The Coevolution of Genes and the Genetic Code

Guy Sella
Dave Ardell

SFI WORKING PAPER: 2001-03-015

SANTA FE INSTITUTE

The Coevolution of Genes and the Genetic Code

Guy Sella [*†‡] David H. Ardell [§]

February 19, 2001

The Standard Genetic Code (SGC) is the mapping of nucleic acids into polypeptides that is employed, some times with minor variations[1], in every organism, organelle and virus. The organization of the SGC is highly non-random[2–8]. In the four decades since the discovery of the SGC a large spectrum of hypotheses have been conceived to explain how its organization came about. These include a variety of load minimizing hypotheses[2,3,5,6,9–15], the frozen accident hypothesis[16], the ambiguity reduction hypothesis[17,18], the stereochemical hypothesis[14,16,19–25], and the metabolic coevolutionary hypothesis[26,27]. None of these hypotheses has laid down a theory that is fully fledged in the sense that it (i) begins from biological or biochemical considerations, (ii) derives the evolutionary mechanisms that follow from such considerations, and (iii) shows how these mechanism can reproduce the patterns in the organization of the SGC. Here we present the first fully fledged theory for the evolution of the SGC. The theory derives from two fundamental observations: first, there are patterns in the SGC that strongly suggest that systematic errors in replication and translation played a causal role in

[*]Department of applied Mathematics and Computer Science, The Weizmann Institute, Rehovot 76100, Israel.
[†]Department of Biological Sciences, Stanford University, Stanford, CA 94305, USA.
[‡]e-mail: sella@wisdom.weizmann.ac.il
[§]Department of Molecular Evolution, Uppsala University, Norbyvägen 18C, 752 36 Uppsala, Sweden.

its evolution[2–8]; and second, the evolution of a genetic code is mediated through the protein-coding genes, where selection acts upon the proteins which are the product of translating these genes with the genetic code[16]. We derive the evolutionary mechanisms of code formation that follow from these observations, and show how these mechanisms reproduce two of the salient organizational patterns of the SGC.

The SGC maps codons, which are three letter words over an alphabet of four letters $\{U, C, A, G\}$, into amino acids, the twenty basic building blocks of proteins. Shortly after the code was discovered[9] many researchers noted that its organization is not random[2,10–13]. The Standard Code associates mutationally close codons with physicochemically similar amino acids. There are at least two patterns in the SGC that are correlated with systematic errors in the processes of replication and translation[2–8] (Fig. 1):

- **Pattern I:** Amino acids are more similar to each other along the first codon position than they are along the second[2–6,8]. This "column-like" pattern corresponds to a higher rate of translational misreading in the first codon position[29–31].

- **Pattern II:** Along the second codon position, amino acids associated with *pyrimidine* bases $Y = \{U, C\}$ or *purine* bases $R = \{A, G\}$ are more similar within these sets than between them[7,8]. This is associated with mutational bias in replication, in which transitions (mutations within these base sets) occur more frequently than transversions (mutations of a base in one set to a base in the other set)[32–34].

The fact that close codons encode similar amino acids implies that the organization of the Standard Code reduces the deleterious effects of errors in replication and translation. This observation led to the *load-minimizing hypothesis* (as it is called in[35]), namely, that the Standard Code may have been selected to correct errors in the processes of replication and translation of protein-coding information[2,12,13].

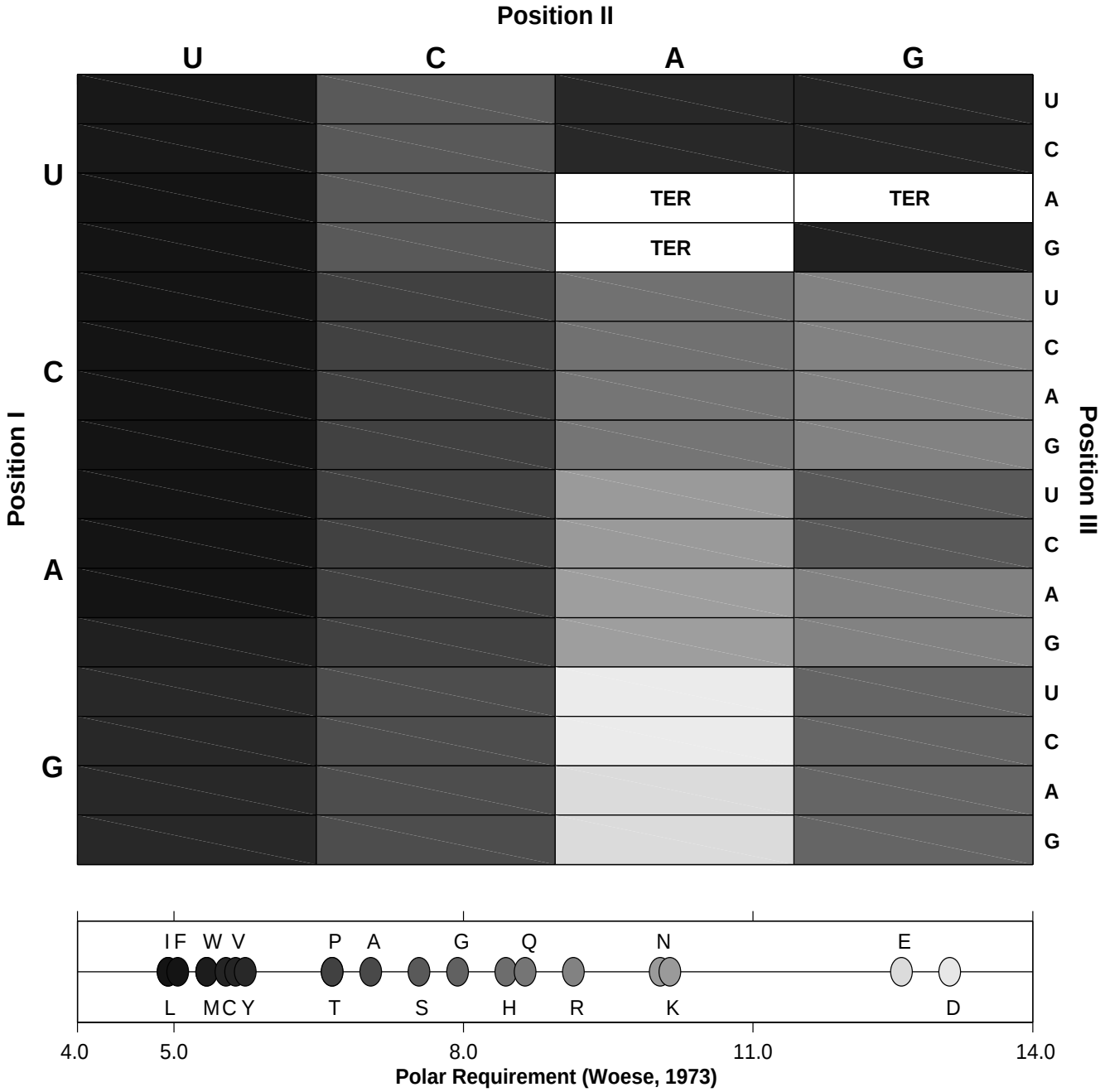

Figure 1: The table of the Standard Genetic Code shaded by a gray scale representing the Polar Requirement of the amino acids encoded. The shades corresponding to each of the 20 amino acids appear below, using the abriviated notation for each of the amino acids. Polar requirement is a physicochemical index proposed by Woese[28] for the purpose of characterizing the average physicochemical requirement on amino acids in protein sites. When two amino acids have more similar polar requirement they can, on average, replace one another more easily at a protein-site, and with less deleterious consequences. The entry corresponding to codon UCG is the rectangle corresponding to Ist position U, IInd position C, and IIIrd position G. The polar requirement of the amino acid encoded by codon UCG, which is Serine (abreviated by S), is represented by the shade of gray in the entry corresponding to codon UCG. The regularities in the organization of the SGC that are apparent in this representation are reviewed in the text.

Later, Crick noted that selection cannot act directly and independently on a genetic code[16]. At a given stage in evolution, the genetic code and the genetic message, referring to the protein-coding regions of the DNA or its precursor, are allied. The "text" in the message has been shaped by selection to code for useful proteins through the prism of the existing genetic code. Similarly, the code is under selection to produce useful proteins with the messages presented to it. Thus, any theory of SGC evolution must consider the coevolution of codes and messages. Crick surmised that code-message coevolution would have led to a frozen random genetic code, which he called a "frozen accident"[16]. Here we show that code-message coevolution in the presence of systematic errors in replication and translation leads to precisely the organizational patterns of the SGC we reviewed above.

Our model of code-message coevolution describes the evolution of an asexual population. The genotype of each individual consists of a message, which is the concatenation of all protein-coding regions, and a genetic code. The phenotype of each individual consists of a protein distribution, which is the outcome of translating the message using the code, but with systematic errors. (Fig. 2). In order to determine the fitness associated with the protein distribution we classify all protein sites into types, within which the fitness contribution of each of the amino acids is pre-defined. For example, when an amino acid a appears at a site of type s it will contribute some fixed increment $w(a, s)$ to the overall fitness. Overall fitness is calculated by multiplying the fitness increments across sites, and arithmetically averaging the products across proteins in the distribution. Individuals in the population reproduce in proportion to their fitness, where systematic mutations in messages and codes occur in replication.

We assume that messages change much faster than codes, and therefore code-message coevolution takes the form of the quasistatic approximation described in Fig. 3: At the initial step $t = 0$ all the individuals in the population have the same initial code c_0 (Fig. 3: Oval 1). The messages

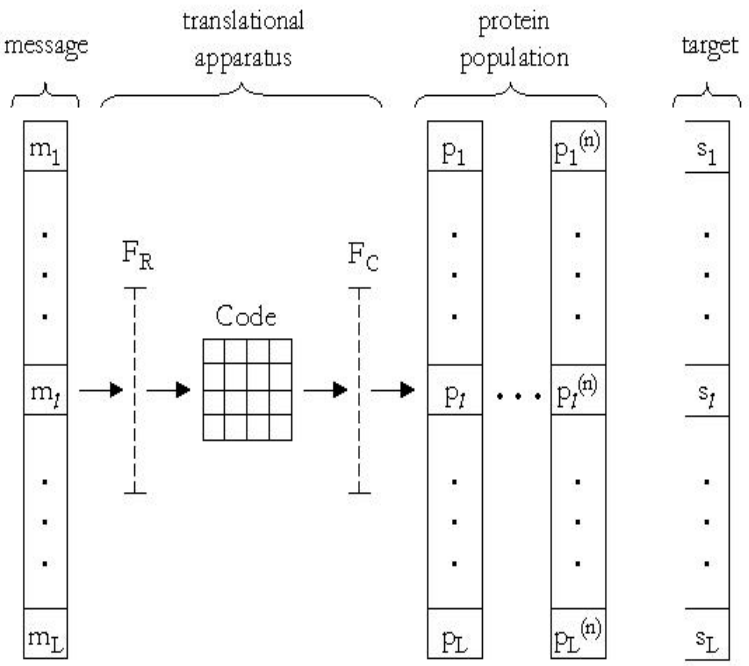

Figure 2: A schematic description of the way the genetic code and message of an individual determine its protein distribution, and consequently its fitness. A codon m_l in the message (on the left) is translated by the translational apparatus to produce an amino acid p_l at the lth protein site. Due to the systematic errors, represented here by the misreading filter F_R and the mischarging filter F_C, translation of codon m_l may be different on different occasions and therefore $p_l, ..., p_l^{(n)}$ are not necessarily the same amino acid. The fitness contribution of an amino acid p_l is determined by the type of site l, s_l, which appears on the right. The overall fitness of the individual is given by multiplying the fitness contributions across sites, and arithmetically averaging the products across the protein distribution.

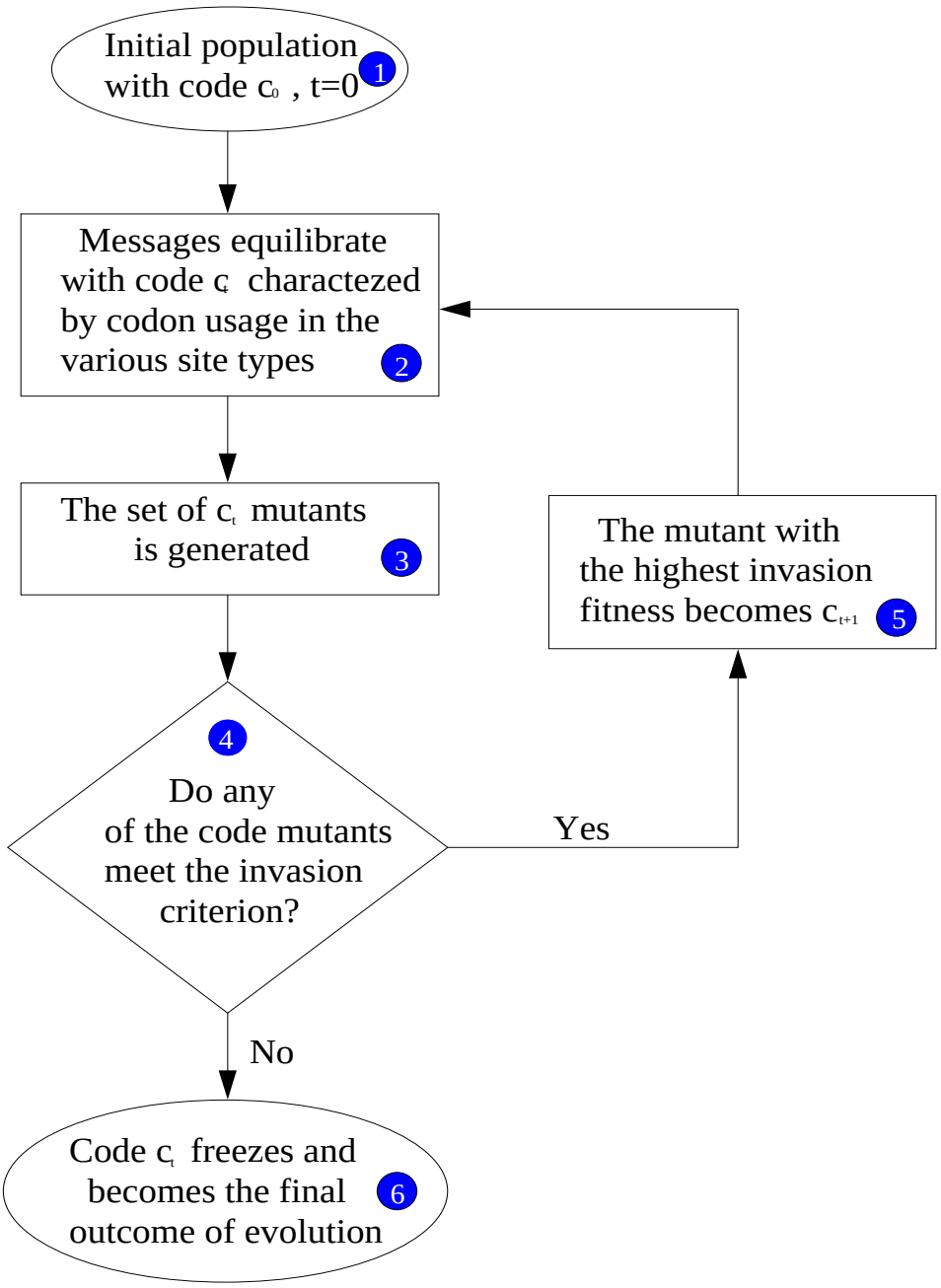

Figure 3: The simplified code coevolutionary dynamics in the quasistatic approximation. The numbers that appear in the boxes refer to the explanation in the text.

of a population with a given code c_t ($t \geq 0$) attain a mutation-selection balance, which can be characterized in terms of the codon usage in the the various site-types (Fig. 3: Box 2). A set of mutant codes that derive from c_t is generated (Fig. 3: Box 3). Given the codon usage at the mutation-selection balance, the fitness of code mutants with the pre-existing messages can be calculated. A mutant code that has a higher fitness with the pre-existing messages than the pre-existing code can invade the population to become the next code (Fig. 3: Diamond 4). From the mutants that meet this invasion criterion, the one that takes over the population depends on our assumptions about population structure. Here we assume it will be the code mutant with the maximal invasion fitness (Fig. 3: Box 5). Once a new code takes over the population the process returns to Box 2. When no mutant code meets the invasion criterion the coevolutionary process freezes, and the code is the final outcome of evolution (Fig. 3: Oval 6).

A logic can be discerned in the process of code-message coevolution and we illustrate this using the double ring toy model presented at the top of Fig. 4. In the double ring model we assume:

- Codon space has the structure of a ring. Each codon can mutate to become each of its neighbors on the ring. The probability of mutation per-generation is $\mu = 0.02$.

- Amino acids and site-types are defined on a ring of circumference 1, which represents a normalized physicochemical index. Site-types are defined in correspondence to amino acids and denoted by the letter associated with the amino acid to which they correspond. For example, the fitness contribution of amino acid a in a site of type d is $w(a, d) = \phi^{d(a,d)}$, where in this example $\phi = 0.25$, and $d(a, d)$ is the distance between amino acids a and d on the ring.

- It is assumed that each of the site-types is present in equal frequency.

- In the initial code, each codon may be translated into each of the amino acids with equal probability.

- Code mutants are defined by changing the meaning of one codon so that it encodes a single amino acid, where previously it was in the initial condition or encoded for a different amino acid.

- Stop codons are not considered in this or the following examples.

Code-message coevolution in the double ring model is described in Fig. 4. We observe three types of evolutionary steps, once the code already encodes some amino acids:

- *A diversifying step.* Step 2 in Fig. 4 is a diversifying step. In this step, codon 5, which is antipodal to codon 1 on the codon ring, is assigned amino acid j, which is antipodal to the encoded amino acid a on the amino acid ring. More generally, for two codons that are far from each other in codon space, if one codes for an amino acid, the other one is assigned an amino acid that is far from the encoded amino acid. Codon usage with the pre-existing code may lead to such a step because the usage of the non-encoding codon is already higher in sites in which the encoded amino acid is undesirable, due to the lack of mutational flow from the encoding codon in these sites. Therefore assigning a distant amino acid to the other codon increases fitness with the pre-existing usage.

- *A load minimizing step.* Steps 3, 4, 5, 7, 8, and 9 in Fig. 4 are load minimizing. In step 3, codon 4, which is a neighbor of codon 5 on the codon ring, is assigned amino acid a, which is similar to the amino acid j, encoded by codon 5. In general, when two codons are neighbors in codon space, and one of the two codes for an amino acid, the other one is assigned a similar amino acid. Codon usage with the pre-existing code may lead to such a step because the usage of the encoding codon's neighbor is already higher in the sites in which the encoded amino acid is desirable, and lower where this amino acid is undesirable, due to the mutational flow from its encoding neighbor. Therefore, assigning a similar amino acid to the other codon

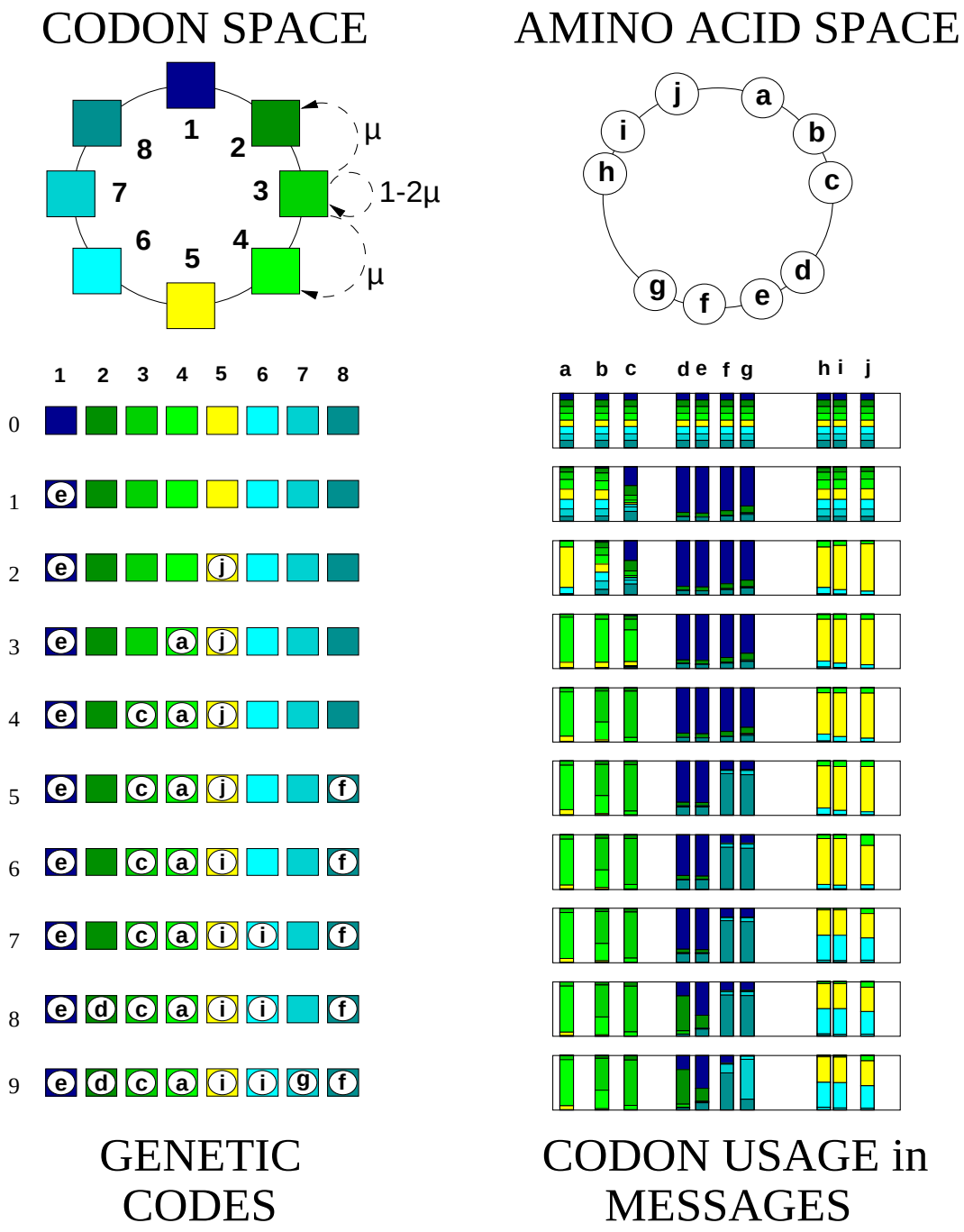

Figure 4: Code-message coevolution in the double ring toy model. The codon and amino acid spaces in the double ring example appear above. Codons are assigned shades from a gray scale corresponding to their position on the ring. Amino acids and the corresponding site-types are assigned letters according to their position on the ring. The codes corresponding to successive evolutionary steps are presented on the left. When a codon becomes encoding, the amino acid assigned to it appears in the codon; when the codon is in its initial state it appears with no letter. For a given code, the equilibrium codon usage at the classes of message sites corresponding to the different site-types appears on its right. The proportional usage of a codon in a site-class is represented by the stacked bar-graph with the corresponding shade.

increases fitness with the pre-existing usage.

- *A codon reassignment.* Step 6 in Fig. 4 is a reassignment. The addition of amino acid a in step 3 reduced the usage of codon 5 in site-type a, this enables codon 5 to be reassigned amino acid i, which better meets the requirements of its modified usage pattern. In general, the assignment of an amino acid to one codon releases usage constraints on other encoding codons, which can then be reassigned to better meet their modified usage requirements.

The notions of load-minimizing and diversifying steps are heuristic articulations of the rules that govern evolutionary pattern formation in genetic codes. These rules are structure-preserving in that amino acids that are similar to each other are associated with codons that are close to each other, and amino acids that are very dissimilar are associated with codons that are distant. As a result, in the final code in Fig 4 the ring in amino acid space is embedded in the ring of codons in a structure-preserving manner, and therefore this code is error-correcting. With these notions we can now explain the evolution of the patterns corresponding to error-correction in the Standard Code.

Fig 5 shows code-message coevolution in a more biologically realistic model of codon and amino acid spaces with transition-bias in mutation. A codon is composed of two letters over the standard alphabet of four bases. Mutations occur among the bases, and we have incorporated transition-bias. The amino acid/site-type spaces, consisting of 20 members, correspond to a distribution of a normalized physicochemical property/requirement along a one dimensional interval. In a transition-biased mutation structure, the codon space consists of four blocks (see step 0 in Fig 5), corresponding to 1st position *pyrimidines* $\{U, C\}$ or *purines* $\{A, G\}$ and second position *pyrimidines* or *purines*. Within a block each codon has two closest neighbors, which are one transition away, and a neighbor which is two transitions away. Each block has two adjacent blocks which are one transversion away, and an antipodal block which is two transversions away. The structure of codon space participates in determining the course of evolution by defining the regions of codon space across which load-

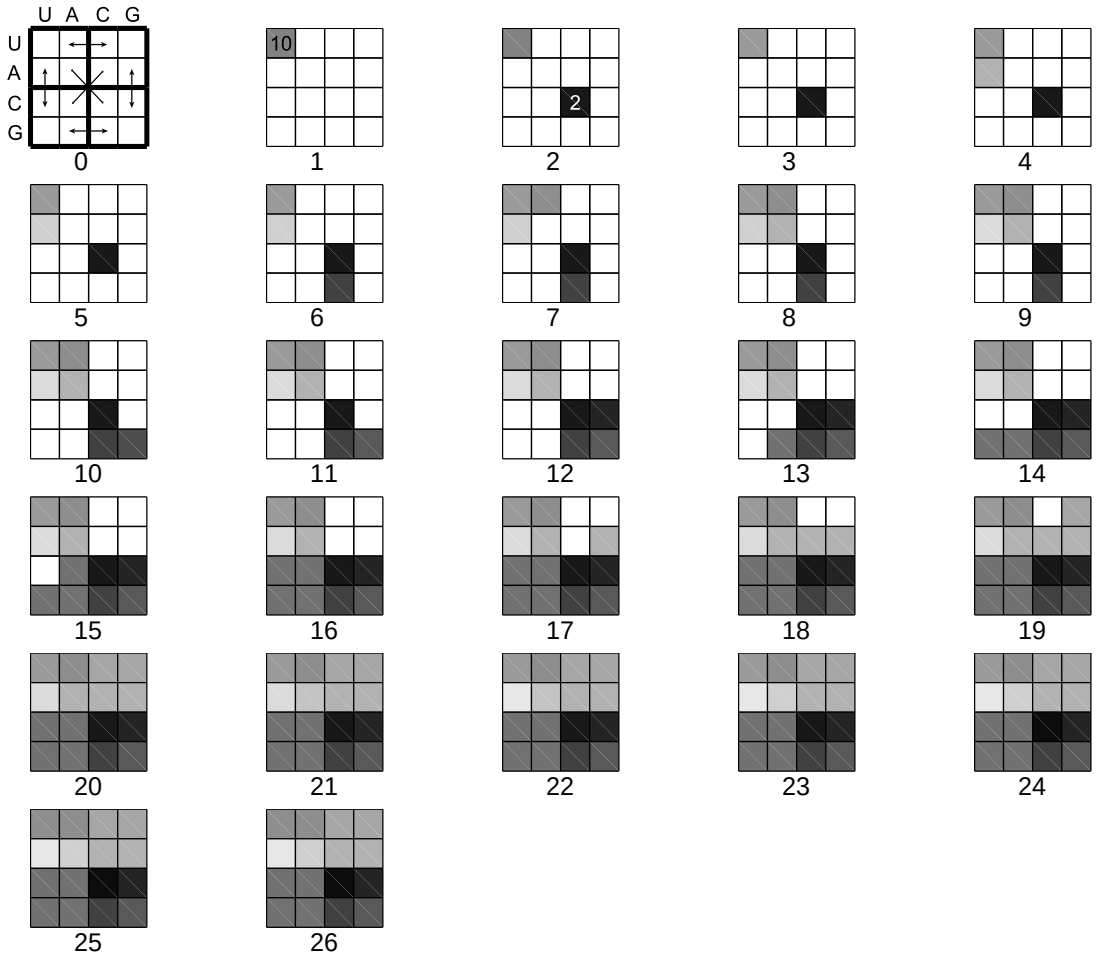

Figure 5: A typical evolution with transitional mutational bias (transition bias was taken to be 7, $\phi = 0.92$, and $\mu = 0.0006$). The 20 amino acids were chosen from a uniform distribution on the $(0, 1)$ interval, and they are represented by a gray scale in which darker shades correspond to a position closer to 1. A white entry in the code corresponds to the initial state in which a codon meaning is uniformly ambiguous across amino acids. The codes corresponding to the sequence of evolutionary steps are displayed above the step number.

minimizing and diversifying steps occur.

Code-message coevolution with transition bias, in Fig. 5, can be explained using the heuristic terms defined above:

- Step 2 is a diversifying step. It associates codon AA, which is antipodal to the existing encoding codon UU, with amino acid 2, which is the end of amino acid space furthest from the existing encoded amino acid 10. Steps 13 and 17 are also diversifying steps, which initiate the encoding in a block by associating amino acids that are far from those encoded by the other blocks.

- Steps 4, 6, 7, 8, 10, 12, 14, 15, 16, 18, 19, and 20, are all load-minimizing steps. In these steps, codons that have encoding neighbors within their block are assigned amino acids similar to those encoded by their neighbors.

- Steps 3, 5, 9, 11 and 21-26 are all reassignments.

The pattern required for the correction of transitionally-biased mutations is precisely that induced by such errors through the local fitness requirements that result in load-minimizing and diversifying steps. It is the pattern of blocks shown at the end of the evolutionary trajectory shown in Fig. 5.

Fig. 6 describes a typical evolution with uniform positional misreading in the first codon position, with the same amino acid and site-type spaces, and the same mutation and selection parameters as in the previous example. 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 terms of load-minimizing and diversifying steps.

When both misreading along the first codon position and transition bias in mutation are introduced into the model, the pattern generated by evolution varies in accordance to the relative magnitudes of misreading and mutation. An array of final codes corresponding to different com-

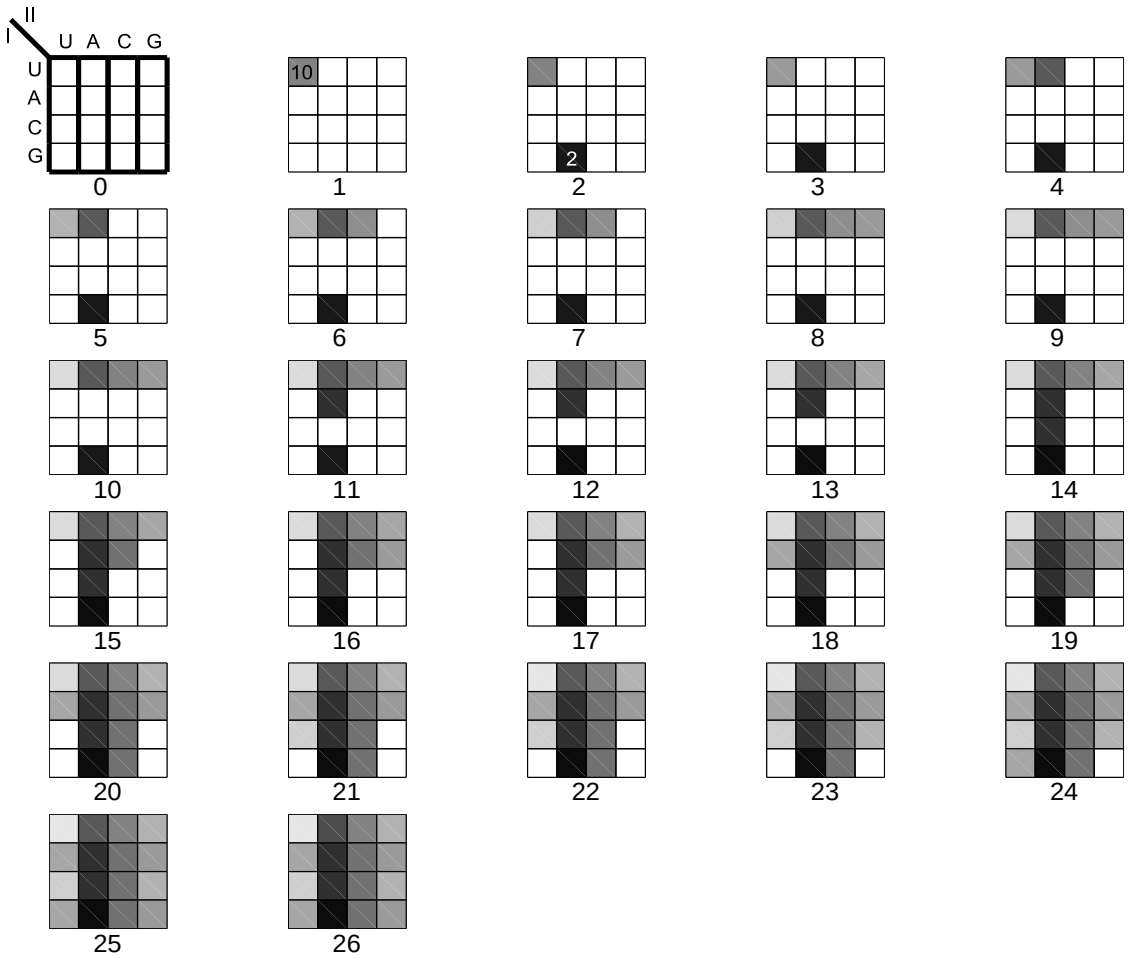

Figure 6: A typical evolutionary trajectory with uniform misreading in the first codon position (the error rate was taken to be 0.006, $\phi = 0.92$, and $\mu = 0.0006$). Steps 2, 6 and 8 are diversifying. Steps 7, 11, 14, 16, 19, 20, 21, 23, 24 and 25 are the misreading analog of load minimizing steps. Steps 3, 5, 7, 9, 10, 12, 13, 22 and 26 are all reassignments.

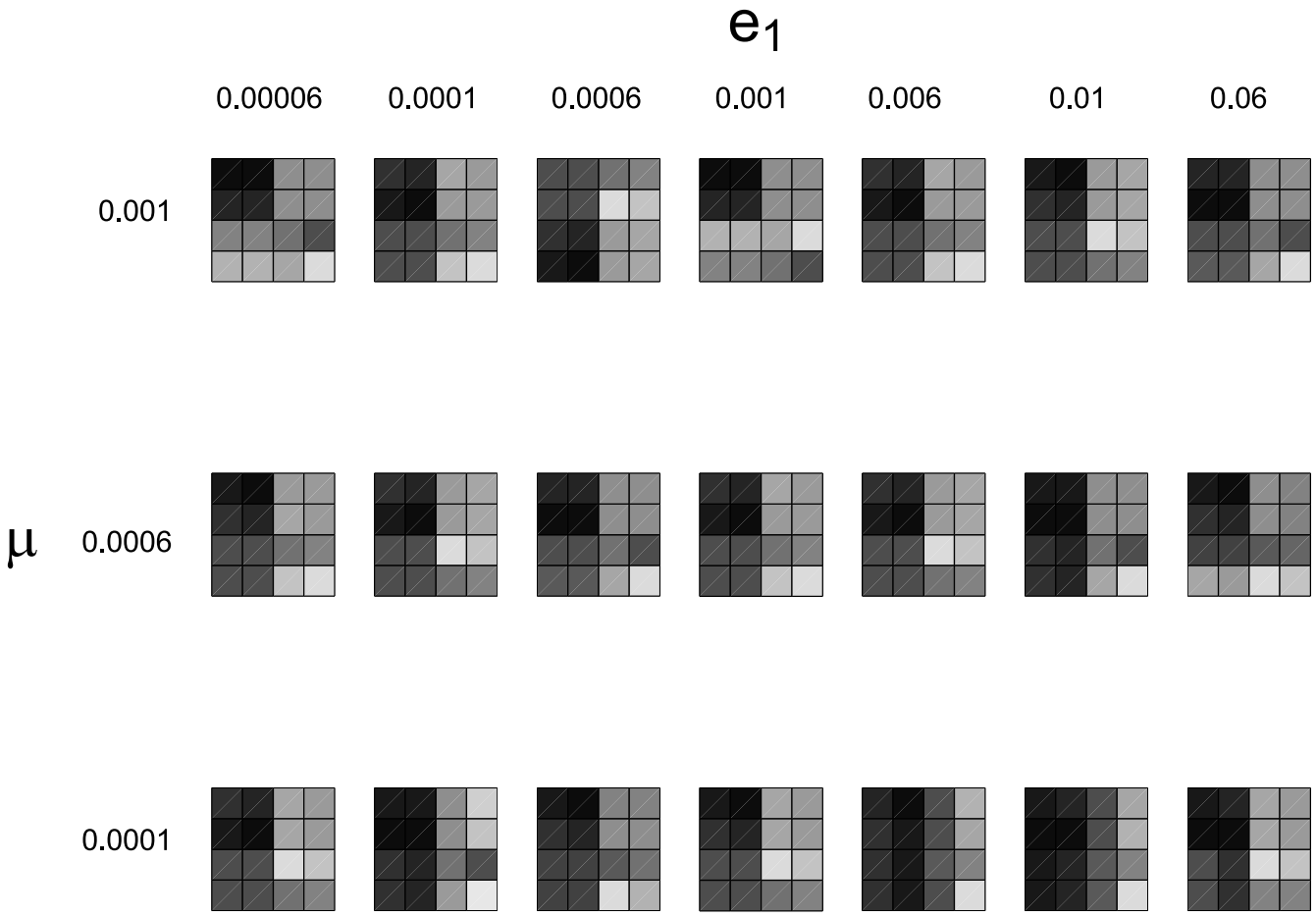

Figure 7: An array of final codes corresponding to different combinations of mutation and positional misreading. The mutation (μ) and misreading (e_1) parameters are noted in the figure, and the transition bias in mutation was taken to be 5.

binations of mutation and misreading parameters is presented in Fig. 7. At the top left corner, where mutation dominates over misreading in the first codon position, the four-block structure corresponding to transition bias in mutation is clearly pronounced. At the bottom right corner, which corresponds to the more realistic case where misreading dominates over mutation in the first codon position, the two-block structure corresponding to properties I and II of the Standard Code (Fig. 1) is reproduced.

The theory of code-message coevolution derives from the two observations noted in the beginning of this paper: first, the coevolution of genetic codes and protein-coding genes, or messages, affected the formation of the SGC; and second, the organization of the SGC suggests that systematic errors in replication and translation played a causal role in its evolution. Given a genetic code, the protein-coding messages are shaped by selection to yield useful proteins under translation with that code. In turn, the protein-coding messages determine which code variants are advantageous to fitness when presented with the existing messages, and can therefore invade and take over the population to form the next code in evolution. When systematic errors in replication and translation are introduced, the constraints of code-message coevolution imply rules for the way a genetic code can change in an evolutionary step. These rules take the form of the load-minimizing and diversifying steps defined above. The iterated application of these steps may be conceived as a process of pattern formation that endows the resulting final code with its organizational properties. Here we show that the theory of code-message coevolution provides a way to understand how and why the salient organizational features of the SGC came about.

Acknowledgement

We thank Marcus W. Feldman, Ilan Eshel, Aaron Hirsh, Michael Lachmann, Tuvik Becker, Ben Kerr, Jenifer Hughes, Susan Ptak, Luaren Ancel, and Emile Zuckerkandl for their substantial help,

support and comments at various stages of this work.

References

1. Osawa, S., Jukes, T., Watanabe, K. & Muto, A. Recent evidence for evolution of the genetic code. *Microbiol. Rev* **56**, 229–264 (1992).

2. Woese, C. On the evolution of the genetic code. *Proc. Natl. Acad. Sci. USA* **54**, 1546–1552 (1965).

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

4. Kimura, M. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *J. Mol. Evol.* **16**, 111–120 (1980).

5. Swanson, R. A unifying concept for the amino acid code. *Bull. Math. Biol.* **46**, 187–203 (1984).

6. Haig, D. & Hurst, L. D. A quantitative measure of error minimization in the genetic code. *J. Mol. Evol.* **33**, 412–417 (1991).

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

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

9. Nirenberg, M., Jones, O., Leder, P., Clark, B., Sly, W. *et al.*. On the coding of genetic information. *Cold Spring Harbor Symposia on Quantitative Biology* **28**, 549–558 (1963).

10. Sonneborn, T. Degeneracy of the genetic code: extent, nature, and genetic implications. In

Evolving Genes and Proteins (eds. Bryson, V. & Vogel, H.), 377–397 (Academic Press, New York) (1965).

11. Zuckerkandl, E. & Pauling, L. Evolutionary divergence and convergence in proteins. In *Evolving Genes and Proteins* (eds. Bryson, V. & Vogel, H.) (Academic Press, New York) (1965).

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

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

14. Woese, C., Dugre, D., Dugre, S., Kondo, M. & Saxinger, W. On the fundamental nature and evolution of the genetic code. *Cold Spring Harbor Symp Quant Biol.* **31**, 723–736 (1966).

15. Goldman, N. Further results on error minimization in the genetic code. *J. Mol. Evol.* **37**, 662–664 (1993).

16. Crick, F. The origin of the genetic code. *J. Mol. Biol.* **38**, 367–379 (1968).

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

18. Fitch, W. & Upper, K. 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 (1987).

19. Woese, C. Evolution of the genetic code. *Naturwissenschaften* **60**, 447–459 (1973).

20. Pelc, S. & Welton, M. Stereochemical relationship between coding triplets and amino acids. *Nature* **209**, 868–872 (1966).

21. Dunnill, P. Triplet nucleotide-amino acid pairing: a stereochemical basis for the division of between protein and nonprotein amino acids. *Nature* **210**, 1267–1268 (1966).

22. Hopfield, J. Origin of the genetic code: a testable hypothesis based on tRNA structure, sequence and kinetic proofreading. *Proc. Natl. Acad. Sci. USA* **75**, 4334–4338 (1978).

23. Shimizu, M. Molecular basis for the genetic code. *J. Mol. Evol.* **18**, 297–303 (1982).

24. Knight, R. & Landweber, L. Rhyme or reason: RNA — arginine interactions and the genetic code. *Chemistry & Biology* **5**, R215–R220 (1998).

25. Knight, R., Freeland, S. & Landweber, L. Selection, history and chemistry: the three faces of the genetic code. *Trends Bioch. Sci.* **24** (1999).

26. Taylor, F. & Coates, D. The code within the codons. *BioSystems* **22**, 177–187 (1989).

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

28. Woese, C. *The Genetic Code: The Molecular Basis for Genetic Expression* (Harper & Row, New York) (1967).

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

30. Davies, J., Jones, D. & Khorana, H. 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 (1966).

31. Parker, J. Errors and alternatives in reading the universal genetic code. *Microbiol. Rev* **53**, 273–298 (1989).

32. Freese, E. . *Proc. Natl. Acad. Sci. USA* **45**, 622 (1959).

33. Topal, M. & Fresco, J. Complementary base pairing and the origin of substitution matricies. *Nature* **263**, 285–293 (1976).

34. Echols, H. & Goodman, M. Fidelity mechanisms in DNA replication. *Ann. Rev. Biochem.* **60**, 477–511 (1991).

35. Maynard-Smith, J. & Szathmary, E. *The Major Evolutionary Transitions in Evolution* (W.H. Freeman, Oxford, UK) (1995).