite population size on messages, also not addressed here, will probably be an additional factor that distributes codons over sites with different selective requirements. We also have not attempted to estimate the conditions of message mutation and selection that reproduce the observed level of redundancy in the standard genetic code.

The interpretations of this paper may possibly be generalized to the evolution of other coding systems. The generalized principle would be that such phenomena as the erroneous transmission of symbols, the multiple use of symbols in different semantic settings, and the ambiguous decoding of symbols may all ultimately restrict later diversification of symbolic meaning.

Acknowledgments. Research supported by NIH grants GM 28016 and GM 28428 to Marcus W. Feldman. We thank Virginia Walbots, Dirk Repsilber, Carolin Frank, and Marcus W. Feldman for critical readings of the manuscript, and Drs. Emile Zuckerkandl, Syozo Osawa, and Takashi Gojobori for the privilege of contributing to this issue in honor of Dr. T. H. Jukes.

References

Amirnovin R (1997) An analysis of the metabolic theory of the origin of the genetic code. J Mol Evol 44:473–476

Anderson S, Kurland C (1995) Genomic evolution drives the evolution of the translation system. Biochem Cell Biol 73:775–787

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

Benner S, Cohen M, Gonnet G (1994) Amino acid substitution during functionally constrained divergent evolution of protein sequences. Protein Eng 7:1323–1332

Chambers L, Frampton J, Goldfarb P, Affara N, McBain W, Harrison P (1986) The structure of the glutathione peroxidase gene: the selenocysteine in the active site is encoded by the "termination" codon, TGA. EMBO J 5:1221–1227

Crick F (1968) The origin of the genetic code. J Mol Biol 38:867–379

Davies J, Jones D, Khorana H (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

DiGiulio 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

DiGiulio M, Medugno M (1999) Physicochemical optimization in genetic code origin as the number of codified amino acids increases. J Mol Evol 49:1–10

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

Fersht A (1986) The charging of tRNA. In: Kirkwood T, Rosenberger R, Galas D (eds). Accuracy of molecular processes. Chapman and Hall, New York, pp 67–82

Fitch W (1966) Evidence suggesting a partial, internal duplication in the ancestral gene for heme-containing globins. J Mol Biol 161

Fitch W, 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, Knight R, Landweber L, Hurst L (2000) Early fixation of an optimal genetic code. Mol Biol Evol 17:511–518

Jukes T (1973) Arginine as an evolutionary intruder into protein synthesis. Biochem Biophys Res Comm 53:709–714

King J, Jukes T (1969) Non-darwinian evolution. Science 164:788–798

Knight R, Landweber L (2000) Guilt by association: the arginine case revisited. RNA 6:499–510

Miller S (1987) What organic compounds could have occurred on the prebiotic earth? Cold Spring Harbor Symp Quant Biol 52:17–27

Osawa S, Jukes T, 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

Press W, Flannery B, Teukolsky S, Vetterling W (1988) Numerical recipes in c: the art of scientific computing. Cambridge University Press, Cambridge

Schön A, Kannangara C, Gough S, Söll D (1988) Protein biosynthesis in organelles requires misaminoacylation of tRNA. Nature 331: 187–190

Schultz D, Yarus M (1994) Transfer-RNA mutation and the malleability of the genetic-code. J Mol Biol 235:1377–1380

Sella G, Ardell D (2001) The impact of message mutation on the fitness of a genetic code. J Mol Evol (in press)

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

Taylor F, Coates D (1989) The code within the codons. BioSystems 22:177–187
Weber A, Miller S (1981) Reasons for the occurrence of the twenty coded amino acids. J Mol Evol 17:273–284
Woese C (1965) On the evolution of the genetic code. Proc Natl Acad Sci USA 54:1546–1552
Woese C, Dugre D, Dugre S, Kondo M, Saxinger W (1966) On the fundamental nature and evolution of the genetic code. Cold Spring Harbor Symp Quant Biol 31:723–736
Wold F (1981) In vivo chemical modification of proteins (post-translational modification). Annu Rev Biochem 50:783–814
Wong J (1975) A co-evolution theory of the genetic code. Proc Natl Acad Sci USA 72:1909–1912
Wong JF (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: Bryson V, Vogel H (eds), Evolving genes and proteins. Academic Press, New York, pp. 97–166

Appendix

Model

This model builds on the assumptions, justifications, and results detailed and discussed in Sella and Ardell (2001). To specify our model, we must define the following quantities: The set of codons and how they mutate; the sets of amino acids, types of sites in 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.

The Codon Set, Codon Mutation Scheme, and Message. We model a set of 16 codons, with 2 positions over the set of 4 bases $B = \{U, C, A, G\}$. The codon set C_{II}^{B} then consists of 16 codons:

$$C_{II}^{B} = B \times B = \{UU, UC, UA, \ldots, GG\}. \tag{2}$$

The codon mutation matrix μ_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 one generation. The corresponding codon mutation is then:

$$\mu_C(zw|xy) = \mu_B(z|x)\mu_B(w|y) \quad x, y, z, w \in B, \quad xy, zw \in C_{II}^{B}. \tag{3}$$

We study a one-parameter model of base mutation, namely:

$$\mu_B = \begin{array}{c} \\ U \\ C \\ A \\ G \end{array} \begin{array}{c} \begin{array}{cccc} U & C & A & G \end{array} \\ \begin{pmatrix} 1-\mu & \frac{\mu}{3} & \frac{\mu}{3} & \frac{\mu}{3} \\ \frac{\mu}{3} & 1-\mu & \frac{\mu}{3} & \frac{\mu}{3} \\ \frac{\mu}{3} & \frac{\mu}{3} & 1-\mu & \frac{\mu}{3} \\ \frac{\mu}{3} & \frac{\mu}{3} & \frac{\mu}{3} & 1-\mu \end{pmatrix} \end{array} \tag{4}$$

where μ is the *message mutation parameter*, the probability of a base being replicated correctly after one generation. The *message* of an individual is a vector of L codons $\vec{m} = \langle m_1, \ldots, m_L\rangle$, $m_i \in C_{II}^{B}$, representing the concatenation of all protein-coding genes.

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 and encodable throughout 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 and fitness in proteins, such as the Polar Requirement measure of Woese et al. (1966). 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 $d(\beta|\alpha) = |\beta - \alpha|$ between them.

A *genetic code* $c(\alpha|i)$ determines the probability of producing amino acid $\alpha \in A$ from codon $i \in C_{II}^{B}$ in a message through translation. Translation is independent by codon, so that the probability $P(\vec{p}|\vec{m})$ of producing any *protein*, which is a vector of L amino acids $\vec{p} = \langle p_1, \ldots, p_L\rangle$, $p_i \in A$, from a given message $\vec{m}$ through translation using code c is

$$P(\vec{p}\,|\,\vec{m}) = \prod_{l=1}^{L} c(p_l|m_l). \tag{5}$$

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. 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 amino acids.

The *target* is the vector of L site-types, $\vec{s} = \langle s_1, \ldots, s_L\rangle$, $s_i \in S$, which determines the type of the lth site of any message $\vec{m}$ in the population or any protein $\vec{p}$ 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 ω here the *physicochemical accuracy scheme*. Denote by $s_\alpha \in S$ the unique and distinct site-type associated with target amino acid α. The physicochemical accuracy scheme makes the fitness of an amino acid β in a site of a given type s_α reflect the accuracy with which its physicochemical properties matches that of the target amino acid α. We thus choose the elementary fitness matrix to reflect the physicochemical distances between amino acid β and α. In mathematical terms, we require that

$$\omega(\beta|s_\alpha) = f(d(\beta|\alpha)), \tag{6}$$

where the function f remains to be defined.

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

$$\omega(\vec{p}\,|\,\vec{s}) = \prod_{l=1}^{L} \omega(p_i|s_i). \tag{7}$$

Denote by α_{s_l} the target amino acid of the lth site-type as determined

by target $\vec{s}$. In the multiplicative scheme for protein fitness we require that

$$\omega(\vec{p}\,|\,\vec{s}) = \prod_{l=1}^{L} f(d(p_l|\alpha_{s_l})) = f\Big(\sum_{l=1}^{L} d(p_l|\alpha_{s_l})\Big). \tag{8}$$

The only functional form that meets this requirement is

$$\omega(\beta\,|\,s_\alpha) = \phi^{d(\beta|\alpha)} \tag{9}$$

where for the fitness to decrease with chemical distance we also require that $0 < \phi < 1$.

We refer to ϕ as the *missense tolerance parameter* because it determines the overall strength of selection against missense in proteins. Increasing ϕ increases the tolerance of selection to missense over all sites in proteins.

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 gene, as created for example through translational ambiguity, are arithmetically averaged to determine overall individual fitness. Let the *usage*, $u(i|s)$, be the frequency of codon $i \in C_{II}^B$ in sites of type $s \in S$ in a given message $\vec{m}$, satisfying $u(i|s) \geq 0$ for all $i \in C_{II}^B$ and $s \in S$ and $\Sigma_{i\in C_{II}^B} u(i|s) = l_s$ for all $s \in S$. The fitness $\omega(c,\vec{m})$ of an individual with message $\vec{m}$ and code c given a target $\vec{s}$ may then be written as

$$\omega(c,\vec{m}) = \prod_{s\in S}\prod_{i\in C_{II}^B}\Big(\sum_{\alpha\in A} c(\alpha\,|\,i)\omega(\alpha\,|\,s)\Big)^{u(i|s)}. \tag{10}$$

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) 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_{II}^B encodes all of the 20 amino acids in A with equal probability, i.e.,

$$c_0(\alpha\,|\,i) = \frac{1}{M} \quad \text{for all} \quad i \in C_{II}^B \quad \text{and} \quad \alpha \in A, \tag{11}$$

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

Code Mutation. We assume *uniform discontinuous code mutation*, in which: (1) the meaning of a codon may only change to *explicitly* encode a single amino acid with unit probability, i.e., $c(\alpha|i) = 1$ for some amino acid α, and $c(\beta|i) = 0$ every other amino acid $\beta \neq \alpha$, (2) codons change meanings independently of one another within and across generations, (3) the probabilities of changes in meaning are identical over all codons and all amino acids, and (4) the probability μ_{cm} of changing the meaning of any one codon is very small, so that the probability of two changes in meaning to a code is negligible.

The probability $\mu_o(c'|c)$ of mutating to a code c' from a code c is then

$$\mu_c(c'\,|\,c) = \begin{cases} 1 - N\mu_{cm} & c' = c \\ \mu_{cm} & c' \text{ differs from } c \text{ by one allowed change} \\ 0 & \text{otherwise,} \end{cases} \tag{12}$$

where $N = 16$ is the number of codons in C_{II}^B.

The Code-Message Coevolutionary Dynamic. With the assumptions outlined above, an infinite-sized asexual population with a unique genetic code c will converge to a unique codon usage distribution $U(c) = \{u_c(i|s)\}_{s\in S}^{i\in C_{II}^B}$ in messages, and a unique growth rate $\lambda(c)$ in mutation-selection equilibrium. Also, assuming that messages are long, the variance in codon usage among individual messages in the population is small (Sella and Ardell 2001). We call the unique code with which an equilibrium population translates its messages the *established code*.

We use these results and Equation 10 to calculate the *invasion fitness* of a mutant code, which is the fitness of an individual using the mutant code to translate a message with the expected codon usage frequencies of messages equilibrated to the established code. The invasion fitness of an individual with an altered genetic code c' and a message $\vec{m}_c$ with the equilibrium codon usage distribution $U(c)$ of the established code c is

$$\omega(c',\vec{m}_c) = \prod_{s\in S}\prod_{i\in C_{II}^B}\Big(\sum_{\alpha\in A} c'(\alpha\,|\,i)\omega(\alpha\,|\,s)\Big)^{u_c(i|s)}. \tag{13}$$

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 take over 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.

More specifically, if at any step $\omega = (c', \vec{m}_c) > \lambda(c)$ for some mutant code c', the mutant code with the greatest invasion fitness is assumed to take over the population and become the new established code (codes with equal maximal invasion fitnesses have equal probability to take over). If $\lambda(c) \geq \omega(c', \vec{m}_c)$ for all mutant codes c', then no mutant code invades and the established code is said to have become *frozen*.