The Origin and Evolution of the Genetic Code:
Statistical and Experimental Investigations

Robin Douglas Knight

A DISSERTATION PRESENTED TO THE FACULTY OF PRINCETON UNIVERSITY IN CANDIDACY FOR THE DEGREE OF DOCTOR OF PHILOSOPHY

RECOMMENDED FOR ACCEPTANCE BY THE DEPARTMENT OF ECOLOGY AND EVOLUTIONARY BIOLOGY

June 2001

Abstract

The structure of the genetic code has puzzled researchers since codon assignments were first elucidated, but recent genome and aptamer sequences allow quantitative testing of hypotheses about the code's origin, evolution, and subsequent diversification. My major findings can briefly be categorized by stage of evolution:

1. There is strong statistical support for two major theories about the evolution of the canonical code. The code is highly resistant to errors relative to random codes, and there is a strong statistical association between at least some codons and the RNA binding sites for their cognate amino acids. I found no support for specific coevolutionary models, although the code presumably evolved from a simpler form.
2. The exact time of code optimization remains unclear, but measures of amino acids that would have been more important early in evolution generally make the code appear more optimal. Thus codon assignments were perhaps determined early, in an RNA world. Aptamer data supports this idea, though equivocally.
3. Modern variant codes are probably recently-derived, nearly neutral mutants of the standard code. They may be adapted to the specific mitochondrial environments where most of them occur, although I did not find any supporting evidence. There is no support for the idea that variant codes are optimized for reduced tRNA number. There is strong support for the idea that code evolution proceeds through an ambiguous intermediate. The molecular basis for certain code changes can be elucidated: in ciliates, the mutations in the release factor eRF1 that prevent stop codon recognition can be recaptured by statistical techniques applied to phylogenies.
4. The genetic code is not just the pattern of codon assignments: it is also the frequency with which each codon is used. Surprisingly, the vast interspecific variation in codon and amino acid usage can largely be recaptured by a neutral model that assumes only purifying selection.

Despite these advances, many fundamental questions — such as the order in which amino acids were added to the code, and the extent to which stereochemistry and natural selection contributed to modern codon assignments — are still to be resolved.

Acknowledgements

Science is fundamentally a collaborative activity: no finding makes sense except in its social and historical context. This thesis is no exception. I thank my advisor, Laura Landweber, for all her help and advice during the four years I spent in her lab, and for bringing together such an exciting group of researchers. Much of my research was done in collaboration with Steve Freeland, who brought a completely different perspective on code evolution: this thesis would have been very different (and much worse!) without Steve's input. Thanks also to Cathy Lozupone for her tireless cloning, sequencing, and phylogeny building, to Tam Horton and Vernadette Simon for advice and discussion, and to Drew Ronneberg, Pete Simon, Laura Katz, Ed Curtis, Irina Pokrovskaya, Dirk Faulhammer, Anthony Cukras, Tai-Chih Kuo, Amy Horton, Wei-Jen Chang, Christina Burch, Man-Kyoon Shin, and other past and present members of the Landweber lab for interesting conversation and fun times. Thanks also to my thesis committee, Henry Horn, Hope Hollocher, Lee Silver, and Michael Hecht, and to Steve Pacala, the director of graduate studies, for advice, direction and guidance.

With interdisciplinary projects, outside collaborations are essential. Here I especially thank Mike Yarus, who put up with me for several months in his lab at the University of Colorado, Boulder, while I learned how to do SELEX experiments correctly, and who contributed greatly to the site/codon association work and to the tests of hypotheses about the evolution of variant codes. Thanks also to members of the Yarus lab: Anastasia Khrorova, Vasant Jadhav, Krishna Kumar, Alexandre Vlassov, Jason Bobe, and, especially, to Irene Majerfeld and Mali Illangasekare, whose patience and helpfulness are simply amazing. While in Boulder, I also had the pleasure of interacting with Noboru Sueoka, Leslie Leinwand, Norm Pace, and Ravinder Singh (and members of the Pace and Singh labs), and I look forward to interesting times in the MCDB department at Colorado. I also thank Jean Lobry at the Universite Claude Bernard, Guy Sella and Michael Lachmann at the Weismann Institute, Stephen Sowerby and David Liberles at the University of Stockholm, Erik Schultes at the Whitehead Institute, Warren Tate at the University of Otago, Michael New at NASA Ames, Dawn Brooks and Jacques Fresco here in Princeton for useful comments and discussion.

I'd also like to acknowledge the many good friends I've made while at Princeton. Special thanks go to the roommates who've put up with me over the years, and who helped make five years in New Jersey tolerable: Amanda Birmingham, Jeremy Peirce, Jared Anderson, David Nadler, Bob Derose, Kiran Kedlaya, and Kyle Morrison. Thanks also to my cohort (also support group) of EEB grad students: Ethan Pride, Beth Macdougal-Shackleton, Kate Rodriguez-Clark, Rae Winfree, Jen Gee, and Nipada Ruankaew; and to other people in the department including (but by no means limited to!) Todd Vision, Diane O'Brien, Mark Roberts, Philippe Tortell, Jon-Paul Rodriguez , Jason Wilder, Stuart Sandin, Eric Dyreson, Tatjana Good, and Jayatri Das.

The EEB administrative staff were absolutely critical for getting things done. Thanks especially to Mona Fazio, Christine Samburg, and Trish Turner-Gillespie for processing innumerable purchase orders and reimbursement forms, and to Mary Guimond for sorting out teaching and funding and for her encyclopedic knowledge of university procedure. Thanks also to Richard Smith for guiding me through the minefield of requirements for finals submission of the thesis, and to the people at the Visa Office for helping me not get deported.

Finally, I'd like to thank my parents for their constant encouragement and support, and my teachers and lecturers in Dunedin, especially Laurie McAuley and Stuart McConnell, who fostered my interest in biology; David Fenby, who rekindled my interest in chemistry, Paul Griffiths, who got me interested in the idea of genetic information; and Warren Tate, who introduced me to the RNA World.

Table of Contents

Preface

This thesis is organized into three main sections. The first section gives a general introduction to the genetic code, the RNA world, and the overall phenomena to be explained. The second section deals with the origin of the canonical genetic code prior to the LUCA, the Last Universal Common Ancestor of modern organisms. The third section explores the ways in which genetic codes have changed since the LUCA, including the diversification of variant genetic codes in both nuclear and mitochondrial lineages.

Many of these chapters have already appeared in print or have already been accepted for publication, as shown in the following list:

1.4 Knight, R. D. and L. F. Landweber (2000). "The Early Evolution of the Genetic Code." Cell **101**: 569-572.

1.5 Knight, R. D., S. J. Freeland, and L. F. Landweber (1999). "Selection, history and chemistry: the three faces of the genetic code." Trends Biochem Sci **24**(6): 241-7.

2.2 Knight, R. D. and L. F. Landweber (1998). "Rhyme or reason: RNA-arginine interactions and the genetic code." Chem Biol 5(9): R215-20.

2.3. Knight, R. D. and L. F. Landweber (2000). "Guilt by association: the arginine case revisited." RNA 6(4): 499-510.

3.1 Knight, R. D., S. J. Freeland and L. F. Landweber (2001). "Rewiring the keyboard: evolvability of the genetic code." Nat Rev Genet 2: 49-58.

3.2` Knight, R. D., L. F. Landweber, and M. Yarus (2001). "How mitochondria redefine the code." J Mol Evol, forthcoming 2001.

3.3 Lozupone, C. A., R. D. Knight and L. F. Landweber (2001). "The molecular basis of genetic code change in ciliates." Current Biology 11: 65-74.

3.4 Knight, R. D., S. J. Freeland, and L. F. Landweber. (2001) "A simple model based on mutation and selection explains compositional trends within and across genomes." GenomeBiology 2:4, http://www.genomebiology.com/2001/2/4/research/0010/.

I have also contributed to the following publications, although my contribution to these was not sufficient to warrant their inclusion in this thesis:

1. Freeland, S. J., R. D. Knight and L. F. Landweber (1999). "Do proteins predate DNA?" Science 286(5440): 690-2.

2. Freeland, S. J., R. D. Knight and L. F. Landweber (2000). "Measuring adaptation within the genetic code." Trends Biochem Sci 25(2): 44-5.

3. Freeland, S. J., R. D. Knight, L. F. Landweber and L. D. Hurst (2000). "Early Fixation of an Optimal Genetic Code." Mol Biol Evol 17(4): 511-518.

I have included a few paragraphs at the start of each section explaining how the different chapters relate to each other. I also preface each chapter with some historical background explaining the motivation for the research, and, where appropriate, the contributions of the various authors. Tables, figures and legends can be found at the end of each section, and are numbered according to their position within the section.

This thesis covers a wide range of material. Because many of the review chapters have already been published, and already incorporate some of my own research, I take this opportunity to give both an overview of the field as it was when I started work on this topic in 1997 and to outline what can be found in each of the review chapters.

Excitement in genetic code research began with cosmologist George Gamow's proposal in 1954 that the genetic code arose from direct interaction between DNA and proteins, reached a climax in the early 1960s as the actual codon assignments were first worked out, but was largely stifled in 1968 when Francis Crick surveyed the speculative modeling and theoretical work and found it lacking, proposing the 'frozen accident' theory of textbook fame. Although Crick's review was a fair and definitive evaluation of the evidence then available, and called for more experimental evidence bearing on the origin and evolution of the genetic code, it unfortunately had the effect of marginalizing research in the field. By 1990, few authors published more than a single paper on the genetic code, focusing instead on more mainstream (and rewarding) areas of inquiry.

All of the fundamental ideas about how the code might have evolved were proposed very early. Tracy Sonneborn, Emile Zuckerkandl, and others suggested that the code was adapted for error minimization, such that point mutations would tend to substitute similar amino acids. Carl Woese argued that the error minimization acted during translation, rather than mutation, because of reading frame-dependent effects. Woese, and a plethora of model builders, also suggested that there might have been some chemical relationship that assigned similar codons to similar amino acids, although evidence beyond the fact of the order in the code itself was not forthcoming. The idea that the code might have evolved from an earlier form, and that metabolic relatedness might have clustered certain amino acids together, was also suggested by Pelc and further elaborated by Crick. Unfortunately, the 'frozen accident' theory allowed molecular biologists uncomfortable with evolution to discontinue active research into the topic. Chapter 1.1 gives an overview of the observed order in the code, and Chapter 1.2 gives an overview of where the components of the translation apparatus might have come from. Chapter 2.4 investigates more extensively whether the choice of components in the translation apparatus might be adaptive.

The next two decades produced only sporadic research into genetic code origins, although the same three themes — selection, chemistry, and history — persisted. In 1969, Alff-Steinberger showed that random codes did far worse than the actual code at minimizing the average effect of point substitutions; this work was ignored for over two decades (although we have been unable to reproduce any of the numerical results, so it may have been neglected for good reason). Stereochemical model-building persisted, providing many conflicting results and incompatible claims. One interesting approach by Lacey's group was to look at chromatographic separation of bases and amino acids on prebiotic mineral surfaces and partitioning between aqueous and organic phases. Although the results were generally inconclusive, there was a striking correlation between the hydrophobicity of the amino acids and the anticodon doublet dinucleosides. This chemical evidence is reviewed in Chapter 2.1.

The major theme of genetic code research in the 1970s and 1980s, however, was that of history and code expansion. In a seminal paper that might have been far more influential had it not been buried in *Botanical Reviews*, Dillon argued that similar amino acids had been clustered in the code because of their metabolic pathways, specified by the first position base; this finding was rediscovered by Taylor and Coates, and by Miseta, in 1989. In 1975, Wong developed the theory further with a quantitative statistical test, by which he estimated that the observed adjacency of codons of metabolic precursors and products could not be explained by chance. Later, he argued against the idea that the code had been optimized by natural selection, on the grounds that codes far better than the actual code could be invented. This method of 'distance minimization' shapes debates about code optimality even today. Coevolutionary theories of code expansion are briefly reviewed in Chapters 1.5 and 2.4.

Although it was largely unrecognized at the time, the discovery of variant genetic codes in 1979 demolished the major argument in favor of the frozen accident theory: that the code was universal and immutable. By 1990 a variety of genetic codes had been discovered, all recent (and relatively minor) variants of the 'standard' code found in the nuclear genomes of most organisms. This should have raised two compelling questions: why did the code change in some organisms and organelles, and why did the code *not* change everywhere else? The first proposal about these variant codes was that they were nonadaptive changes driven by biased, directional mutation: by this process of 'codon capture', codons could disappear from the genome and be neutrally reassigned. Not surprisingly, this theory was co-invented by Syozo Osawa and Thomas Jukes, one of the major proponents of the Neutral Theory of molecular evolution. One important detail that remained (and remains) unclear is whether the evolution of variant genetic codes has anything to do with the evolution of the canonical code: the modern tRNA/aminoacyl-tRNA synthetase system intermediates between codon and amino acid, so there is no possibility for the direct pairing that could conceivably have been important in primordial systems. Variant genetic codes are covered extensively in Section 3, and briefly in Chapters 1.5 and 2.5.

The 1990s marked a rebirth of interest in genetic code evolution. There were two enabling technologies. The first was the easy availability of powerful desktop computers for extensive statistical and simulation studies. The second was the idea of catalytic RNA and, more specifically, the invention in 1990 of SELEX (a technique for evolving RNA molecules with arbitrary functions) in the labs of Larry Gold, Gerald Joyce, and Jack Szostak. The discovery that RNA could act as a catalyst as well as a store of information implied the possibility of an RNA World, a period before proteins in which RNA was the only macromolecule (reviewed in Chapter 1.3). Many of the components of the translation apparatus are made of RNA, and there was suggestive evidence as early as 1992 that the catalytic component of the ribosome might actually be a ribozyme (an RNA enzyme), although it was not until eight years later that this was finally confirmed. This made the RNA world a compelling milieu for genetic code evolution, especially since it provided a solution to the fundamental problem that many essential components of the translation apparatus are made of protein, and cannot themselves have predated coded protein synthesis. SELEX experiments that isolated functional RNA molecules provided a means of testing whether particular activities could have been available in an RNA world.

All three explanations for the structure of the genetic code, selection, history, and chemistry, were being actively defended by 1997. Haig and Hurst repeated Alff-Steinberger's experiment in 1991, using a range of measures of amino acid similarity, and found highly significant conservation of polarity. The following year, Szathmary and Zintzaras showed by matrix correlation analysis that tRNAs with similar amino acids were likely to have similar sequences and anticodons, and that the amino acids were likely to be connected by fewer metabolic steps than would be expected by chance. Szathmary also showed that codon swapping by directional mutation pressure could potentially act as a pathway for swapping arbitrary codons without deleterious effects, allowing adaptive evolution of the genetic code. However, these results were controversial: at the same time, Di Giulio used Wong's metric of distance minimization to argue that the code was in fact only weakly optimized by natural selection, and that the primary factor shaping codon assignments was metabolic relatedness. This metabolic theory received support from the finding that several of the aminoacyl-tRNA synthetases were closely related to each other, in particular product/precursor pairs such as Glu/Gln and Asp/Asn: in fact, glutamyl-tRNA synthetase is paraphyletic, the glutamine-charging enzyme being derived from within the eukaryotes.

Meanwhile, Yarus had observed that the *Tetrahymena* Group I intron had a binding site for arginine that was composed almost entirely of arginine codons, proposed that this was an example of a type of intrinsic affinity that led to the modern genetic code, and selected RNA aptamers to several amino acids to test whether these interactions were reliably recovered from random sequence. Several amino acid aptamers from other groups were also available,

and the first NMR structure (of one of Famulok's aptamers) was published in 1996, giving a clear picture of the binding interactions for the first time. Amazingly, the arginine aptamer had been evolved from a citrulline aptamer, from which it differed by only three point mutations, which contributed to the formation of two new Arg codons (though the authors had not noted this fact).

There were also several competing theories to explain the origin of variant genetic codes. The orthodox view was Osawa and Jukes's Codon Capture hypothesis (outlined above), but Andersson and Kurland proposed that most variants were selected for genome minimization, and Schultz and Yarus proposed that variants arose through ambiguous intermediate tRNAs that could translate a single codon with two meanings. Although debate had been acrimonious, few quantitative tests had been brought to bear on the question. The vast influx of complete mitochondrial genomes at the close of the decade finally allowed such tests to be formulated and carried out (Chapters 3.2 and 3.3). Additionally, the molecular basis of many of the changes in tRNAs that contribute to variant codes have now been worked out biochemically, providing insight into the specific mechanisms involved.

Thus, when I started work on the genetic code, there were two entirely separate domains of inquiry: the origin of the canonical code, and the evolution of variant codes. Although the existence of the latter has profound implications for the former, the fact had not been widely incorporated into work on code evolution. Worse, although there was considerable evidence for both stereochemical and adaptive effects on code structure, these (and code expansion) were almost invariably presented in the literature as competing, mutually irreconcilable explanations for the *entire* set of codon assignments: the assumption was that only one had acted, or, if more than one, then all traces of the others must have been erased. In this thesis, I have tried to apply rigorous, quantitative tests to as many aspects of code evolution as possible, and, where there is evidence for several independent lines of explanation, to allow pluralistic explanations. I hope that fostering détente among the various factions will contribute to a more productive atmosphere, and perhaps encourage more researchers to enter the field at this uniquely exciting time.

1 Introduction

This section provides an overview of the genetic code as an object of study, and provides some historical and biochemical background. I have not included much detail on mechanisms of translation, as these are adequately covered by many biochemistry textbooks, but rather focus on more speculative material.

Chapter 1.1 gives an overview of what the genetic code is, how the codon assignments were first discovered, and some of the patterns in the genetic code that need to be explained. Chapter 1.2 gives an overview of where the metabolites essential to life might have come from originally, especially the seminal work of Miller on prebiotic synthesis of amino acids, and reviews some ideas about how these simple organic monomers could have evolved into a system capable of replication with 'unlimited inheritance'. Chapter 1.3 covers the RNA World hypothesis, the idea that RNA preceded both proteins and amino acids, playing both the catalytic role of the former and the informational role of the latter. Chapter 1.4, which appeared in Cell in June 2000, provides a concise and relatively gentle summary of recent evidence bearing on the early evolution of the genetic code. Finally, Chapter 1.5, which appeared in TiBS in June 1999, reviews the evidence for the three main hypotheses about the evolution of the canonical genetic code, and sets the stage for the more detailed investigation of these hypotheses in Section 2.

1.1 The Genetic Code

This chapter introduces the star of this thesis, the 'universal' or 'canonical' genetic code found in most organisms. I give a brief overview of the experimental techniques used to work out the genetic code table, the chemical ordering of the genetic code table, and briefly introduce the idea that the genetic code has changed in several recent lineages (although this idea is more fully explored in Section 3).

1.1.1 Introduction

In modern organisms, the genetic code provides the link between inheritance and development. By establishing a one-to-one correspondence between nucleic acids and proteins, the genetic code allows stable inheritance of the phenotypic variation on which selection acts. Before the genetic code evolved, however, primitive organisms must have used either of two simpler strategies: (a) inheritance of metabolic states rather than of physical carriers of information (limited replicators), or (b) restriction of phenotypes to nucleic acids and their reactions (unlimited replicators) (Szathmáry and Maynard Smith 1995). The former limits the transmission of variability, while the latter limits the range.

The stage at which the genetic code developed should affect its properties in predictable ways. Theories of the origin of the genetic code can be broadly divided into "early" theories, which imply that the genetic code developed before macromolecules were common, and "late" theories, which imply that it developed after macromolecules were widely available. Early development of the genetic code implies greater reliance on simple stereochemical interactions, since complex catalysts would have been absent. For instance, the genetic code may have been established at the start of life, when most "metabolites" were actually synthesized by abiotic processes. If so, there must have been some stereochemical mechanism allowing specific pairing between the limited repertoire of amino acids and nucleic acids available at the time. Alternatively, the genetic code may have developed in the context of a chemoton (Gánti 1975), an autocatalytic chemical system composed primarily of small molecules. If so, the choice of amino acids would not be restricted to those with prebiotically plausible syntheses, but the code could only be established if simple stereochemical relationships exist between those amino acids and oligonucleotides.

Late theories, in contrast, are necessarily statistical because macromolecules acting as adaptors could enhance any arbitrary pairing between codons and amino acids. Thus, the genetic code might be a “frozen accident” (Crick 1968), persisting because any change would be deleterious. Perhaps the most popular of the late theories is the hypothesis that the genetic code arose from the RNA world (Gilbert 1986), a hypothetical metabolism relying exclusively on RNA as a catalyst. If so, (a) RNA catalysts (ribozymes) must be capable of catalyzing amino acid biosynthesis and peptide condensation, and (b) RNA must be capable of discriminating between and binding to amino acids. In this case, any stereochemical relationship between amino acids and oligonucleotides would influence the genetic code. Alternatively, the genetic code may have been established by progressive refinements of the translation apparatus (Woese 1967), with each generation of proteins providing greater accuracy by discriminating more finely between related amino acids (some of which may only become available after early protein catalysts became active (Wong 1975)). If so, there should be similarities between codons for related amino acids, although the block of codons assigned to each group might be arbitrary if the placement of new amino acids were determined by the adaptive consequences of taking over a group of codons in preexisting proteins rather than by intrinsic affinities between amino acids and RNA. In Chapter 1.5 I explore how these two models, which have traditionally been presented in the literature as mutually exclusive accounts of the evolution of codon assignments, may in fact both be true to some extent.

Besides time of establishment, the various theories of the origin of the genetic code differ in the degree of adaptationism they invoke. Although it is clear that the genetic code is a complex, evolved mechanism, it is unclear whether the observed codon assignments are themselves historical accidents, specific adaptations, or “spandrels” (Gould and Lewontin 1979) resulting from the laws of chemistry. In this thesis, I evaluate theories of the evolution of the present codon assignments in terms of the assumptions they make both about the availability of precursors and the selective advantage of each step in the development of the translation apparatus. I also present statistical and experimental findings that bear on the relative likelihood of different hypotheses about the genetic code’s origin, evolution, and diversification.

1.1.2 The Genetic Code: History and Experimental Determination

In 1943, Schrödinger proposed that the hereditary material must take the form of an “aperiodic crystal,” a macromolecular structure in which the sites are fixed, but the subunit at each site is free to vary (Schrödinger 1945). Such a system is capable of storing information in proportion to the number of sites it contains, since the decrease in uncertainty (defined as the frequency-weighted sum of the logarithm of the probability of each outcome (Shannon and Weaver 1949)) after determination of each site is additive. By portraying genes as messages in a formal code, Schrödinger set the stage for investigation of the mechanisms by which this code might be translated into the phenotypes of organisms.

Early hypotheses about the relationship between DNA and protein sequences (extensively reviewed in Woese 1967; Ycas 1969; Osawa 1995: see these sources for references) relied on direct templating. Alexander Dounce proposed in 1952 that amino acids paired with specific nucleotides, depending on the nucleotides surrounding them on each side. A year after Watson and Crick’s discovery of the structure of DNA in 1953, George Gamow proposed the “diamond code,” which relied on a key-and-lock mechanism pairing amino acids with diamond-shaped “holes” formed by a base pair plus the next base on each chain. This theory had the advantage that it predicted the occurrence of the 20 proteinaceous amino acids, since if the holes were the same when reflected or when rotated 180° there would be exactly 20 distinct shapes into which amino acid side-chains might fit. However, it was soon discovered that proteins are synthesized in the cytoplasm rather than the nucleus in eukaryotes. This discredited direct templating on DNA as a plausible mechanism for protein synthesis in modern organisms, although it could in principle have played a role in early evolution. To take account the fact that proteins are translated from RNA, not DNA, Gamow proposed the

“triangle code,” which relied on stereochemical fit between amino acids and their codons in mRNA. If the composition, but not the sequence, of each codon were important, there would be exactly 20 classes of codon — one for each amino acid. By a perverse coincidence, the frequency distribution of the codon classes in tobacco mosaic virus RNA matched the frequency distribution of the amino acids in tobacco mosaic virus proteins.

Both the diamond code and the triangle code required overlap between adjacent codons, which in turn implied that only a restricted set of dipeptides could be produced. In 1957, Sydney Brenner showed that the observed diversity of dipeptides in proteins could not be explained by any overlapping code. This observation caused direct templating hypotheses to fall from favor, since they do not allow reading frames to be maintained. In 1957, Crick proposed the adaptor hypothesis, which stated that another molecule (which later turned out to be tRNA) acted as an intermediary between mRNA and amino acid. These adaptors would allow a non-overlapping code (since adjacent adaptors would bind simultaneously to adjacent codons). To explain the maintenance of reading frame, Crick proposed the ingenious “commaless code,” in which codons are assigned such that no out-of-frame trinucleotide can be read as a valid codon. Thus, for instance, if ACG is assigned as a codon then CGA and **GAC** cannot be used anywhere because they could be read out of frame if two ACG codons were adjacent as A**CGAC**G. The maximum size of such a commaless code is, as for the diamond code and the triangle code, 20 — one for each amino acid. As Carl Woese remarked 10 years later, “In retrospect, however, it seems the wildest, almost cruelest joke, that nature would pick for the number of kinds of amino acids in protein the same number that is derived so easily through any of a variety of simple mathematical operations” (Woese 1967).

The codon assignments were finally determined from 1961 to 1966 using Nirenberg and Matthei’s in vitro translation system with synthetic polynucleotides of known sequence. However, considerable progress had already been made by other approaches. In 1961, Crick found strong evidence for a triplet code by examining recombination between different frameshift mutants of T4 bacteriophage. The rIIB cistron was not active in (+) or (–) alone, or in (+ +), (– –), (+ – –) or (+ + –) recombinants, but was active in (+ –), (+ + +), and (– – –) recombinants. Analysis of amino acid replacements under exposure to specific mutagens, such as nitrite (which causes deamination and thus only C→U and A→I transitions), significantly reduced the number of possible codon assignments. Frameshift analysis provided similar data, revealing amino acids for which codons overlapped and differed by only one base. Finally, statistical approaches related frequencies of dinucleotides in DNA to frequencies of amino acids. On the basis of these limited data, Roberts proposed a 4x3x2 code in 1962, suggesting degeneracy of up to four codons per amino acid and predicting many of the correct codon assignments.

1.1.3 Order in the Genetic Code

The “universal” genetic code shows considerable order in the assignment of codons both within and between amino acids (Table 1). Perhaps the most obvious feature of the genetic code is its degeneracy (Sonneborn 1965; Woese 1965; Zuckerkandl and Pauling 1965). All amino acids except Met and Trp are assigned to more than one codon. Furthermore, the codons for a single amino acid are clustered together, rather than being randomly distributed throughout the code. In cases where an amino acid has two codons, those codons are the same in the first two positions and differ only by a transition (a change from one purine to another purine or from one pyrimidine to another pyrimidine; A and G are purines, while U, C and T are pyrimidines) at the third position. In cases where an amino acid has four codons, those codons vary only in the third position. In cases where an amino acid has six codons, these form one four-codon box and one two-codon box.

The degeneracy of the code appears to be controlled by the (G,C) content of codons. Watson-Crick base pairs between C and G involve three hydrogen bonds, while those between A and U or T involve only two (Fig. 1). Thus, GC base pairs are stronger. All codons in which the

doublet (the first two bases) is composed solely of G and C form four-codon boxes, while those in which the doublet is composed solely of A and U form split boxes (either two two-codon boxes or one three-codon box and one one-codon box). This pattern might arise because all-GC doublets bind sufficiently strongly to their cognate anticodons that the third base is irrelevant, while all-AU doublets bind weakly enough to allow discrimination (Lagerkvist 1978; Lagerkvist 1980; Lagerkvist 1981). Mixed doublets form a four-codon box if the second base is a pyrimidine, but form split boxes if the second base is a purine. Presumably, the larger purine at the second position reduces binding at the third position, resulting in finer discrimination (Davydov 1995). (These observations are sufficient to explain Jayaram's finding that if a doublet forms a four-codon box its 'conjugate' forms a split box (Jayaram 1997). The conjugate of a base is the opposite size and forms the opposite number of hydrogen bonds: thus C is the conjugate of A, and G is the conjugate of U.)

The origin of these patterns, and their maintenance, may require separate explanations: even if primordial stereochemical interactions assigned codons in blocks of 2 or 4 depending on GC content, modern tRNAs need not abide by the same constraints. Interestingly, Lagerkvist's rules hold true for almost all variant genetic codes known to date with the exceptions that CUN is split between Ser and Leu in *Candida*, and that the CGN box is sometimes split between arginine and nonsense codons. However, the dinucleotide CG is rare in protein-coding DNA (at least in mammals, which contributed the first few sequences to be determined), and so limited selection pressure against loss-of-function mutations in the CGN box may explain the latter deviation from the general pattern. This consistency may imply that the degeneracy in the code is largely fixed by chemical considerations rather than being a specific adaptation, especially if differences in the strength of codon/anticodon paring between modern tRNAs and mRNAs affect the ribosome's ability to discriminate certain codons efficiently. The fact that the 'wobble' position of tRNAs is sterically configured to promote G-U mispairing may be a contingent, rather than necessary, feature of tRNAs (Szathmáry 1991). If a primordial code based on direct interaction between codons and amino acids were influenced by factors such as GC content, it is possible that these patterns would be reinforced in the evolution of later components of the translation apparatus. On the other hand, it is possible that Lagerkvist's rules arose with the invention of tRNA, and that they partially obscure the primordial codon assignments.

The hydrophobicity of amino acids varies regularly within the code. Five of the most hydrophobic free amino acids — Phe, Leu, Ile, Met, and Val — have U at the second position of their codons; the three most similar amino acids, Leu, Ile, and Val, are connected by single-base mutations at the first position. Six of the most hydrophilic amino acids — His, Gln, Asn, Lys, Asp, and Glu — have A at the second position (Tyr, which is hydrophobic, also has A at the second position, however) (Woese 1965; Woese 1965; Volkenstein 1966; Woese, Dugre et al. 1966; Woese, Dugre et al. 1966). As a result of this, amino acids with complementary anticodons tend to have opposite hydrophobicities (Volkenstein 1966; Blalock and Smith 1984). Amino acids with C in the second position are generally intermediate in hydrophilicity between those with A and U at the second position, while those with G at the second position show no particular pattern. Finally, amino acids that share a doublet always have very similar polar requirements (measured as the ratio of the log relative mobility to the log mole fraction water in a water-pyridine mixture) except perhaps for Cys/Trp (Woese, Dugre et al. 1966a, but see Woese, Dugre et al. 1966b). Principal components analysis on a 20-variable data set later confirmed these general findings, indicating that isoelectric point and other electronic properties might also vary regularly with the second-position base (Sjöström and Wold 1985).

Codons for amino acids with similar chemical properties tend to be highly connected. The two acidic amino acids, Asp and Glu, share their doublet. Their amide derivatives, Asn and Gln, do not share a doublet, but the chemically related pairs are connected by single changes at the first position (GAR Glu↔ CAR Gln; GAY Asp ↔ AAY Asn). The three basic amino acids Lys, Arg and His are connected by single-base mutations (AAR Lys ↔ AGR Arg ↔ CGR Arg ↔ CGY Arg ↔ CAY His). Similarly, the three aromatic amino acids Phe, Tyr and Trp are

connected (UUY Phe ↔ UAY Tyr ↔ CAY; UGG Trp), as are the three hydroxyl-containing amino acids Ser, Thr, and Tyr (UAY Tyr ↔ UCY Ser ↔ UCN Ser ↔ ACN Thr ↔ AGY Ser; CCN, which is also connected, encodes Pro, which is hydroxylated in some proteins). The three stop codons, UAA, UAG, and UGA, are also connected.

It has also been proposed that there is a formal code for general side-chain composition. All O-ended amino acids (Ser, Asp, Glu, Tyr), N-ended (Lys, Arg), and ON-ended (Asn, Gln) amino acids have A-containing doublets. All C-ended amino acids (Val, Ile, Leu, Met, Phe, Trp) have U-containing doublets. The first rule is dominant over the second when side-chains are branched (Davydov 1995; Davydov 1996; Davydov 1998). However, some caution is warranted here: the genetic code is a small, highly connected set, and in some cases it is possible to find equally good "patterns" in randomly generated codes (Amirnovin 1997).

Section 2 deals with the origin of the canonical genetic code, including explanations for its apparently high degree of order.

1.1.4 Variation in the Genetic Code

When the genetic code was found to be the same in humans and *E. coli*, it was natural to assume that it was universal to all organisms. Since any change in the code would be equivalent to introducing mutations throughout the genome, Crick proposed that the codon assignments were a "frozen accident" that became fixed once proteins played crucial roles in metabolism (Crick 1967; Crick 1968). The discovery that, in human mitochondria, the genetic code differs by several codons thus came as rather a surprise (Barrell, Bankier et al. 1979). However, with increasingly detailed comparisons of DNA and protein sequences in diverse taxa, it is clear that the genetic code is still evolving in many lineages. Section III deals with changes in codon usages and codon assignments in modern cells and organelles.

There has also been one artificial change in the genetic code, resulting in the genome-wide substitution of one amino acid for another. In this experiment, Trp-auxotrophic strains of *B. subtilis* were selected on 4-fluorotryptophan medium. Mutant strains were selected that reduced the Trp to 4-fluorotryptophan incorporation ratio by a factor of 2×10^4, and actually grew better in 4-fluorotryptophan than in Trp. This proves that the standard complement of 20 amino acids is not totally inflexible (Wong 1983).

Interestingly, all known changes in the genetic code are recently derived compared to the last common ancestor. In particular, the fact that amino acids are arbitrarily linked to codons by a system of tRNAs and aminoacyl-tRNA synthetase that can themselves evolve means that stereochemical direct-templating effects, if they ever played a role in determining codon assignments, can no longer do so. Thus the processes that led to the establishment of the "universal" genetic code may not be the same as those that led to modern deviations.

Table 1: The "Universal" Genetic Code

	U		C		A		G	
U	UUU	Phe	UCU	Ser	UAU	Tyr	UGU	Cys
	UUC	Phe	UCC	Ser	UAC	Tyr	UGC	Cys
	UUA	Leu	UCA	Ser	UAA	TER	UGA	TER
	UUG	Leu	UCG	Ser	UAG	TER	UGG	Trp
C	CUU	Leu	CCU	Pro	CAU	His	**CGU**	**Arg**
	CUC	Leu	CCC	Pro	CAC	His	**CGC**	**Arg**
	CUA	Leu	CCA	Pro	CAA	Gln	**CGA**	**Arg**
	CUG	Leu	CCG	Pro	CAG	Gln	**CGG**	**Arg**
A	AUU	Ile	ACU	Thr	AAU	Asn	AGU	Ser
	AUC	Ile	ACC	Thr	AAC	Asn	AGC	Ser
	AUA	Ile	ACA	Thr	AAA	Lys	**AGA**	**Arg**
	AUG	Met	ACG	Thr	AAG	Lys	**AGG**	**Arg**
G	GUU	Val	GCU	Ala	GAU	Asp	GGU	Gly
	GUC	Val	GCC	Ala	GAC	Asp	GGC	Gly
	GUA	Val	GCA	Ala	GAA	Glu	GGA	Gly
	GUG	Val	GCG	Ala	GAG	Glu	GGG	Gly

Key

Saturation reflects molecular volume (Grantham 1974). Colourful = bigger.
Brightness reflects polar requirement (Woese et al. 1966). Lighter = hydrophobic.
Hue reflects side-chain composition. 0° (red) = acid; 30° (orange) = amide;
60° (yellow) = sulphur; 120° (green) = alcohol; 180° (cyan) = aromatic;
240° (blue) = basic; 270° (purple) = hydrophobic.

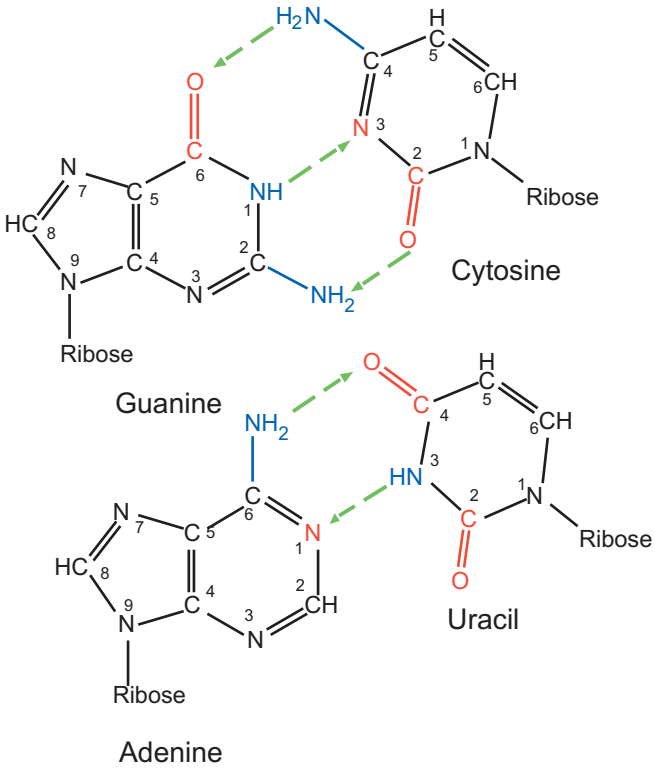


Fig. 1: The four canonical RNA bases, U, C, A, and G, in their Watson-Crick base pairs. T differs from U in that it has a methyl group at C5. Hydrogen bonds (green) point from donors (blue) to acceptors (red).

1.2 Early Metabolism and the Origin of Coded Information

Where did the building blocks for life come from? This chapter outlines ways in which amino acids and nucleotide bases might have been formed on the early earth, and how they might have come together in such a way as to code information. Organic molecules are easy to come by (and have been isolated from extraterrestrial sources), but how they managed to arrange themselves into macromolecules and, ultimately, organisms, is far less clear. Fig. 1 kindly provided by S. J. Freeland.

1.2.1 Prebiotic Synthesis

The first organisms must have relied on abiotic sources of chemicals as sources of both energy and components for replication. The subsequent evolution of metabolism may have been influenced by initial gluts and scarcities of particular compounds. For instance, adenine is produced spontaneously by oligomerization of HCN in dilute solution, and is present in many essential coenzymes such as NADH, FAD, SAM, and ATP (Oró and Kimball 1961; Oró and Kimball 1962). Many important intermediates in purine and pyrimidine metabolism, such as orotic acid and 4-aminoimidazole-5-carboxamide, can be formed in similar reactions, indicating that early organisms may have been selected to synthesize these compounds when prebiotic supplies were depleted (Ferris, Joshi et al. 1978).

For nucleotides, the primary selective pressure was probably template-directed synthesis. Biological nucleic acids have far fewer types of functional groups than amino acids, and so their catalytic range is limited. However, by accelerating spontaneous complementary pairing (A with T or U, and G with C), a single nucleic acid replicase can replicate indefinitely many nucleic acid sequences. Although a self-replicating RNA molecule has not yet been isolated, ribozymes can ligate several bases to themselves in a template-directed fashion (Ekland and Bartel 1996; Bergman, Johnston et al. 2000; Glasner, Yen et al. 2000). Protein replicases, if they are possible at all, would probably have to be unique to each protein sequence. The work of Ghadiri ((Lee, Severin et al. 1997; Severin, Lee et al. 1997), and refs cited therein) on “replicating peptides” supports the view that universal peptide replicases are unlikely, since they catalyze only the last step of their own specific synthesis. These peptide ligases are 32-mers that ligate together a particular electrophilic 17-mer and a particular nucleophilic 15-mer, which together comprise the original peptide ligase sequence. Thus the peptide replicates, given a pure pool of its two already complex subunits. Since the peptide contains glutamine and arginine, and since specific 15-mers and 17-mers would be extremely rare as substrates in a randomly synthesized pool, these peptides do not imply that peptide self-replication is a plausible mechanism for the origin of life.

The natural nucleotides are not the only “solutions” to the complementary pairing problem, however. Non-standard base pairs, such as between the nucleotide analogs κ and χ, can be incorporated with high fidelity by certain DNA and RNA polymerases (Piccirilli, Krauch et al. 1990). Since each of the three positions on the Watson-Crick face of a base can be either a hydrogen bond donor or acceptor, and since the base on a given strand can be either a purine or a pyrimidine, there are potentially sixteen base-pairs that could be formed on the basis of Watson-Crick pairing. Although some of these are unstable due to tautomerization, others must either have been selected against (perhaps due to differences in catalytic activity) or never have been “discovered” by early life forms (Orgel 1990).

Although amino acids are far more diverse than nucleotide bases, they still lack many functional groups. For instance, no translationally incorporated amino acid contains phosphate, sulfate, or sulfonate groups, aldehydes or ketones, nitriles, halides, multiple amines or hydroxyls, etc. (see Fig. 1). However, some of these functional groups are added by posttranslational modification, underscoring the fact that amino acids *can* have these groups,

and that they can be useful in functional proteins. The set of amino acids in proteins only partially overlaps the set found in spark-tube experiments (Miller 1953; Ring, Wolman et al. 1972; Wolman, Haverland et al. 1972; Weber and Miller 1981; Miller 1987) and the Murchison meteorite (Kvenvolden, Lawless et al. 1970; Kvenvolden, Lawless et al. 1971) (Figure 1), indicating that factors other than initial availability must have been important in selecting them. Since Gly, Ala, Asp, Glu and Ser are typically formed in such syntheses, but His, Trp, Met, Arg, Asn and Gln are not, the genetic code may have evolved from a small initial set of prebiotic amino acids by a process of "codon expansion" (Wong and Bronskill 1979). Alternatively, the choice of amino acids may be constrained by chemical stability in the polypeptide chain, which would explain the absence of most of the "missing" functional groups. In an extensive review of the possible monomers for biological catalysts, Weber and Miller argue that about 15 of the 20 amino acids would be identical between independent origins of life (the main exceptions being the puzzling absence of norleucine, norvaline, and α-amino-n-butyric acid) (Weber and Miller 1981).

1.2.2 Theories of the Origin of Metabolism

The relationship between the genetic code and metabolism has rarely been explicitly recognized in research on the two topics. However, both the form and content of the genetic code depend sensitively on the milieu in which it evolved. For instance, theories assuming that the genetic code was determined by chromatographic association between nucleotides and amino acids (see Section II) make the assumptions (a) that all and only the 20 protein-coding amino acids and four RNA nucleotides were present, (b) that association and reaction in a compartment allows the evolution of specific mechanisms for enhancing specific pairing, and (c) that any metabolism in the compartments was sufficiently limited that interactions with macromolecules or metabolic derivatives would not override the association due to pairing. In contrast, theories assuming that the genetic code has been optimized in some respect (see Section II) make the assumptions (a) that replication fidelity with the first genetic code is sufficient that selection between lineages with variant genetic codes is possible, (b) that lineages with other genetic codes existed and were selected against (assuming a large number of alternate, viable codes), and (c) that recognition between codons and amino acids is sufficiently nonspecific that the appropriate adaptor could catalyze any relationship. More generally, late development of the genetic code allows for greater complexity and arbitrariness in the reactions leading to codon-amino acid pairing, and hence an increased role for selective constraints over stereochemical constraints. The following overview of theories of the origin of metabolism will provide context for the various theories of the origin of the genetic code.

Amino acids can be synthesized relatively easily (Miller 1953; Ring, Wolman et al. 1972; Wolman, Haverland et al. 1972; Weber and Miller 1981; Miller 1987) and have been isolated from extraterrestrial sources (Kvenvolden, Lawless et al. 1970; Kvenvolden, Lawless et al. 1971) (Figure 1). In addition, "thermal peptides" produced by heating mixtures of amino acids demonstrate a wide range of weak catalytic activities (reviewed in Fox and Dose 1977). Consequently, the view that an all-protein metabolism evolved prior to a nucleic acid metabolism is attractive. However, it is unclear how proteins could replicate, since amino acid residues lack the self-complementarity that characterizes nucleotide bases. Although it has been suggested that any sufficiently complex mixture of peptides should form an autocatalytic set (Kauffman 1993), the chemical plausibility of this hypothesis remains unsupported. An alternative proteins-first view of metabolism suggests that amino acids and hydroxy acids were activated by sulfur as thioesters, which would spontaneously polymerize into heterogeneous macromolecules (de Duve 1995). The primary advantage of this scheme is that the thioester bond contains sufficient energy to catalyze ATP formation, allowing later evolution of nucleic acids. However, it shares the difficulty that template-directed replication is impossible.

Nucleic acid-first models of metabolism have the advantage that replication is simple. However, the image of life evolving as a "naked replicator" (Dawkins 1976), later "clothing"

itself in metabolism, is powerful but chemically implausible. In particular, ribose is not a plausible prebiotic molecule due to the nonspecificity of the formose reaction and to its high rate of decomposition; also, pyrimidines do not efficiently bind to ribose under prebiotic conditions (Joyce 1989; Schwartz and de Graaf 1993; Larralde, Robertson et al. 1995) (reviewed in last of these). Furthermore, the polyphosphates suggested as a source of phosphate probably did not exist (Keefe and Miller 1995). Peptide-nucleic acid (PNA), a stable nucleic acid analog that substitutes an uncharged peptide backbone for the ribose-phosphate backbone of DNA and RNA, has received considerable attention as a possible substitute (Egholm, Buchardt et al. 1993; Wittung, Nielsen et al. 1994). However, no prebiotic synthesis of PNA has been demonstrated, its backbone is uncharged, and it lacks the hydroxyl groups that many ribozymes use for catalysis. Thus, it is difficult to see how PNA could interact with simple metabolites in a prebiotic setting. Another possibility is that early nucleic acids relied on a sugar other than ribose for the backbone (Joyce, Schwartz et al. 1987; Schwartz 1997). In one variant of this, it is suggested that the original sugar was glycerol and that the original bases were all purines (including two purines with the H-bonding patterns of U and C, which must subsequently have disappeared from metabolism), due to the relative ease of synthesis of these components (Wächtershäuser 1988). Again, no plausible prebiotic syntheses have been demonstrated.

Another suggestion is that the first genetic information was carried by clay (Cairns-Smith 1982). In this theory, lattice defects in clays such as kaolinite might differentially catalyze particular chemical reactions (giving them phenotypes), and could be inherited by direct surface templating (giving them genotypes). In particular, clay-directed synthesis of particular nucleic acid sequences would allow information transfer between the clay and nucleic acid genetic systems, allowing the latter to “take over” control of metabolism after the development of RNA and/or protein. Again, however, this theory is without experimental support.

Interestingly, the actual components of intermediary metabolism may be a large subset of the easily synthesized components. Morowitz et al. searched Beilstein Online, a large index of chemical compounds, and found that a simple set of physical and chemical constraints picked out 153 molecules from the 3.5 million entries, including all 11 members of the reductive citric acid cycle (Morowitz, Kostelnik et al. 2000). However, some caution is required in interpreting these results, since the rules were derived with full knowledge of the citric acid cycle intermediates, and some of them are rather arbitrary (compounds containing only C, H, and O; special ratios of compositional ranges). Additionally, the compounds in the database are likely to be biased towards those with direct biological significance (Orgel 2000).

At present, there is a gap in understanding the events that led from prebiotically synthesized monomers and oligomers to metabolic systems capable of catalyzing their own replication. However, several things seem fairly clear. First, primordial life forms probably did not have any macromolecular information storage system, as do modern organisms. Instead, they probably relied on the “limited replication” (Szathmáry and Maynard Smith 1995; Szathmáry and Maynard Smith 1997) provided by alternative metabolic hypercycles. Second, primordial life forms probably did not use RNA, although this has no bearing on the likelihood of the “RNA World”. Instead, early metabolism probably relied on surface-catalyzed interconversion of small molecules, aided by any macromolecules able to condense under those conditions (Wächtershäuser 1990). Finally, the “crystallization” of information out of metabolism and into stable genetic resources is an open problem. Pinpointing the timing of the origin of the genetic code may help decide whether RNA was a late addition to a protein metabolism or proteins were a late addition to an RNA metabolism.

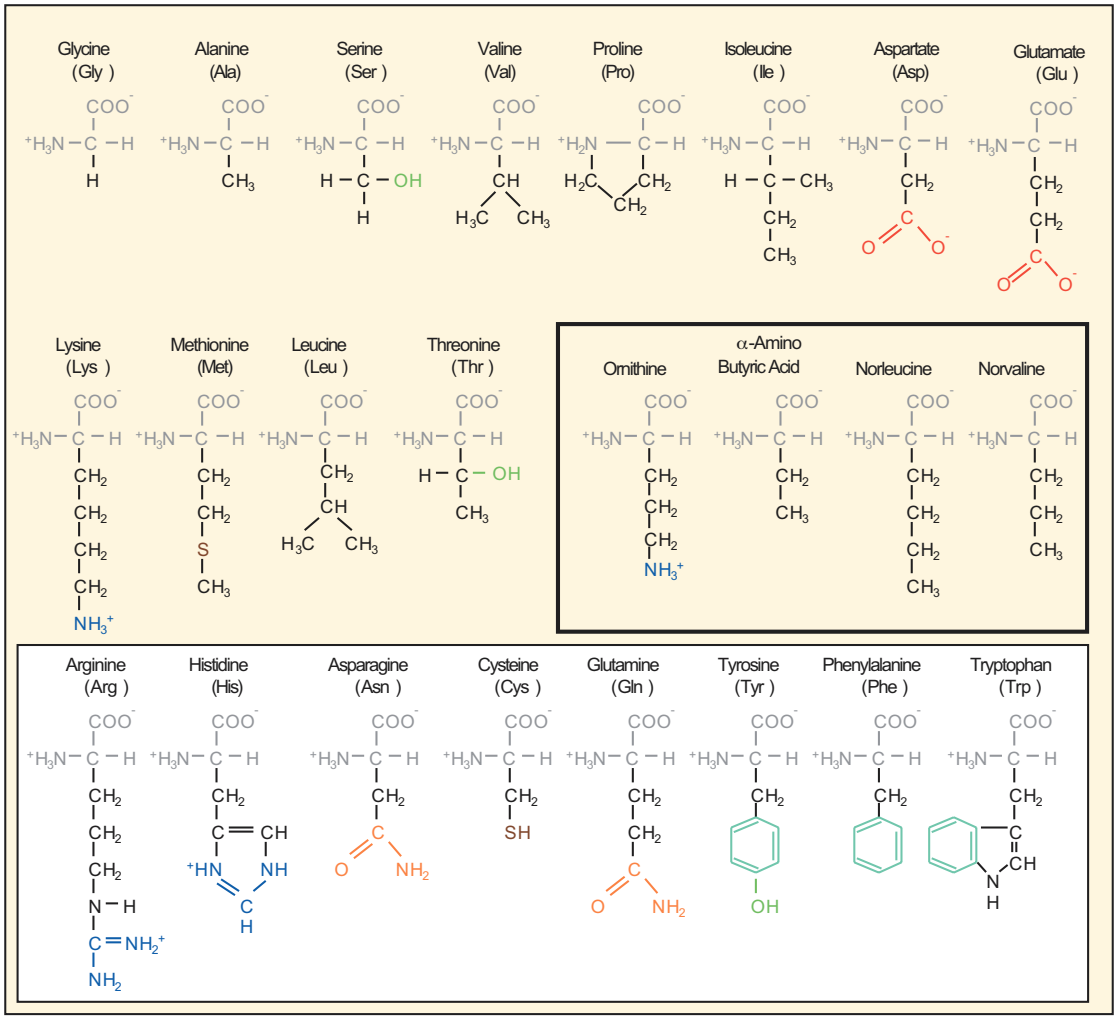


Figure 1: Amino acid structures; beige shading indicates those found either in the products of pre-biotic simulation ('spark experiments': see Miller 1987 and references therein) or in the Murchison meteorite (Kvenvolden, Lawless et al. 1971). Boxed amino acids are a sub-set of those found in the Murchison satellite which are *not* found within the genetic code. *In vitro* selection can determine which RNA triplets associate most strongly with each amino acid, and may reveal why certain prebiotic amino acids are absent from the code.

1.3 The RNA World

Modern protein synthesis suffers from a 'Chicken or Egg?' problem: proteins are essential for life, yet protein synthesis itself requires many proteins. One way around this is to suppose that proteins were a late invention in metabolism, and that RNA (perhaps with small-molecule cofactors) originally performed all necessary catalysis. This chapter covers the RNA world hypothesis, which has been the paradigm for early evolution since the amazing discoveries by Cech and Altman in 1982 showing that RNA could act as a catalyst as well as a message. The chapter also suggests several ways to test whether the genetic code arose in such a milieu.

1.3.1 Overview

The "RNA world" (Gilbert 1986), a hypothetical period in evolution during which RNA preceded both DNA and protein, has become the favored explanation for the origin of metabolism. Its strong form, in which "One can contemplate an RNA world, containing only RNA molecules that serve to catalyze the synthesis of themselves," (Gilbert 1986) is probably false due the difficulties in synthesizing ribose noted above. Its weak form, in which "…after a period of chemical evolution, with or without further augmentation by a genetic system completely unrelated to RNA, a genetic system evolved based on some simple RNA-like molecule," (Joyce 1989) is supported by a wide range of evidence.

The main virtue of the RNA world is that it resolves the following contradiction: DNA requires protein catalysts for its replication, yet proteins cannot themselves replicate. RNA, however, can perform both functions. Some viruses have RNA genomes, which demonstrates that RNA can act as a genetic material. The discovery that RNA could act as a catalyst demonstrated that proteins were not the only biological macromolecules that could effect specific reactions. The explanatory power of RNA catalysis is so compelling that it was proposed long before there was any evidence for it, solely on the basis of structure and convenience (Woese 1967; Crick 1968; Orgel 1968). The concept crystallized in 1986, with a series of reviews postulating the existence of "ribo-organisms," organisms that used RNA as their sole catalyst (Alberts 1986; Cech 1986; Gilbert 1986; Lewin 1986; Orgel 1986; Lazcano and Miller 1996).

One intriguing suggestion is that, rather than an RNA self-replicase, the primordial metabolism consisted of a hypercycle of coevolving RNA molecules (Eigen 1971; Eigen and Schuster 1979; Eigen, Gardiner et al. 1981). According to this view, hereditary information was originally carried as a dynamic set of transformations between molecules rather than in a single molecule, as in present genomes. A hypercyclic metabolism is one in which each molecule catalyzes the formation of at least one other molecule in the system: the system as a whole thus forms an autocatalytic set, and replicates together. In fact, the plant mitochondrial genome may replicate as a hypercycle of various linear DNA fragments that interconvert by replication (Albert, Godelle et al. 1996). Models of hypercyclic systems (Dyson 1985; Kauffman 1993) indicate that it may be easier for proteins than for RNA to form such a metabolism. However, hypercyclic metabolism provides an alternative to the supposition that life began with an RNA molecule that could catalyze its own replication.

1.3.2 Translation Apparatus Structure and Function

The first evidence that RNA had a role beyond messenger was the discovery of tRNA, which "looks like Nature's attempt to make RNA do the job of a protein" (Crick 1968). Similarly, rRNA makes up more than 60% of the mass of the ribosome (Lake 1985), and even confers its peptidyltransferase activity (Nissen, Hansen et al. 2000). This exemplifies the use of RNA in the type of structural role that proteins typically assume. Since it is unlikely that relatively inefficient RNA solutions would evolve in a protein world, it is likely that these functional RNA molecules are holdovers from an RNA world, especially where they interact with so many

other enzymes that it would be difficult to change them. Thus, the primitive translation apparatus was probably made entirely of RNA (Orgel 1968), preceding protein entirely.

One interesting feature of tRNA is that it contains many modified bases (reviewed in Osawa 1995, Knight, Landweber et al. 2001), including a conserved change from adenine to inosine whenever adenine is at the first position of the anticodon. It is possible that these modifications are relics of an initial widespread modification system that expanded the catalytic range of RNA by adding functional groups, similar to contemporary posttranslational modification of proteins. If so, (a) ribozymes selected from an expanded set of bases should be capable of a wider range of catalytic tasks (Piccirilli, Krauch et al. 1990), and (b) it should be possible to select catalytic RNAs (ribozymes) that perform each of the modifications that are preserved back to the last common ancestor of extant life.

Considerable evidence suggests that the anticodon loop and acceptor stem of tRNA (Fig. 1) have separate evolutionary histories. This finding is directly relevant to the origin of the genetic code, since it implies that the ability of tRNA to be aminoacylated is separate from its ability to specifically pair with codons in mRNA. Thus there should be no necessary interrelation between the two (Maizels and Weiner 1987). The first line of evidence comes from sequence analysis, which showed that the 3' and 5' halves of tRNA are roughly symmetrical (Eigen and Winkler-Oswatitsch 1981). Thus, tRNA may have been the product of duplication of an existing RNA gene. Second, tRNA can prime reverse transcription of a variety of retroviruses and retrotransposons, indicating that tRNA-like structures may have originally been selected for their role in replication (reviewed in Maizels and Weiner 1987). Third, the top half of modern tRNA is a unit that is separately recognized by a numerous enzymes including RNase P, elongation factor Tu, and tRNA synthetases (reviewed in Maizels and Weiner 1994). Furthermore, the bottom half of tRNA (containing the anticodon loop) interacts only with 16S rRNA, while the top half of tRNA (containing the acceptor stem) interacts only with the 23S rRNA (Noller 1993). Thus, the two halves of tRNA interact independently with the two subunits of ribosomes. Fourth, tRNA genes are often split by introns towards the 5' end. These may be molecular fossils from a time when the two halves were separate entities (Dick and Schamel 1995). Fifth, ATP(CTP):tRNA nucleotidyltransferase, which adds the terminal CCA to mature tRNAs, will accept the top half alone as a substrate (Shi, Weiner et al. 1998). Finally, very short minihelices containing the terminal CCA are able to accept amino acids in the presence of tRNA-aminoacyl synthetases (reviewed in Schimmel, Giege et al. 1993; Schimmel 1995), or even in the presence of specific Asp-containing dipeptides (Shimizu 1995), indicating that the aminoacylation activity can be separated from the rest of the tRNA. Perhaps amino acids were employed as coenzymes for ribozymes before the evolution of coded protein synthesis (Szathmáry 1993; Szathmáry and Maynard Smith 1997).

The recent discovery that the catalytic molecule in the ribosome is the rRNA (Cech 2000; Nissen, Hansen et al. 2000) provides the most compelling evidence that catalytic RNA molecules really did shape the evolution of modern metabolism. The crystal structure of ribosomes from *Haloarcula marismortui* with peptidyl-tRNA substrate analogs shows contacts only with two bases, G2482 and A2486, in the large subunit rRNA. The nearest protein side-chain atom is more than 18 Angstroms distant, too far to be involved in the catalysis. Thus, as proposed eight years earlier (Noller, Hoffarth et al. 1992), the ribosome itself is a ribozyme. Like tRNAs, rRNAs contain many modified bases, some of them dating back to the last common ancestor of extant life (Cermakian and Cedergren 1998). The recent synthesis of modified purines under prebiotic conditions (Levy and Miller 1999) suggests that noncanonical bases may have been significant in ribozyme catalysis since its inception.

1.3.3 Other Evidence for the RNA World

RNA appears to be primary to DNA in metabolism, since deoxythymidine monophosphate is formed by 5' methylation of deoxyuridine monophosphate, deoxyribonucleotide diphosphates are formed by reduction of ribonucleotide diphosphates, and DNA polymerase extends the 3'

end of an RNA primer (reviewed by Joyce 1989). Histidine, one of the most important amino acids in catalysis (and, interestingly, not produced by prebiotic syntheses), also appears to be an RNA derivative. Its biosynthesis begins with condensation between ATP and phopshoribosylpyrophosphate, removes most of the ATP as 5-aminoimidazole-4-carboxamide ribonucleotide, and forms a new imidazole moiety by condensation with an amino group from glutamine that is integrated into protein as an amino acid side-chain (reviewed in Lamond and Gibson 1990; Voet and Voet 1995). Many essential coenzymes, such as NAD, FAD, FMN, CoA, and SAM, also contain nucleotides or nucleotide derivatives (Joyce 1989), despite the fact that simpler chemical equivalents would work equally well (Benner, Allemann et al. 1987). Since it is unlikely that simple coenzymes would unnecessarily evolve nucleotide moieties, these may well be molecular fossils from a time when most reactions were catalyzed by RNA (Woese 1967; Crick 1968; Orgel 1968; White 1976; Visser and Kellog 1978).

The most important discovery supporting the RNA world, however, was the fact that RNA could enhance reactions with enzymatic specificity. The first such system to be discovered was the self-splicing group I intron in an rRNA gene in the ciliate *Tetrahymena thermophila* (Kruger, Grabowski et al. 1982). This intron excises itself from the primary transcript by recruiting guanosine as a nucleophile, cleaving itself at the 5' end, and joining the surrounding exon sequences. Group I introns are found in various eukaryotic mitochondria, nuclei, and chloroplasts, in bacteriophage T4, and in eubacteria, and interrupt mRNA, rRNA and tRNA genes (reviewed in Cech 1993). The widespread distribution of the group I intron might be taken to imply that it dates back to the last common ancestor, but its distribution is more likely due to horizontal transmission (Sogin, Ingold et al. 1986). The group I intron is not a true catalyst, however: because it evolved to act in *cis*, the reaction occurs only once per molecule. Moreover, the ribozyme is changed by the reaction, so cannot catalyze the reaction indefinitely (as would a true enzyme).

The demonstration that RNA could act as a true catalyst came almost immediately after the discovery of the group I intron. RNase P, a ribonuclease that cleaves a short 5' leader sequence during tRNA processing, has both protein and RNA components. In certain eubacterial RNase P enzymes, the RNA component alone can catalyze the reaction if the salt concentration is sufficiently high (Guerrier-Takada, Gardiner et al. 1983). The RNA subunit of RNase P from vertebrate and fungi nuclei and mitochondria cannot act alone, even though it is essential for activity (Brown and Pace 1992), indicating that protein may have taken over more of the catalytic role in these taxa than in eubacteria. A variety of other ribozymes have subsequently been discovered *in vivo*, including group II introns in eukaryotic organelles and eubacteria, hammerhead ribozymes in viroids and newt satellite DNA, hairpin ribozymes in the satellite RNA of tobacco ringspot virus, and unique ribozymes in hepatitis delta virus and in *Neurospora* VS RNA (reviewed in Landweber, Simon et al. 1998).

1.3.4 The RNA World and In Vitro Selection

The main problem faced by RNA world scenarios is explaining how RNA catalysts can sustain a complex metabolism (Benner, Ellington et al. 1989; Di Giulio 1997). *In vitro* selection, a technique in which RNA (or DNA) molecules that satisfy certain criteria are amplified from a large randomized pool (Fig. 2) (Joyce 1989; Ellington and Szostak 1990; Tuerk and Gold 1990), demonstrates that RNA can perform tasks other than those exhibited in extant organisms. Although no trace remains of the presumptive RNA catalysts that performed such tasks as nucleotide and amino acid biosynthesis, replication, etc., if they existed then it should be possible to reproduce them in the laboratory. The existence of a ribozyme that catalyzes a specific reaction does not imply that ribo-organisms actually used the particular ribozyme selected in the relevant experiment. However, it does demonstrate that it is possible for RNA alone to carry out the metabolic task (Szostak and Ellington 1993; Hirao and Ellington 1995).

The power of *in vitro* selection comes from the fact that it is an iterative enrichment procedure. In a combinatorial library of 10^{13} –10^{16} sequences, only a few molecules might be able to

catalyze an arbitrary reaction. However, by coupling each sequence's ability to replicate with its ability to perform a catalytic task, repeated rounds of selection can amplify even very rare sequences to fixation. The selection step can be any process that isolates molecules that perform a desired task. For instance, affinity chromatography selects molecules by the extent to which a target ligand retards their progress through a column or gel, and is widely used to select aptamers (nucleic acid molecules that bind a specific target, but do not catalyze a reaction). With ribozyme selections, more cunning schemes are possible: for example, a ligase might be assayed for its ability to ligate itself to a target sequence bound to avidin beads, which could then be recovered from the bulk solution. After the selection step, the "surviving" sequences are amplified by PCR, and the resulting enriched pool enters the next round of selection. Even if the amplification is only 1000-fold at each step, after 5 rounds of selection the overall amplification and enrichment is 10^{15}-fold. Mutagenic PCR might increase the effective number of possible sequences still further by introducing new diversity into an already selected pool, making it possible to search larger parts of sequence-space.

Ribozymes have now been selected to catalyze activities as diverse as sulfur alkylation, RNA and DNA cleavage and ligation, isomerization of a bridged biphenyl, self-biotinylation, carbon-carbon bond formation, and porphyrin metalation (reviewed in Landweber, Simon et al. 1998). These ribozymes demonstrate many of the functions that would be necessary to sustain metabolism in an RNA world. Several activities particularly relevant to the origin of the genetic code have been isolated. The first of these is amide bond cleavage, constructed by inserting an amide bond at the cleavage site of the DNA target of an existing ribozyme (Dai, Mesmaeker et al. 1995). The second is self-aminoacylation with phenylalanine, which demonstrates that RNA can act as an aminoacyl-tRNA synthetase (Illangasekare, Sanchez et al. 1995), and can even be faster and more stereoselective than protein PheRS (Illangasekare and Yarus 1999). Aminoacylation can proceed with as few as 29 nucleotides, making this a plausible primitive RNA reaction (Illangasekare and Yarus 1999). A second, independently derived aminoacylator can charge wild-type tRNA in trans (Lee, Bessho et al. 2000). The third is aminoacyltransferase activity, in which the selected RNA transferred N-biotinylated methionine from a fragment of the natural tRNA to itself (Lohse and Szostak 1996). Finally, ribozymes can catalyze both amide and peptide bond formation (Wiegand, Janssen et al. 1997; Zhang and Cech 1997). Taken together, these results show that RNA can catalyze all of the reactions required for coded peptide synthesis. The final step, which may be possible in the near future, is to make an artificial translation system entirely out of RNA.

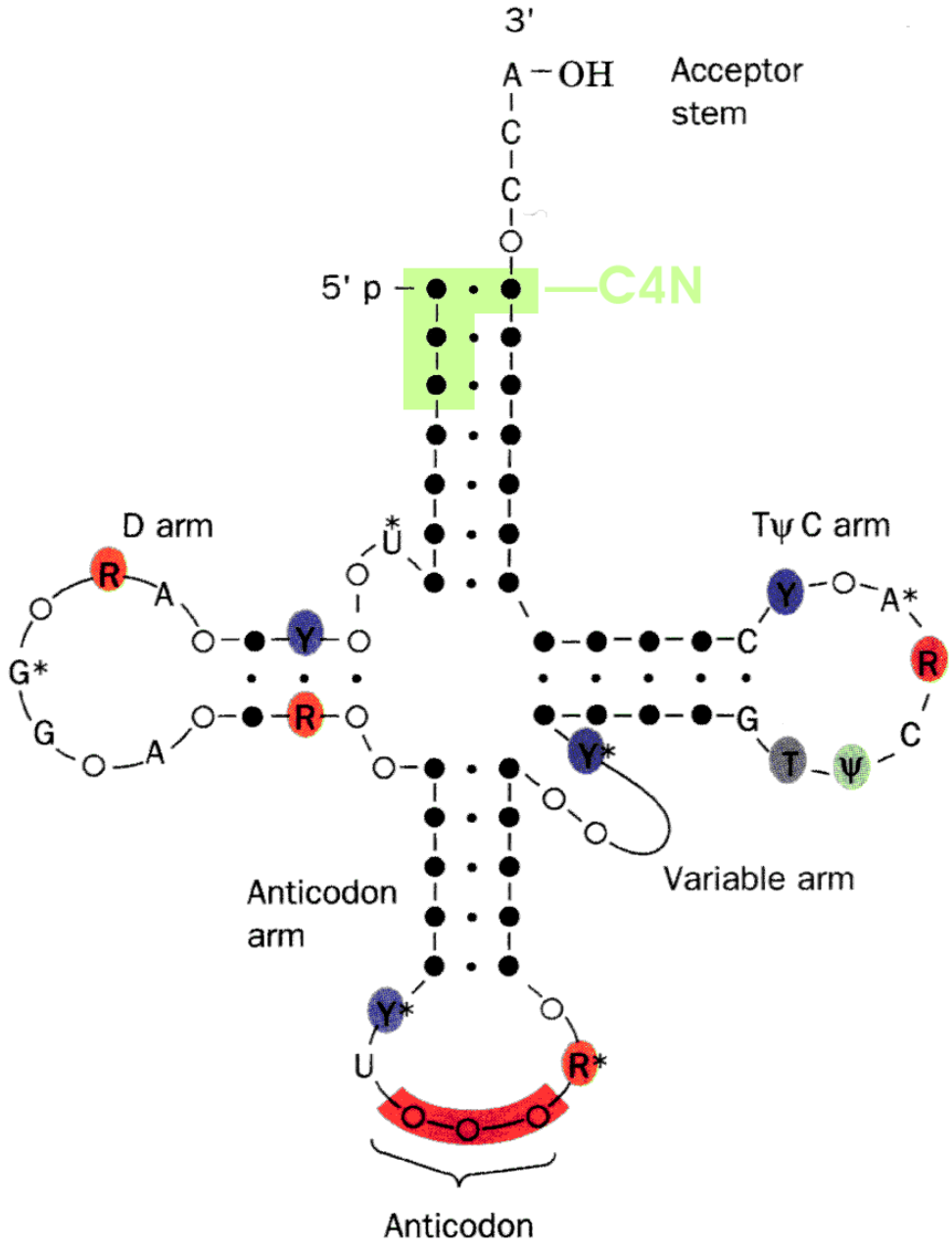


Fig. 1. Secondary structure of tRNA. Filled circles are bases involved in Watson-Crick pairing. Dashed regions indicate variable number of bases. C4N indicates Shimizu's "Complex of Four Nucleotides" (Shimizu 1982). Adapted from Voet and Voet (1995).

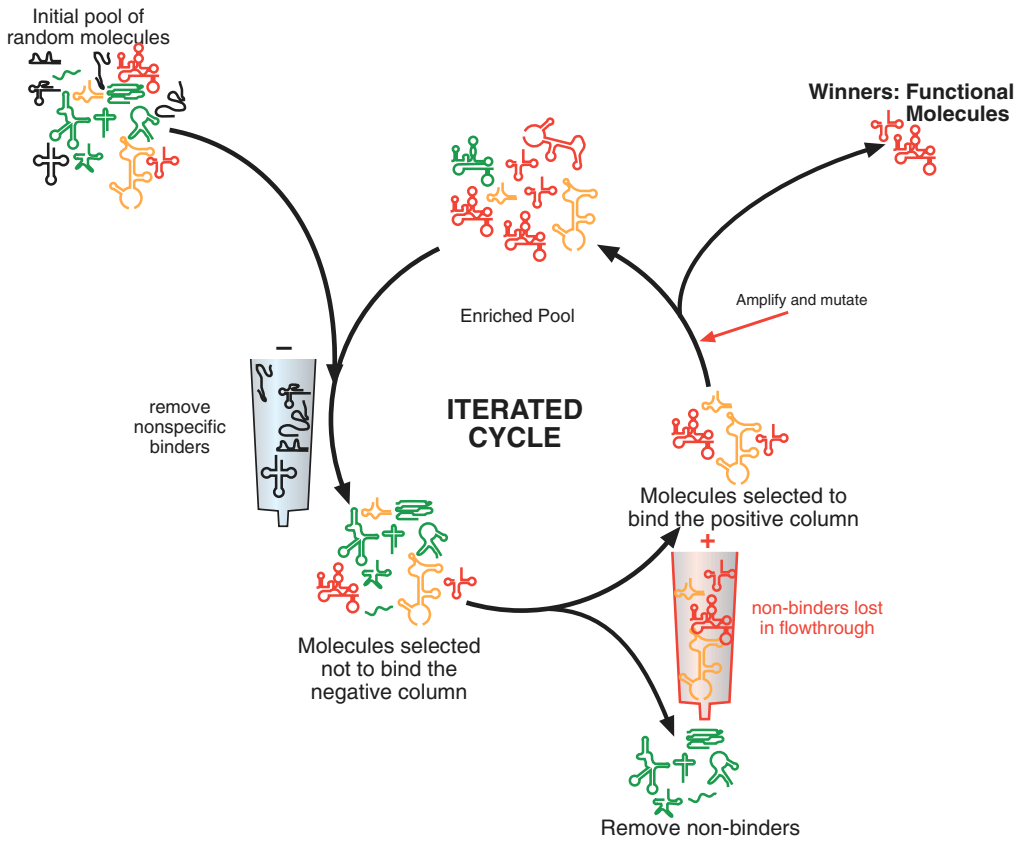


Fig. 2 Overview of In Vitro Selection (SELEX).

1.4 The Early Evolution of the Genetic Code

This chapter distills the information in the two previous chapters, together with some recent findings about several components of the translation apparatus, to argue that the code did in fact evolve in an RNA world. First, I summarize the evidence for direct interaction between codons and binding sites, explored in much more detail in Section 2. Then I highlight recent ribozyme selections for aminoacyl-tRNA synthetases made entirely out of RNA, and suggest that the phylogeny of protein aminoacyl-tRNA synthetases may yield the order in which later amino acids were added to the code after proteins took over from ribozymes. Finally, I summarize the ways in which modern code evolution might differ from ancient code evolution, giving as an example the change of the release factor eRF1 in ciliates to prevent recognition of UAR as a termination codon. Unfortunately, the change in specificity did not turn out to be as simple as the single amino acid alteration highlighted here, but further investigation provided interesting results (see Chapter 3.3). This chapter appeared in Cell as:

Knight, R. D. and L. F. Landweber (2000). "The Early Evolution of the Genetic Code." Cell 101: 569-572.

Cell, Vol. 101, 569–572, June 9, 2000, Copyright ©2000 by Cell Press

The Early Evolution of the Genetic Code

Minireview

Robin D. Knight* and Laura F. Landweber*
Department of Ecology and Evolutionary Biology
Princeton University
Princeton, New Jersey 08544

Evolutionary inferences rely on diversity. The source of differences among organisms is accumulated divergence from a common ancestor, which may be random or selected. When a system is adaptive yet highly complex, one can follow its evolution from a simpler state in one of two ways: from fossilized transitional forms, or from early-diverging extant organisms. This is how, for example, we can trace the evolution of trichromatic vision in primates or flowers in angiosperms.

The problem becomes harder when no intermediate states exist. In particular, hypotheses about evolution prior to the Last Universal Common Ancestor of extant life (LUCA) defy standard techniques. Biochemical pathways do not fossilize, precluding direct inferences about ancestral states, and by definition no lineages diverging before the LUCA survive. Thus the diversity of extant life reveals little about general principles, since biochemical necessities mingle with quirks inherited from the shared ancestor. Consequently, it is difficult to explain why highly conserved and universal systems such as the translation apparatus are the way they are.

Early Evolution of the Code: Extraordinary Techniques for Extraordinary Problems

In the absence of evidence, many of the most interesting questions about the genetic code have fallen into a twilight zone of speculation and controversy. Although it is generally accepted that the modern code evolved from a simpler form, there has been no consensus about when the initial code evolved or what it was like, how and when particular amino acids were added, how and when the modern tRNA/synthetase system arose, or the processes by which the code could have expanded. Now, detailed study of the components of the translation apparatus is at last making these questions tractable.

Three general approaches have recently yielded surprising intimations about how the genetic code evolved. The first is to appeal to general principles at a primary level, in this case the chemistry of nucleic acids and amino acids, to infer how a translation system might be constrained. The second is to alter parts of the translation apparatus in vitro in ways that might reflect earlier states, showing what changes are possible. The third is to examine the phylogeny of particular components, revealing how they have changed since the LUCA (or, in the case of paralogous genes, even before the LUCA), and to extrapolate backward from the principles thus revealed. Here we show how key applications of these approaches begin to provide a general framework for understanding the origin and development of the code.

*E-mail: rdknight@princeton.edu (R. D. K.), lfl@princeton.edu (L. F. L.)

Amino Acid/Nucleotide Interactions in the RNA World and Earlier

Because RNA is unstable and difficult to synthesize, the first genetic material may have used a simpler backbone than ribose. One candidate is peptide nucleic acid (PNA), in which the backbone is polymeric N-(2-aminoethyl)glycine (AEG) and the N-acetic acids of the bases (N_9 for purines, N_1 for pyrimidines) are linked via amide bonds (Figure 1). This is an attractive scenario because AEG forms in spark-tube experiments that also produce amino acids (Nelson et al., 2000), and may spontaneously polymerize at 100°. The N-acetic acids of the bases are also accessible in prebiotic syntheses, which suggests that PNA could have been an early genetic material (although the evidence is far from conclusive).

The prebiotic plausibility of PNA implies that amino acids and a genetic system based on purines and pyrimidines could have been coproduced and then coevolved on the early earth. Does the genetic code in modern organisms reflect such ancient interactions, or have all traces been erased by subsequent evolution of the translation apparatus?

One can approach this question statistically, asking whether chemistry has influenced codon assignments. SELEX, the selective amplification of nucleic acid molecules that perform particular tasks, can identify specific RNA sequences that bind amino acids (Connell et al.,

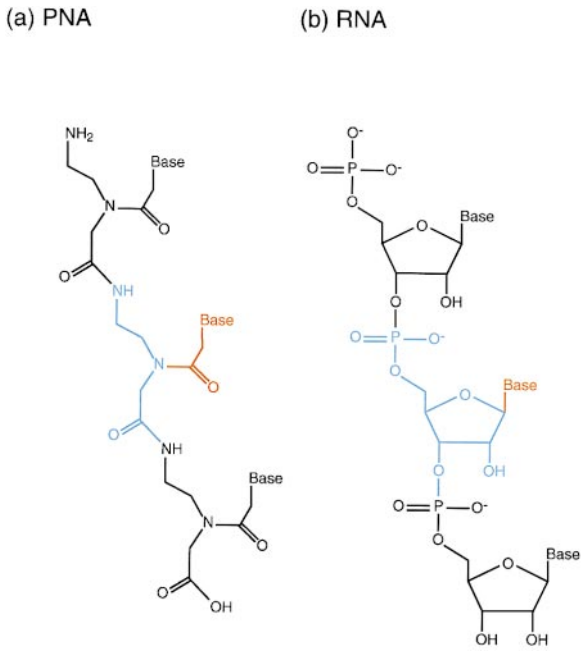


Figure 1. PNA or RNA First?

PNA (a) has a peptide backbone instead of the sugar-phosphate backbone of RNA (b). Unlike ribose, N-(2-aminoethyl)-glycine is formed at high yields under prebiotic conditions and spontaneously produces a stable polymer. However, its uncharged rigid backbone may limit possibilities for catalysis. One backbone monomer is highlighted in blue, with the informational unit highlighted in red.

1993). Thus, one can test whether codons that specify a particular amino acid in the canonical genetic code occur disproportionately often at RNA sites that bind it. For instance, we found that arginine binding sites are predominantly composed of Arg codons, even in aptamers selected in different labs using different protocols (Knight and Landweber, 1998).

The recent isolation of aptamers to tyrosine, which bear more Tyr codons than expected at their binding sites (Mannironi et al., 2000), prompted a debate in the April issue of *RNA* over the robustness and interpretation of the statistical evidence for such associations. Yarus extends this analysis to other amino acids for which aptamers are now available (arginine, isoleucine, and tyrosine), and concludes that the overall probability that the observed codon/binding site association would occur by chance is 3.3×10^{-7} (Yarus, 2000). Ellington et al. raise methodological concerns, showing that the choice of statistical techniques and sequences can affect the level of significance of the association (Ellington et al., 2000). We test the robustness of the result by examining all possible combinations of sequences for binding site associations with each codon set (Knight and Landweber, 2000), and show that Arg codons alone significantly associate with arginine binding sites.

Taken together, these papers show that the amino acid aptamers that have been structurally characterized do overrepresent their cognate codons at their binding sites. Although as Ellington et al. point out there are grounds for caution (the structurally characterized aptamers could be a nonrepresentative sample, and the relationship does not hold independently for each of the four nucleotides), we can tentatively conclude that amino acid binding sites are preferentially enriched in certain trinucleotides, which correspond curiously to modern codon assignments. Even DNA aptamers for arginine showed significant codon/binding site associations (Knight and Landweber, 2000), indicating that the backbone is not critical; this is consistent with the idea that an alternative backbone, such as PNA, might have been the original genetic molecule, and may suggest that elements of the modern genetic code predate the RNA world.

If the genetic code evolved from a simpler form, stereochemistry may have produced an initial code that later expanded. The modern code appears highly optimized for resistance to various types of error (Freeland et al., 2000), which complicates the situation. How could stereochemical codon assignments reflect the same properties that affect substitution rates of amino acids within proteins? Clearly, more aptamer data are needed to establish the generality of codon/binding site associations, and to assess the relative roles of chemistry and selection in shaping the earliest genetic codes.

The Origins of Aminoacyl-tRNA Synthesis

Although protein enzymes catalyze tRNA aminoacylation today, they cannot have existed before protein synthesis itself. It is widely accepted that ribozymes predated proteins, and several labs have recently isolated ribozymes with peptidyltransferase activity. This shows that specific peptide synthesis could have arisen in an RNA world. Two ribozymes of particular interest come from the Yarus lab and the Szostak lab.

Illangasekare and Yarus selected a self-aminoacylating ribozyme using Phe-AMP as a substrate (Illangasekare and Yarus, 1999a). One 95-mer from this pool was highly specific for Phe, accelerating the reaction 6×10^7-fold over background and preferring Phe-AMP 10^4-fold over other aminoacyladenylates. This compares favorably to yeast PheRS on both counts, indicating that RNA can catalyze aminoacylation at least as well as do proteins. A 29 nt aminoacylating RNA was later constructed (Illangasekare and Yarus, 1999b). Although this tiny ribozyme is not specific for any amino acid, it can catalyze peptide bond formation as well, suggesting that both these reactions may have been easily accessible to RNA world metabolisms.

Lee et al. selected a self-aminoacylating ribozyme using two different substrates: first, a hexanucleotide complementary to a 3′ guide sequence and derivatized with Phe-biotin, and second, cyanomethyl-activated glutamine (Lee et al., 2000). This produced 170 nt "ambidextrous" ribozymes with two independent active domains. Because the transfer from a 5′-OH to a 3′-OH is energetically neutral, a ribozyme that catalyzes transfer from the 3′-OH of another molecule to its own 5′-OH should also perform the reverse reaction if its 5′-OH is already aminoacylated. Thus, when provided with Gln-CME, the ambidextrous ribozymes aminoacylated tRNA molecules that bound the guide sequence. Although these ribozymes are not as fast or selective as that isolated in the Yarus lab, they can specifically aminoacylate tRNA in *trans*, as do modern synthetases.

Because several aminoacylation specificities (Lys, Gly, possibly Tyr/Trp) appear to have evolved several times in independent lineages, it may also be relatively easy for proteins to evolve aminoacyl-tRNA synthetase (aaRS) activity. Chihade and Schimmel attempted to reconstruct a primitive aaRS by linking a minimal aminoacyladenylate-forming domain of Ala-RS to a nonspecific RNA binding domain (Chihade and Schimmel, 1999). Although the resulting protein could aminoacylate a tRNA-Ala-derived microhelix at rates comparable to an aaRS that permits cell growth in yeast, this construct was still large: over 600 amino acids long. Thus, protein synthesis was presumably highly developed by the time protein aaRSs began to replace ribozymes.

Aminoacyl-tRNA Synthetases and the Expanding Code

Were all 20 amino acids in our genetic code present in the RNA world? If so, ribozymes must catalyze a tremendous range of reactions; alternatively, the RNA world might have relied on the few prebiotically available amino acids. Although the aaRSs within each class are related to each other, and hence arose by duplication and divergence of two original synthetases, these duplications could reflect either addition of new amino acids to the code or takeover of existing amino acids from ribozyme synthetases. Although SELEX may suggest an ancient stereochemically determined code, the intermediate transitions are unclear.

New amino acids may have been initially synthesized from metabolic precursors by tRNA-dependent processes, with synthetases capable of directly charging them to tRNAs evolving only later. The new synthetase would capture some of the tRNAs, and hence some of the codons, of its ancestor, assigning metabolically

Figure 2. Atavistic GlnRS

GlnRS (Q) → GluRS (E) from *E. coli* and *H. sapiens*, shown with an example of GluRS from each of the three domains. Residues conserved across both specificities are highlighted in green; those conserved only across GlnRS are highlighted in blue, and those conserved but differing between GlnRS and GluRS are highlighted in yellow. The changes that remove Gln specificity are marked in red. Note that different residues changed in the two cases: this may be because the *E. coli* experiment selected against efficient mischargers, since mischarging of wild-type $tRNA^{Gln}$ with Glu inhibited protein synthesis.

related amino acids to adjacent codons (Wong, 1981). This type of code expansion requires that aaRSs acquire new specificities. Although suppressor mutants are typically altered tRNAs, and never aaRSs, two recent studies show that aaRSs can be engineered to retrace their evolutionary history.

Although all organisms have a dedicated GluRS, GlnRS appears to have arisen as a paralog of GluRS in eukaryotes, with subsequent lateral transfer to a few other lineages. Most bacteria and archaea use GluRS to charge tRNA-Gln with glutamate, and then convert it to glutamine on the tRNA by a transamidase. Agou et al. analyzed a structure-based alignment of GluRS and GlnRS from different taxa, and identified two residues invariant in all GlnRS but absent from GluRS (Agou et al., 1998). Altering these residues to match eukaryotic GluRS reduced selectivity for Gln more than 10,000-fold.

This rational mutagenesis approach is limited to testing the effects of a few mutations. Hong et al. instead randomized sections of GlnRS and selected the variants best conferring GluRS specificity in vivo, using *E. coli* GlnRS as a starting point (Hong et al., 1998). Two changes, though interestingly different from the ones noted in Agou et al., improved Glu recognition 3- to 5-fold (Figure 2). This GlxRS was inefficient, probably because it mischarged wild-type $tRNA^{Gln}$ with Glu. Combining both approaches by mutating and selecting an orthogonal tRNA/synthetase pair that does not affect the components already in the cell, such as human GlnRS and tRNA-Gln in *E. coli*, might allow a more complete identity switch.

To add amino acids to the code, the original aaRS must relinquish some of its isoacceptor tRNAs to its new paralog. Li et al. take the first steps toward achieving this process experimentally in an insertion mutant of *E. coli*, LeuRS, which prefers one tRNA-Leu isoacceptor 3-fold over another instead of charging both at equal rates (Li et al., 1999). In nature, a similar process has taken place in *Thermus thermophilus*, which has two independent pathways for tRNA asparaginylation (Becker et al., 2000). The first is direct formation of Asn-$tRNA^{Asn}$ by an archaeal-type AsnRS; the second is indirect formation, first producing Asp-$tRNA^{Asn}$ by a eubacterial-type AspRS and then transamidating this aminoacyl-tRNA to Asn-$tRNA^{Asn}$. *T. thermophilus* also has an archaeal-type AspRS, which recognizes and aspartylates only tRNA-Asp, in contrast to the eubacterial AspRS, which recognizes and aspartylates tRNA-Asn as well. Clearly, AspRS has lost the ability to recognize a subset of its tRNA substrates in lineages that have an independent AsnRS. This may indicate that Asn was a relatively recent addition to the code, perhaps postdating the origin of most aaRSs.

Recent Code Evolution: Release Factors and Modified Bases

Thus far we have covered processes that led to the code in the LUCA but did not contribute to its subsequent diversification. Recent variant codes are predominantly changes in a few tRNAs and release factors. Examples of the former are numerous, and are often changes in RNA editing or base modification at the anticodon rather than mutations in the anticodons of tRNA genes themselves. For example, Met is encoded by AUG alone in the standard code, but by AUA and AUG in metazoan mitochondria. $tRNA^{Met}$ normally has anticodon CAU: a mutation to UAU would allow recognition of both A and G at the third codon position by wobble pairing. This would seem the easiest way to effect this change, as UNN anticodons commonly read NNR 2-codon sets. However, *Drosophila*, bovine, and squid $tRNA^{Met}$ instead retain the CAU anticodon sequence but modify the C to 5-formylcytidine, which recognizes both A and G

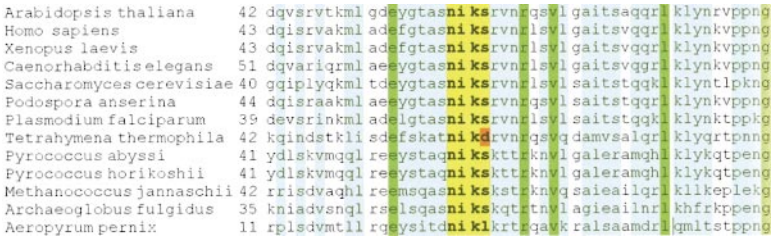


Figure 3. Comparison of eRF1 Homologs from Different Taxa

Release factors are highly conserved compared to aaRSs (see Figure 2 for comparison). Highly conserved residues (>50% identity) are blue, and absolutely conserved residues are green. The NIKS motif, yellow, is involved in stop codon recognition and is conserved except in *Tetrahymena* (nonconservative Ser → Asp, red). Interestingly, of the species shown only *Tetrahymena* uses a noncanonical set of termination codons.

(Tomita et al., 1999). Changes in base modification may indeed be a widespread mechanism of producing alternative genetic codes.

Stop codons are the most labile, changing independently in many lineages. This mutability may reflect their rarity—occurring only once per reading frame—or the ease of losing or altering release factors. The sequence of the release factor eRF1 from *Tetrahymena* (Karamyshev et al., 1999), which uses UAA and UAG for Gln instead of stop, may illuminate this question, since the crystal structure of human eRF1 was recently solved (Song et al., 2000) and several homologs are available from other eukaryotes and archaea. Thus, we can form a specific hypothesis about the molecular basis for this change. The NIKS domain is universally conserved and involved in codon recognition, and mutations immediately adjacent to it produce a univeral suppressor. However, in *Tetrahymena*, the Ser undergoes a nonconservative mutation to Asp, which may generate the new specificity (Figure 3). Examination of other ciliate, diplomonad, and algal lineages, with parallel changes in termination, will indicate whether this residue is universally important in stop codon recognition.

Conclusions

Together, research into different components of the translation apparatus is beginning to paint a consistent picture of how the genetic code might have evolved. The primordial code, influenced by direct interactions between bases and amino acids probably dates back to the RNA world or earlier. The invention of tRNAs and ribozyme-based aaRSs made this mapping indirect, allowing swapping of amino acids between codons and hence a level of optimization. Additionally, the code probably underwent a process of expansion from relatively few amino acids to the modern complement of 20. By the time protein aaRSs took over, translation was probably well developed; however, some amino acids, such as Gln, Asn, and Trp, may postdate the first protein aaRSs. Today, laboratory experiments that alter the specificity of aaRSs for amino acids and/or tRNA isoacceptors recapitulate some of these processes. Finally, changes to both tRNAs and release factors produced the range of modern codes, particularly through post-transcriptional base modification and changes in release factors. This diversity of events suggests that an explanation for the fixation of the canonical code in the LUCA will require more historical reconstruction than reasoning from chemical principles.

Selected Reading

Agou, F., Quevillon, S., Kerjan, P., and Mirande, M. (1998). Biochemistry *37*, 11309–11314.

Becker, H.D., Roy, H., Moulinier, L., Mazauric, M.H., Keith, G., and Kern, D. (2000). Biochemistry *39*, 3216–3230.

Chihade, J.W., and Schimmel, P. (1999). Proc. Natl. Acad. Sci. USA *96*, 12316–12321.

Connell, G.J., Illangsekare, M., and Yarus, M. (1993). Biochemistry *32*, 5497–5502.

Ellington, A.D., Khrapov, M., and Shaw, C.A. (2000). RNA *6*, 485–498.

Freeland, S.J., Knight, R.D., Landweber, L.F., and Hurst, L.D. (2000). Mol. Biol. Evol. *17*, 511–518.

Hong, K.W., Ibba, M., and Soll, D. (1998). FEBS Lett. *434*, 149–154.

Illangasekare, M., and Yarus, M. (1999a). Proc. Natl. Acad. Sci. USA *96*, 5470–5475.

Illangasekare, M., and Yarus, M. (1999b). RNA *5*, 1482–1489.

Karamyshev, A.L., Ito, K., and Nakamura, Y. (1999). FEBS Lett. *457*, 483–488.

Knight, R.D., and Landweber, L.F. (1998). Chem. Biol. *5*, R215–R220.

Knight, R.D., and Landweber, L.F. (2000). RNA *6*, 499–510.

Lee, N., Bessho, Y., Wei, K., Szostak, J.W., and Suga, H. (2000). Nat. Struct. Biol. *7*, 28–33.

Li, T., Li, Y., Guo, N., Wang, E., and Wang, Y. (1999). Biochemistry *38*, 9084–9088.

Mannironi, C., Scerch, C., Fruscoloni, P., and Tocchini-Valentini, G.P. (2000). RNA *6*, 520–527.

Nelson, K.E., Levy, M., and Miller, S.L. (2000). Proc. Natl. Acad. Sci. USA *97*, 3868–3871.

Song, H., Mugnier, P., Das, A.K., Webb, H.M., Evans, D.R., Tuite, M.F., Hemmings, B.A., and Barford, D. (2000). Cell *100*, 311–321.

Tomita, K., Ueda, T., Ishiwa, S., Crain, P.F., McCloskey, J.A., and Watanabe, K. (1999). Nucleic Acids Res. *27*, 4291–4297.

Wong, J.T.-F. (1981). Trends Biochem. Sci. *6*, 33–36.

Yarus, M. (2000). RNA *6*, 475–484.

1.5 Selection, Chemistry, and History: Three Faces of the Genetic Code

This chapter, although written earlier than the previous chapter, builds on it by elaborating the three specific hypotheses that have been developed to explain the codon assignments in the canonical genetic code. Crick's (1968) famous 'Frozen Accident' model of code evolution argued that there was no good evidence for adaptation or stereochemistry in shaping the modern code structure, and that the difficulty of changing the code would cause even a random structure to become fixed once sufficiently useful. We argue that the frozen accident is no longer a tenable view, by evaluating recent findings supporting the idea that (a) there are traces of the pathways of the code's evolution still visible in its structure; (b) the code has been selected over the possible alternatives because it minimizes genetic errors; and (c) direct chemical interactions shaped at least some codon assignments. This chapter provides the general framework for Section 2, in which these hypotheses are explored in far greater detail.

The chapter was published in Trends in Biochemical Sciences *in 1999, as:*

Knight, R. D., S. J. Freeland and L. F. Landweber (1999). "Selection, history and chemistry: the three faces of the genetic code." Trends Biochem Sci 24(6): 241-7.

Most of this material came directly from my thesis proposal, which I prepared as part of the requirements for the General Examination in mid-1998. Dr. Freeland contributed several of the figures, the suggestion that the code might have adapted by choosing amino acids that fit both chemical and selective constraints, and the distinction between 'engineering' and 'statistical' approaches for measuring code optimality, and suggested many other useful changes to the manuscript.

Selection, history and chemistry: the three faces of the genetic code

Robin D. Knight, Stephen J. Freeland and Laura F. Landweber

The genetic code might be a historical accident that was fixed in the last common ancestor of modern organisms. 'Adaptive', 'historical' and 'chemical' arguments, however, challenge such a 'frozen accident' model. These arguments propose that the current code is somehow optimal, reflects the expansion of a more primitive code to include more amino acids, or is a consequence of direct chemical interactions between RNA and amino acids, respectively. Such models are not mutually exclusive, however. They can be reconciled by an evolutionary model whereby stereochemical interactions shaped the initial code, which subsequently expanded through biosynthetic modification of encoded amino acids and, finally, was optimized through codon reassignment. Alternatively, all three forces might have acted in concert to assign the 20 'natural' amino acids to their present positions in the genetic code.

THE GENETIC CODE remains an enigma, even though the full codon catalog was deciphered over 30 years ago. Although we know which base triplets encode which amino acids, and even how these assignments vary among taxa, we do not know why the specific codon assignments take their actual form[1]. Why, for instance, does the AUU triplet encode isoleucine rather than some other amino acid? Why do some amino acids have more codons than others? And why do amino acids that have similar chemical properties tend to have similar codons (Fig. 1)?

The simplest answer is that codon assignments were historical accidents that became fixed in the last common ancestor of all modern organisms and that, therefore, their pattern requires no further explanation[2]. This 'frozen accident' hypothesis is a useful null model against which other models can be tested, but does not predict the observed order in the genetic code. The model has also been criticized because we now know that the code is not universal, and thus variant codes might have existed before the last common ancestor, as well as at present.

There are three main challenges to the frozen-accident model, which are based on 'adaptive', 'historical' and 'chemical' arguments. All three deal only with the genetic code present in the last universal ancestor and might not apply to more-recent changes. The 'adaptive' challenge suggests that the pattern of codon assignments is an adaptation that optimizes some function, such as minimization of errors caused by mutation or mistranslation. The 'historical' challenge suggests that the genetic code accumulated amino acids over a long period of time and that codon assignments reflect this pattern of incremental expansion. The 'chemical' challenge suggests that certain codon assignments were directly influenced by favorable chemical interactions between particular amino acids and short nucleic acid sequences, whereas lack of such interactions excluded other amino acids from proteins entirely. Here, we evaluate the evidence for these three views and suggest how they might be combined into a coherent synthesis of code evolution.

R. D. Knight, **S. J. Freeland** and **L. F. Landweber** are at the Dept of Ecology and Evolutionary Biology, Guyot Hall, Princeton University, Princeton, NJ 08544-1003, USA.
Email: lfl@princeton.edu

Adaptation – the best of all possible codes?

The earliest explanations for the observed order in the genetic code, such as Crick's ingenious commaless code[3], assumed that natural selection somehow optimized the codon catalog. Given that more changes to a protein are deleterious than beneficial, the genetic code should reduce the impact of errors: the pattern of degeneracy, which groups together codons for the same amino acid, certainly has this effect (Fig. 1). The 'lethal mutation' model[4] proposed that the genetic code reduces the effects of point mutation, whereas the 'translation error' model[5] proposed that the code structure instead reduces the effects of errors during translation.

The principal evidence that supported these early models came from inspection of the genetic code itself: (1) codons for the same amino acid typically vary only at the third position; (2) amino acids that have U at the second position of their codon are hydrophobic, whereas those that have A at the second position are hydrophilic; and (3) the genetic code initially appeared to be universal[5]. This evidence is neither compelling nor unequivocal. Crick's wobble hypothesis[6] explained much of the degeneracy of the code in terms of simple chemical considerations: a single tRNA anticodon can recognize multiple codons by nonstandard base pairing. The association between second-position base and amino acid hydrophobicity holds only for two of the four bases

0968–0004/99/$ – See front matter PII: S0968-0004(99)01392-4

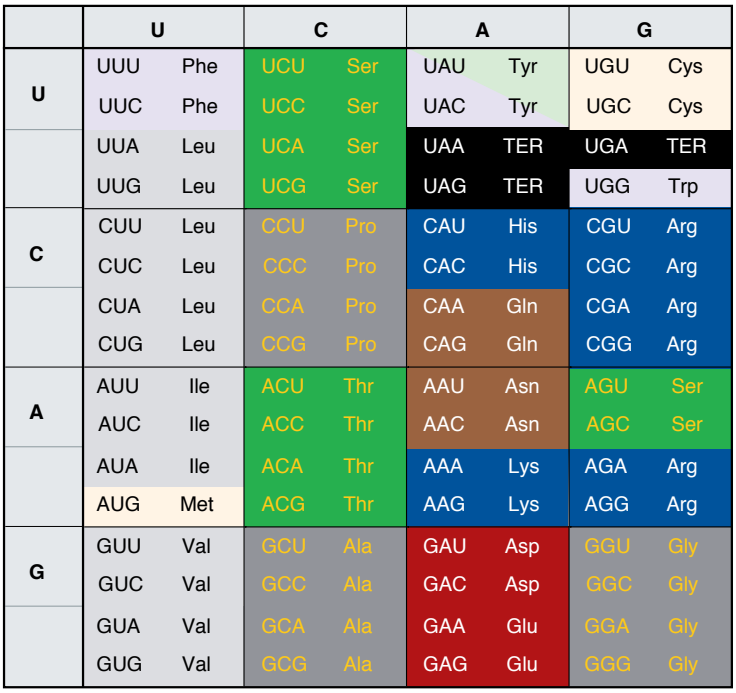


Figure 1

The 'universal' genetic code. Shading indicates polar requirement (PR)[1]: lighter shades (black text), PR < 6 (hydrophobic); medium shades (yellow text), PR = 6–8 (medium); darker shades (white text) PR > 8 (hydrophilic). Amino acids whose codons have U at the second position tend to be unusually hydrophobic; those whose codons have A at the second position tend to be hydrophilic. Amino acids that share structural similarity tend to share codon sets connected by single point mutations: for instance, the basic amino acids arginine, lysine and histidine are connected. Ter, termination codon.

(Fig. 1). Finally, if code optimization had actually occurred, then the present genetic code must have been selected from a large pool of alternative genetic codes (a problem when the code was thought to be absolutely invariant). These shortcomings, given the choice of the frozen-accident theory as an alternative, probably account for the decline of adaptive explanations towards the end of the 1960s.

A variety of criteria have been used to assess whether the genetic code is in some sense optimal. These analyses fall into two main classes: 'statistical' and 'engineering'. The statistical approaches[7–11] compare the natural code with many randomly generated alternative codes and typically have concluded that the genetic code conserves amino acid properties far better than would a random code. In contrast, the engineering approaches[12–16] compare the natural code with only the best possible alternative (i.e. the code that formally minimizes the change in amino acid properties following an average single point mutation), and conclude that the genetic code is still far from optimal.

The statistical approach provides a more realistic representation of the variability available to selection than does the engineering approach. Because the engineering approach measures optimality on a linear scale as a fraction of the distance between the mean and optimal codes, it ignores the distribution of possible codes. This distribution is roughly Gaussian: increasingly optimal codes are increasingly rare, and the difference between successively more optimal codes decreases as optimality increases. Consequently, the globally optimal code might be unattainable, whereas the most optimal code accessible by point mutations is still closer to optimal than almost all alternatives. In fact, our unpublished results indicate that the canonical genetic code is closer to optimal than practically all alternatives, and this conclusion holds for differences in both measurement of optimality and distribution of possible codes. However, the evolutionary plasticity of the code might have been limited by unknown chemical or historical constraints.

The principal objection to optimization theories has been that a change in the genetic code causes mutations in every protein, most of which are likely to be deleterious. Consequently, once cells relied on a particular genetic code to any appreciable extent, the further changes required by the optimization process would have become increasingly unlikely[2]. The ability of the genetic code to change is a prerequisite for theories that involve optimization through a stepwise evolutionary process. The discovery that the genetic code is not invariant[17] removed this objection: if the genetic code recently has changed in apparently nonadaptive ways, then similar changes might have facilitated adaptation in the past. Actual changes in the nuclear genomes of eukaryotes (Fig. 2a) indicate that, even in metabolically complex organisms, the code is far from frozen.

Two mechanisms account for the codon swapping evident in a variety of species, and in both nuclear and mitochondrial genomes (Fig. 2). In the Osawa–Jukes mechanism[18], particular codons vanish from the genome because of mutational pressure on the genome for changes in A.T or G.C composition, and the corresponding tRNAs are lost. When the mutational pressure later reverses, codons that lack cognate tRNAs inhibit translation. Consequently, any mutation that allows translation of these codons is advantageous. Such a mutation can occur through duplication of an existing tRNA gene and subsequent mutation of the anticodon to recognize a different codon. If the mutated tRNA still retains its original aminoacyl-tRNA synthetase specificity, the codon will encode an amino acid that differs from that used by the canonical code.

The Schultz–Yarus mechanism[19] is similar but does not require the complete disappearance of a codon from the

genome before the transfer takes place. Instead, a mutation in a duplicated tRNA that generates either a new anticodon or a new aminoacyl-charging specificity leads to ambiguous translation of one or more codons. If this new specificity confers an advantage, selection will fix the new codon set. The fact that certain *Candida* species have ambiguous translation – depending on the circumstances, CUG will encode either serine or leucine – supports the model[20].

History – searching for footprints of the code's ancestors

Historical theories propose that the present code evolved from a simpler ancestral form: proteins produced by the initial, limited, set of amino acids synthesized new amino acids that could in turn be incorporated into the code. Recently introduced amino acids presumably would take over codons from their metabolic precursors; this could happen only if the resulting changes in protein structure were not widely deleterious[2]. Consequently, historical theories often predict that similar amino acids would be assigned to similar codons even without explicit selection for error minimization.

The principal evidence for coevolution of amino acids and the code through stepwise expansion comes from cases in which dissimilar amino acids from related biosynthetic pathways also share similar codons (Fig. 3). Several authors argue that a disproportionate number of biosynthetically related amino acids have codons connected by single point mutations[14,16,21,22]; however, because many amino acids are interconvertible, even randomized codes show similar associations between biosynthetically related amino acids and single base changes in codons[23].

One intriguing suggestion is that the first- and second-position bases have different functions: the second-position bases connect amino acids that have similar properties; and the first-position bases connect amino acids from the same biosynthetic pathway[24]. Codons of the form GNN correspond to amino acids thought to be most primitive for several reasons[24]; this might suggest that UNN, CNN and ANN codons were transferred to novel amino acids as their synthesis became possible. This hypothesis constrains the set of possible codes considerably, but does not explain the near optimality of the code[11].

Another approach looks at the phylogenies of tRNAs and of aminoacyl-tRNA synthetases (the enzymes that specifically link amino acids to their cognate tRNAs). If amino acids were added sequentially to the code, then tRNA and aminoacyl-tRNA synthetase phylogenies should be congruent; this would reflect duplication and divergence of a tRNA and its cognate synthetase as each amino acid was added. Unfortunately, most studies that examined tRNA phylogenies[25–27] have assumed that trees derived from the set of tRNAs in different species are congruent, which is not the case[28]. Because tRNAs can change either their anticodons or their amino acid specificity remarkably easily[29], modern tRNA phylogenies are unlikely to reveal anything about the phylogeny of tRNAs in the last common ancestor. Furthermore, tRNA phylogenies are likely to become increasingly unstable as more sequences are added: this apparent tRNA flexibility is consistent with the requirement of the adaptive theories that the code be able to change.

Phylogenies of aminoacyl-tRNA synthetases prove slightly more revealing. Aminoacyl-tRNA synthetases fall into two main classes. Some of those for related amino acids cluster together[30], and phylogenies are similar among widely separated taxa[31]. Interestingly,

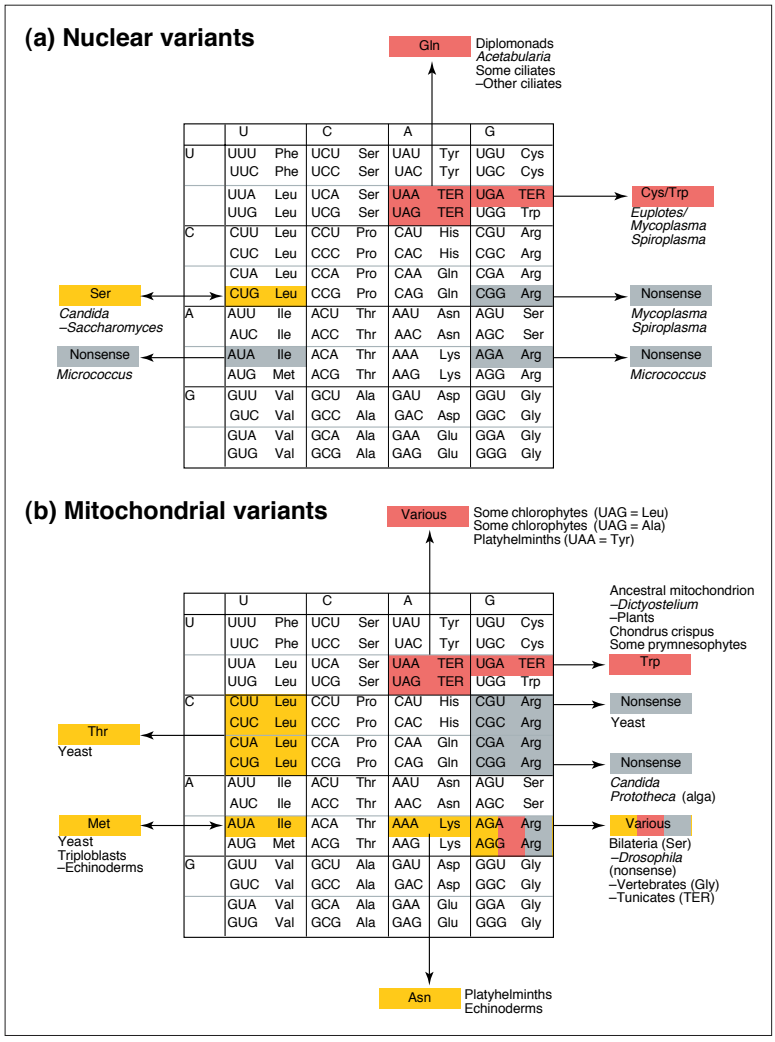


Figure 2

Naturally occurring variants of the canonical genetic code. **(a)** Nuclear variants (including changes effective within bacterial genomes)[34,48,49]. **(b)** Mitochondrial variants[48,50,51] (yeast variants are from http://www.ncbi.nlm.nih.gov/htbin-post/Taxonomy/wprintgc?mode=c). Missense changes are shown in yellow; nonsense changes are shown in gray; changes in termination codons are shown in red. '–' indicates a reversal of a change in a particular lineage.

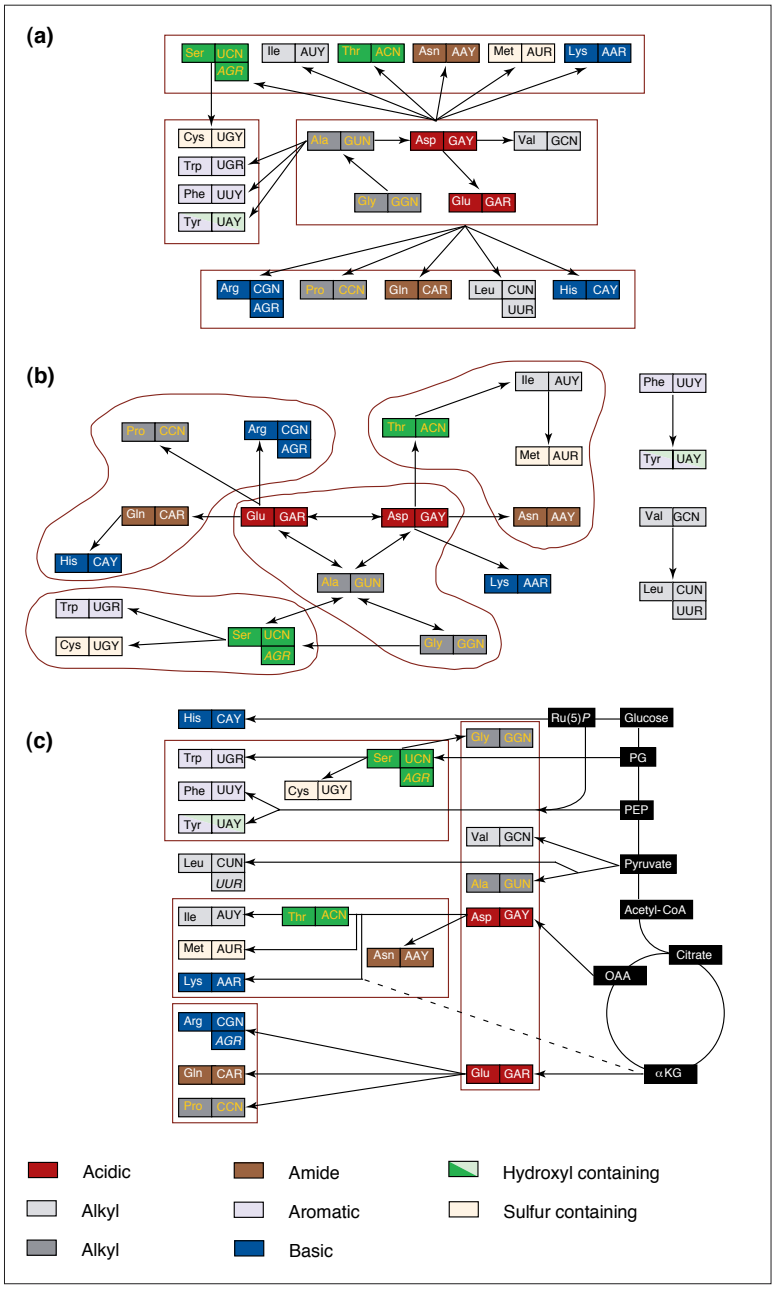


Figure 3

Biosynthetic pathways and code assignments. **(a)** Primitive sulfur-metabolizing bacteria (hypothetical)[47]. **(b)** Generalized prokaryotes[21]. **(c)** *Escherichia coli*[24]. Shading indicates polar requirement (PR)[1]: lighter shades (black text), PR < 6 (hydrophobic); medium shades (yellow text), PR = 6–8 (medium); darker shades (white text) PR > 8 (hydrophilic). Bounded areas highlight codons that share the same first base identity. αKG, α-ketoglutarate; OAA, oxaloacetic acid; PEP, phosphoenolpyruvate; PG, phosphoglycerate; Ru(5)*P*, ribulose 5-phosphate.

although most organisms have a class II lysyl-tRNA synthetase, some archaea and spirochetes have a class I lysyl-tRNA synthetase[32]. Given that the class I lysyl-tRNA synthetases are monophyletic and cluster within the other type I synthetases[31], the last common ancestor of all organisms probably contained both types of synthetase, and all lineages probably lost one or the other at a later stage[33]. However, because the complete set of tRNA synthetases and tRNAs was present in the last common ancestor, phylogenetic analysis alone cannot discriminate between stepwise introduction of amino acids into translation and stepwise takeover of aminoacylation by protein aminoacyl-tRNA synthetases from more-primitive catalysts. Although congruence between tRNA and synthetase phylogenies would have provided striking evidence for sequential amino acid incorporation, the lack of such congruence provides evidence against expansion of the code during synthetase evolution. The present synthetases might have usurped the roles of earlier ribozymes that had the same functions, erasing the information in the original synthetases about the order in which amino acids were added to the code.

Stereochemistry – does it fit the evidence?

Stereochemical theories propose that amino acids are assigned to particular codons because of direct chemical interactions between RNA and amino acids. If these interactions follow consistent patterns, similar amino acids should bind to similar short RNA motifs and should therefore have similar codons. Although the resulting pattern of codon assignments might be adaptive, relative to randomized codes (because a point mutation would tend to substitute a relatively similar amino acid), it need not have been explicitly selected for this effect. Thus, the rules that constrain the set of chemically plausible codes might also lead to apparent error minimization.

The fact that the genetic code initially appeared to be universal provided the strongest support for stereochemical theories, because it suggested that the actual code is the only possible code. However, the known variations in the code do not disprove the stereochemical theories. All deviations from the canonical code appeared recently in comparison with the last common ancestor: the first surviving change probably appeared in the lineage leading to diplomonads[34], and most are much more recent. Furthermore, no known code differs by more than a few amino acids from the standard code. Because translation pairs codons with amino acids through a tRNA adaptor, the mechanisms that allowed recent changes in the genetic code might be entirely different from those that generated the code initially. All stereochemical theories have dealt only with the canonical code found in the last common ancestor,

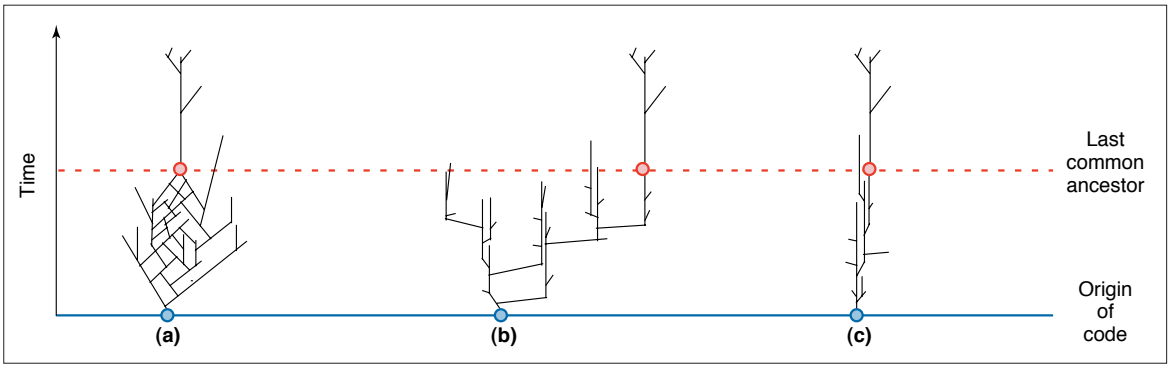


Figure 4

Three models of early code evolution. The 'universal' genetic code found in the last common ancestor (pink circle) might or might not be similar to the first genetic code that evolved (blue circle). **(a)** The primordial genetic code is maintained by lineage merging in a reticulate network: there is little competition between lineages, and lineages that share the majority genetic code have the advantage of using novel proteins from other lineages when protocells merge. **(b)** Strong selection for increased code efficiency among lineages drives the code in the last common ancestor far from the primordial code. Most lineages with variant codes become extinct, but a few successfully reach new local optima. **(c)** Despite competition among lineages, the chemical factors leading to the establishment of the original genetic code are much the same as the factors that influence the error in a given amino acid substitution; therefore the final code remains similar to the initial code. Aptamer experiments can distinguish (b) from (a) and (c) by providing evidence for a primordial code that might or might not be similar to the code in the last common ancestor.

because later changes probably were unaffected by stereochemical constraints.

The first stereochemical theories about the origin of the code relied on chemical models. These provided weak support for a variety of possible pairing mechanisms: amino acids might bind to their cognate codons[35], anticodons[36], reversed codons[37], codon–anticodon double helices[38] or a complex of four nucleotides containing the anticodon at the end of the acceptor stem[39]. Unfortunately, the diversity of results reduces their significance: the apparent freedom inherent in the building and

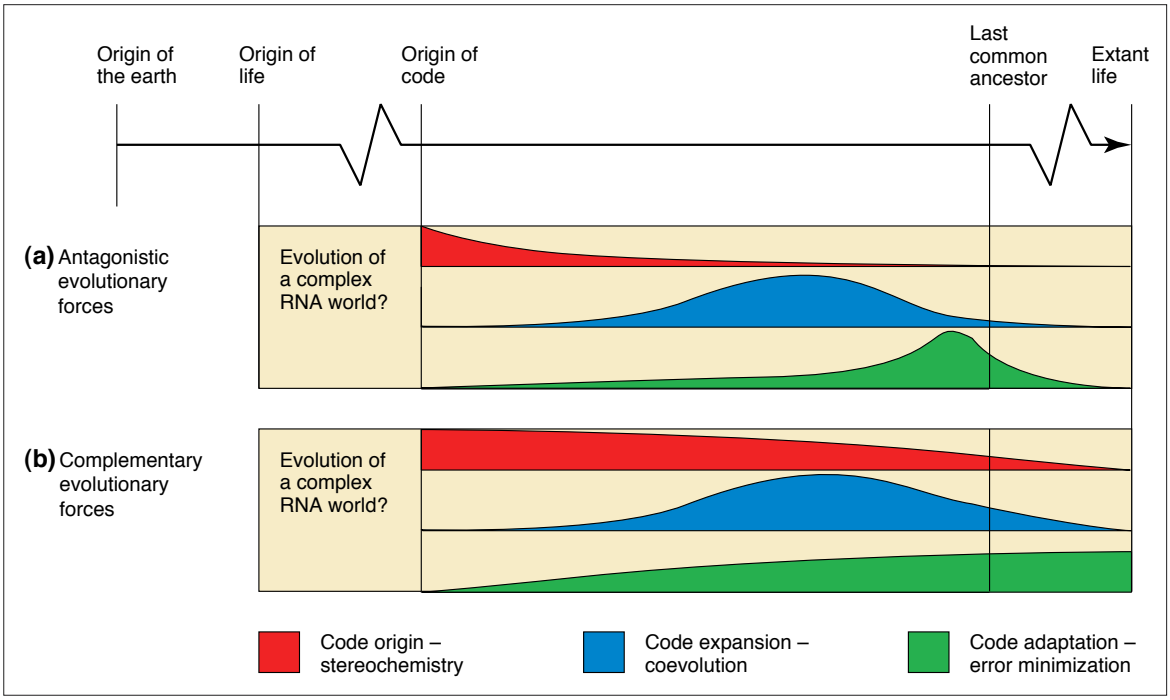


Figure 5

Three facets of code evolution. The genetic code probably originated through stereochemical interactions and, then, underwent a period of expansion in which new amino acids were incorporated. The evolution of the tRNA system, which separated codons from direct interaction with amino acids, then allowed reassignment of codons and, therefore, adaptive evolution. Traditionally, these forces have been assumed to be antagonistic **(a)**, but they might actually have been complementary **(b)**; for example, current codon assignments might assign biosynthetically similar amino acids to similar codons, which would meet both stereochemical and adaptive criteria.

interpretation of these models has undermined the significance of any particular model, especially in the absence of empirical predictions.

Another approach has been to examine interactions between amino acids and individual bases or nucleotides. Early studies showed that 'polar requirement,' a partitioning coefficient of a water–pyridine system that reflects hydrophobicity, varies among second-position bases[1]. Other approaches included tests for the following: (1) correlations between the hydrophobicity of an amino acid and particular nucleotides or dinucleotides; (2) correlations between the partitioning coefficients of amino acids and nucleotides on various surfaces; and (3) differential effects of particular amino acids on nucleotide solubility. These studies tend to show weak associations between anticodons and amino acids[40].

The most direct test of RNA–amino-acid interactions is to determine the precise RNA sequences that bind most strongly to each amino acid. *In vitro* selection, which isolates nucleic acid molecules that bind to a particular target by selective amplification over several generations[41], has generated aptamers (RNA ligands) for several amino acids. Interactions between arginine and RNA have been studied in most detail: several laboratories have selected and characterized the binding of arginine to arginine aptamers. The set of codons assigned to arginine occurs far more often at arginine-binding sites than would be expected by chance: arginine anticodons, and the codon sets assigned to other amino acids, do not show this association[42]. We propose that this is also the case for at least some other amino acids and their codons, and that arginine interacts with its codons in other contexts, such as in RNA-binding proteins. Such intrinsic affinities between codons and amino acids might have influenced early codon assignments. Information about RNA molecules that bind to other amino acids will test the generality of this hypothesis. The first isoleucine aptamers seem to have critical isoleucine codons at their binding sites, although the first valine aptamers do not[43].

The RNA world: the milieu of code evolution?

Translation presents a 'chicken or egg' problem: given that many crucial components of the translation apparatus (including aminoacyl-tRNA synthetases, release factors and much of the ribosome) are made of protein, how could translation ever have evolved? The RNA-world hypothesis[44] avoids this problem by suggesting that RNA preceded DNA and protein and acted as both genetic material and catalyst. The structure of the genetic code might contain information about the chemical environment in which the code evolved.

Two plausible pathways explain how a genetic code arose in an RNA world. First, RNA catalysts might have built specific peptides residue by residue, much in the way that short peptides are now constructed by specific enzymes. Once a general translation system evolved, it would have supplanted these early peptide-synthesis pathways. Second, some ribozymes might have used amino acids, and later peptides, as cofactors[45]. As peptide synthesis became more feasible, the peptide parts of the hybrid catalysts would increasingly have replaced the RNA components; the final result was a protein world in which a few essential nucleotide cofactors remained as molecular fossils. In either case, specific interactions between RNA and amino acids would have been necessary to establish the initial coding system.

Compelling evidence (see above) supports the idea that arginine, and perhaps isoleucine, interacts with its codons in RNA aptamers and that the genetic code is highly optimal with respect to error minimization. When sequences for aptamers for more amino acids are available, we will be able to test whether chemical factors influenced the choice of amino acids and their codon assignments in the canonical genetic code. Assuming that each amino acid was originally assigned those codons for which it has greatest chemical affinity, it would be possible to reconstruct this primordial genetic code. The divergence between this primordial code and the code found in the last common ancestor of all life could test models of early code evolution (Fig. 4).

We envisage a series of definite, although perhaps overlapping, stages in the evolution of the code (Fig. 5). At first, in the RNA world, stereochemical interactions would have largely determined the correspondence between certain RNA-sequence tags and amino acids. Such early peptides, generated by direct templating[43] or similar mechanisms, need not have had catalytic function: for instance, short positively charged arginine repeats might have neutralized the phosphate backbones of RNA molecules, potentially allowing uptake of the latter through membranes[46] and/or their refolding into active structures. As amino acid and peptide cofactors, and eventually catalysts, became more prevalent at the onset of the RNA–protein world, coevolution of the code and the amino acid set might have led to expansion of the code on the basis of metabolic relatedness[47]. This expansion would also have preserved the rules initially established by stereochemical interactions in order to continue making the original templated protein or proteins. Finally, after the evolution of the mRNA–tRNA–aminoacyl-tRNA-synthetase system removed direct interaction between amino acids and codons, codon swapping in different lineages would have permitted some degree of code optimization by codon reassignment.

Code optimization, however, need not be limited to this late stage: error minimization might have acted in concert both with stereochemical considerations and with biosynthetically driven code expansion to produce the canonical code (Fig. 5b). Recent evidence that suggests that the code has a highly optimized structure[7–11] highlights the crucial gap in our understanding of its evolution: the pattern of chemical interactions between the 64 codons and 20 amino acids remains largely unknown. Only when these interactions are known will we be able to understand the relative importance of selection, history and chemistry in code evolution.

References

1 Woese, C. R., Dugre, D. H., Saxinger, W. C. and Dugre, S. A. (1966) *Proc. Natl. Acad. Sci. U. S. A.* 55, 966–974
2 Crick, F. H. C. (1968) *J. Mol. Biol.* 38, 367–379
3 Crick, F. H. C. (1957) *Biochem. Soc. Symp.* 14, 25–26
4 Sonneborn, T. M. (1965) in *Evolving Genes and Proteins* (Bryson, V. and Vogel, H. J., eds), pp. 377–297, Academic Press
5 Woese, C. R. (1967) *The Genetic Code: The Molecular Basis for Genetic Expression*, Harper and Row
6 Crick, F. H. (1966) *J. Mol. Biol.* 19, 548–555
7 Alff-Steinberger, C. (1969) *Proc. Natl. Acad. Sci. U. S. A.* 64, 584–591
8 Haig, D. and Hurst, L. D. (1991) *J. Mol. Evol.* 33, 412–417
9 Ardell, D. H. (1998) *J. Mol. Evol.* 47, 1–13
10 Freeland, S. J. and Hurst, L. D. (1998) *J. Mol. Evol.* 47, 238–248
11 Freeland, S. J. and Hurst, L. D. (1998) *Proc. R. Soc. London Ser. B* 265, 2111–2119
12 Wong, J. T. (1980) *Proc. Natl. Acad. Sci. U. S. A.* 77, 1083–1086
13 Di Giulio, M. (1989) *J. Mol. Evol.* 29, 288–293
14 Di Giulio, M. (1991) *Z. Naturforsch.* 46c, 305–312
15 Di Giulio, M., Capobianco, M. R. and Medugno, M. (1994) *J. Theor. Biol.* 168, 43–51
16 Di Giulio, M. (1998) *J. Mol. Evol.* 46, 615–621
17 Barrell, B. G., Bankier, A. T. and Drouin, J. (1979) *Nature* 282, 189–194

18 Osawa, S. and Jukes, T. H. (1988) *Trends Genet.* 4, 191–198
19 Schultz, D. W. and Yarus, M. (1994) *J. Mol. Biol.* 235, 1377–1380
20 Yarus, M. and Schultz, D. W. (1997) *J. Mol. Evol.* 45, 1–8
21 Wong, J. T-F. (1975) *Proc. Natl. Acad. Sci. U. S. A.* 72, 1909–1912
22 Miseta, A. (1989) *Physiol. Chem. Phys. Med. NMR* 21, 237–242
23 Amirnovin, R. (1997) *J. Mol. Evol.* 44, 473–476
24 Taylor, F. J. R. and Coates, D. (1989) *Biosystems* 22, 177–187
25 Eigen, M. and Winkler-Oswatitsch, R. (1981) *Naturwissenschaften* 68, 282–292
26 Fitch, W. M. and Upper, K. (1987) *Cold Spring Harbor Symp. Quant. Biol.* 52, 759–767
27 Eigen, M. *et al.* (1989) *Science* 244, 673–679
28 Saks, M. E. and Sampson, J. R. (1995) *J. Mol. Evol.* 40, 509–518
29 Saks, M. E., Sampson, J. R. and Abelson, J. (1998) *Science* 279, 1665–1670
30 Nagel, G. M. and Doolittle, R. F. (1995) *J. Mol. Evol.* 40, 487–498
31 Ribas de Pouplana, L., Turner, R. J., Steer, B. A. and Schimmel, P. (1998) *Proc. Natl. Acad. Sci. U. S. A.* 95, 11295–11300
32 Ibba, M., Bono, J. L., Rosa, P. A. and Soll, D. (1997) *Proc. Natl. Acad. Sci. U. S. A.* 94, 14383–14388
33 Landweber, L. F. and Katz, L. A. (1998) *Trends Ecol. Evol.* 13, 93–94
34 Keeling, P. J. and Doolittle, W. F. (1997) *Mol. Biol. Evol.* 14, 895–901
35 Pelc, S. R. and Welton, M. G. E. (1966) *Nature* 209, 868–872
36 Dunnill, P. (1966) *Nature* 210, 1267–1268
37 Root-Bernstein, R. S. (1982) *J. Theor. Biol.* 94, 895–904
38 Hendry, L. B. and Whitham, F. H. (1979) *Perspect. Biol. Med.* 22, 333–345
39 Shimizu, M. (1982) *J. Mol. Evol.* 18, 297–303
40 Lacey, J. C., Jr (1992) *Orig. Life Evol. Biosph.* 22, 243–275
41 Landweber, L. F., Simon, P. J. and Wagner, T. A. (1998) *BioScience* 48, 94–103
42 Knight, R. D. and Landweber, L. F. (1998) *Chem. Biol.* 5, R215–R220
43 Yarus, M. (1998) *J. Mol. Evol.* 47, 109–117
44 Gilbert, W. (1986) *Nature* 319, 618
45 Szathmáry, E. (1993) *Proc. Natl. Acad. Sci. U. S. A.* 90, 9916–9920
46 Jay, D. G. and Gilbert, W. (1987) *Proc. Natl. Acad. Sci. U. S. A.* 84, 1978–1980
47 Dillon, L. S. (1973) *Bot. Rev.* 39, 301–345
48 Osawa, S. (1995) *Evolution of the Genetic Code*, Oxford University Press
49 Tourancheau, A. B. *et al.* (1995) *EMBO J.* 14, 3262–3267
50 Hayashi-Ishimaru, Y. *et al.* (1996) *Curr. Genet.* 30, 29–33
51 Hayashi-Ishimaru, Y., Ehara, M., Inagaki, Y. and Ohama, T. (1997) *Curr. Genet.* 32, 296–299

2 Evolution of the Canonical Genetic Code

This section explores, in detail, the evolution of the genetic code from the RNA world to the LUCA (the last common ancestor of extant life). As outlined in Chapter 1.5, there are three basic models for the evolution of the code (besides the null hypothesis that it is a frozen accident): that codon assignments are determined by selection for error minimization, chemical interactions between amino acids and RNA, or historical relationships within metabolic pathways. Of the three hypotheses, my primary interest has been in testing chemical models: my personal viewpoint is that showing that a biological system is optim**al** **in some respect is insufficient to demonstrate that it has been optim**ized** **for that function, and that to show that the genetic code was a necessary consequence of chemical interactions would, besides being aesthetically pleasing, underscore the role of self-organization in even the most critical of biochemical systems.

However, the true situation seems to be more complex. While there is evidence that at least some amino acids prefer binding sites made of their codons, others do not, and the evidence that the code is optimized to prevent mistranslation is far too strong to ignore. Additionally, there is no reason to believe that the code sprang forth fully formed, like Athena from the forehead of Zeus: rather, it is likely that some amino acids were there from the beginning, while others were later inventions. The difficulty lies in deciding which are which, since there is no agreement in the literature.

Chapter 2.1 outlines the history of stereochemical theories of the code's origin, and provides the most up-to-date summary of the evidence in favor of direct association between triplet motifs and amino acids. Chapter 2.2 presents the first strong statistical evidence for a link between an amino acid, arginine, and its codons (but not its anticodons or other related motifs, or the codons of other amino acids!), while Chapter 2.3 explores the statistical robustness of this result and speculates about how the information contained in a binding preference in the RNA world could have been transmitted to genetic codes in modern organisms.

Chapter 2.4 summarizes the evidence in favor of an adaptive side to code evolution, and examines which features of the code, such as choice of components, pattern of degeneracy, and amino acid identities of codon blocks, seem to have been chosen from a number of possible alternatives by natural selection (in contrast to those that are more likely to be directly chemical, or historical, in origin). In particular, the fact that genetic codes have changed both their codon assignments and degeneracy in modern times implies that adaptation could well have occurred prior to the LUCA. I also explain the conceptual issues underlying several current controversies in the field, and outline fruitful areas for future research.

Chapter 2.5 resolves several questions raised by this review. In particular, I test whether the apparent optimality of the code is due to the specific chemical measurements used, and whether the code is best at conserving the properties of free amino acids, amino acid side-chains, or conformational characteristics in proteins. The later the code evolved, the more likely it is that it would be optimized for properties relevant to modern structural contexts. I also update Woese's Polar Requirement measure using thin-layer chromatography, and measure values for a variety of nonprotein amino acids, in an attempt to explain (a) why the code minimizes errors in Polar Requirement so well, and (b) why some amino acids available to metabolism or on the early earth were excluded from the code.

2.1 Stereochemistry and the Origin of the Genetic Code

This chapter was originally intended to consolidate the evidence for codon/binding site associations, given the recent influx of new aptamer data. Previous work had shown that Tyr, and perhaps Ile, showed significant codon/site interactions as did Arg (Yarus 2000). However, it had not been determined whether these other amino acids show preferences for all and only their cognate codons at binding sites, as does Arg.

The reanalysis revealed a big surprise: when all amino acid binding sites are considered, there Is a strong association with the cognate anticodons *as well as the cognate codons. Both results appear about equally robust, although amino acids typically show a preference for one or the other.*

Does this imply that two *distinct mechanisms of assignment were operative in the RNA world, and that traces of both survive to the present? It is probably premature to draw strong conclusions at present: although the Arg codon/site association suggested that the primitive binding sites evolved into ribozyme aminoacyl-tRNA synthetases rather than tRNAs (see Chapter 2.3), the new data make this straightforward interpretation problematic. At this point, it is probably safest to delay judgment until several independently selected aptamers are available for each of the amino acids.*

Many thanks to Mike Yarus for providing sequence and binding site information prior to publication.

2.1.1 Abstract

Does the genetic code assign similar codons to similar amino acids because of fundamental chemical interactions? Unlike adaptive explanations, which can only explain the relationships of the amino acids to each other, a stereochemical explanation could potentially explain the identity of particular codon assignments. However, in modern systems the tRNA and aminoacyl-tRNA synthetase mediate the codon/amino acid pairing and allow codon reassignment, so we might expect such relationships to be obscured even if they had played a crucial role in the RNA world. Here we examine the evidence that direct interactions between amino acids and nucleic acids recapture some of the relationships in modern genetic codes, implying that selection for error minimization has not erased all traces of these primordial relationships.

2.1.2 Adaptors and Adaptation

Perhaps the most intuitively obvious explanation for the observed order in the genetic code (Pelc 1965; Volkenstein 1965; Woese 1965; Epstein 1966) is that the codon assignments were determined by stereochemical association between nucleotides and amino acids (Dunnill 1966; Pelc and Welton 1966; Woese, Dugre et al. 1966; Woese, Dugre et al. 1966). This mechanism would assign similar amino acids to similar codons because of intrinsic affinity, rather than as a result of natural selection among alternative codes. Although the resulting codon assignments might appear adaptive, in that they reduce various errors relative to other possible codes, they would not be an evolved adaptation.

In modern organisms, there is no direct interaction between codon (or anticodon) and amino acid. The specific coding between codon and amino acid takes place in a two-step process. In the first step, the tRNA is charged with an amino acid by a specific enzyme, the aminoacyl-tRNA synthetase (aaRS). One aaRS exists for each amino acid, and each aaRS simultaneously recognizes the correct amino acid and the correct tRNA (which may be any one of several isoacceptor tRNAs, for amino acids that have more than two codons). In some tRNAs the anticodon is recognized by the aaRS, but many aaRSs do not recognize the codon

at all. In the second step, the tRNA anticodon pairs with the mRNA codon in the ribosome. This takes place largely by Watson-Crick pairing, although many bases in the tRNA are modified (for instance, A is always modified to I when it occurs in the first position of the anticodon). At no stage is the amino acid ever explicitly paired with the codon (Translation process reviewed in Voet and Voet 1995, Chapter 30). However, primitive translation mechanisms may have relied on direct nucleotide-amino acid pairing.

Several different specific pairing schemes have been suggested. The most important criterion is that of continuity: the transition from a primordial coding scheme to the present coding scheme must not destroy the utility of the information already stored at the time of the transition. Thus, the primordial codons with which pairing occurred must either be the actual codons, or some simple transform thereof (Woese, Dugre et al. 1966). Interactions have been proposed between amino acids and codons (Pelc and Welton 1966), anticodons (Dunnill 1966; Ralph 1968), codons read 3' → 5' instead of 5' → 3' (Root-Bernstein 1982; Root-Bernstein 1982), a complex of four nucleotides (C4N) formed by the three 5' nucleotides of tRNA with the fourth nucleotide from the 3' end (Shimizu 1982), or a double-stranded complex of the codon and anticodon (Hendry and Whitham 1979; Alberti 1997).

The fundamental problem that these models share is that codons and amino acids are never stereochemically linked in modern translation. Even if direct association were important in an RNA world, how did it get transmitted to the present? Assuming the original amino acid binders were made of RNA, they could have evolved into any of the components of modern translation: tRNA, rRNA, mRNA, or primitive aminoacyl-tRNA synthetases (subsequently replaced by protein versions). Depending on which component of the translation apparatus is the descendent of amino acid binding sites of the type isolated by modern SELEX experiments, we would predict different associations between trinucleotides and active sites: if binding sites evolved into tRNAs, for instance, the anticodons should be overrepresented, whereas if they evolved into mRNA the codons should be overrepresented (Knight and Landweber 2000). Given a sufficiently large number of RNA molecules that bind specific amino acids, and a sufficiently powerful statistical test, it should be possible to discriminate among these hypotheses.

The existence of adaptors, tRNAs and aminoacyl-tRNA synthetases, in the modern system allows codon assignments to be shuffled among amino acids and hence for adaptive evolution that would erase primordial codon assignments. Additionally, we would only expect *some* amino acids to show codon/site associations, especially if some were added to the code later. Consequently, it is remarkable that *any* associations persist to the present (Yarus 2000).

2.1.3 Chemical Associations: A Historical Perspective

The idea that the genetic code might be stereochemically determined predates the determination of the code itself: Gamow's 'diamond code', in which amino acids would fit into specific pockets formed by four bases in the DNA duplex, gave a model for direct interaction between amino acids and nucleic acids leading to modern codon assignments (Gamow 1954). Although the discovery that tRNA acted as an adaptor between codon and amino acid dampened the enthusiasm for such theories, mathematical (and even numerological) schemes for solving the coding problem abounded before the actual codon assignments were fully uncovered (Woese 1967; Ycas 1969).

The structure of the code showed clear patterns, which required some sort of explanation. Explanations based on chemistry fell into two general classes. Physicochemical theories (Woese, Dugre et al. 1966; Woese, Dugre et al. 1966) assumed that some property of the amino acids led to some sort of interaction between bases and amino acids, perhaps by chromatographic co-partitioning on the early earth, or by direct interaction as measured by NMR. In contrast, stereochemical theories (Pelc 1965; Dunnill 1966) assumed that molecular modeling could reveal direct molecular fits between amino acids and coding triplets, which would explain the individual assignments in the modern code.

Stereochemistry/Molecular Models: The first line of evidence that chemical interactions might have determined modern codon assignments came from molecular modeling studies. According to various authors, molecular models "prove" that the genetic code was established in its modern form by pairing between amino acids and codons in the tRNA (Pelc and Welton 1966), between amino acids and anticodons in the tRNA (Dunnill 1966), between codonic mononucleotides and α-helical homopolymeric amino acids (Lacey and Pruitt 1969) (this model "correctly predicts the glycine codon GGG", although it fails to predict any other), between free glycine and free nucleotides (Rendell, Harlos et al. 1971), by intercalation of the free amino acid into adjacent bases in the anticodon doublet through H-bonding between methylene groups and the π-electrons of the bases (Melcher 1974), by specific 2' aminoacylation of the second position anticodon base mediated by the first position anticodon base (Nelsesteuen 1978), by intercalation of amino acids between first and second position bases in double-stranded RNA molecules (Hendry and Whitham 1979), by insertion of amino acids into cavities caused by removal of the second-position codon base in B-DNA (Hendry, Bransome Jr et al. 1981), by nestling of amino acids into a pentanucleotide cup with the anticodon in the center (Balasubramanian 1982), by pairing between amino acid side-chains with holes in a complex of four nucleotides (C4N) on the acceptor stem of tRNA (Shimizu 1982), and by pairing between amino acids and their codons transposed 3' → 5' (Root-Bernstein 1982; Root-Bernstein 1982).

The modeling approach was tarnished early on, when the claimed association between codons and amino acids (Pelc and Welton 1966) relied on models that had been built backwards, 3' to 5' (Crick 1967) (although see Root-Bernstein 1982; Root-Bernstein 1982 for a relatively recent defense of the idea that there really is a relationship between these reversed codons and amino acids). The main problem with model building is one of statistics: it is difficult to assess whether it would be possible to find equally good associations with a random code, and it is difficult to assess whether a particular amino acid fits all and only its coding triplets (codons, anticodons, etc.). Additionally, these approaches tend to assume that the code was uniquely determined by stereochemical fit (or even that modern variant codes reflect fits induced by different environmental conditions (Mellersh 1993)); if amino acids were added to the code over time, as seems probable (Crick 1968), such explanations are untenable.

Physicochemical Effects/Chromatography: The second line of evidence comes from chromatography. If the chromatographic properties of amino acids (usually, some measure of hydrophobicity) show regular variation in the genetic code, then this might provide some clue to the code's organization. Various studies have shown that the code conserves certain properties, such as polarity, although such evidence cannot tell us whether the code was optimized to minimize errors or established by direct chemical interactions (Knight, Freeland et al. 1999). The polar requirement of amino acids (measured as the ratio of the log relative mobility to the log mole fraction water in a water-pyridine mixture) varies such that amino acids with U in the second position of their codon are hydrophobic while those with A are hydrophilic; those with C are intermediate, and those with G are mixed. Furthermore, codons that share a doublet have almost identical polar requirements even if not otherwise related (e.g. His and Gln; possibly Cys and Trp) (Woese, Dugre et al. 1966; Woese, Dugre et al. 1966; Woese 1967; Woese 1973). However, this model does not provide a mechanism for the actual codon assignments. Partitioning of amino acids and nucleotides between aqueous and organic phases, as in a primordial oil slick, might have associated AAA codons with Lys and UUU codons with Phe (Nagyvary and Fendler 1974), if any of these molecules had existed prebiotically (none of them are produced in prebiotic syntheses (Miller 1987)) and if these chromatographic association had any direct bearing on codon assignment. Analysis with two further chromatographic systems, water/micellar sodium dodecanoate and hexane/ dodecylammonium propionate-trapped water, confirmed the previous hydrophobicity scales in a context closer to prebiotic conditions (Fendler, Nome et al. 1975). Relative hydrophobicity of the homocodonic amino acids (Phe UUU, Pro CCC, Lys AAA, Gly GGG) and the four

nucleotides in an ammonium acetate/ammonium sulfate system showed an anticodonic association, and for dinucleoside monophosphates the association was also with the anticodon, rather than the codon, doublets (Weber and Lacey Jr 1978). Multivariate analysis of the properties of dinucleoside monophosphates and amino acids, focusing on hydrophobicity, revealed many strong ($p < 0.001$) correlations between anticodons and amino acids, but not between codons and amino acids (Jungck 1978).

Thus, the chromatography data suggest anticodonic, rather than codonic, interactions (assuming that molecules with similar properties should interact with one another). However, although chemical partitioning on the early earth could conceivably have led to specific associations between particular nucleotides (or oligonucleotides) and those amino acids that were prebiotically available, there do not seem to be consistent correlations in behavior. Chromatographic separation on various prebiotic surfaces (silicates, clays, hydroxyapatite, calcium carbonate, etc.) showed that, on a silica surface under an aqueous solution of $MgCl_2$ and $(NH_4)H_2PO_4$, Ala comigrates with CMP and Gly comigrates with GMP (Lehmann 1985). Ala is assigned the GCN codon class, while Gly has the GGN codon class. However, there was no strong separation between GMP and UMP or between AMP and CMP even on silica, and many prebiotic amino acids (Pro, Ile, Leu, Val) fell well outside the range of the nucleotides. The situation was even worse on other surfaces, which did not provide any amino acid-nucleotide concordances. Thus, the data do not support the conclusion that copartitioning of nucleotides and amino acids led to the genetic code (Lehmann 1985), especially in the absence of a plausible mechanism for translating this copartitioning into modern codon assignments.

Physicochemical Effects/Direct Interactions: The third line of evidence comes from studies that test for direct interactions between nucleotides and amino acids. Mononucleotides interact nonspecifically and charge-dependently with polyamino acid chains, as measured by the change in turbidity of the solution (Lacey and Pruitt 1969). Affinity chromatography, which tested retardation of the four nucleotide monophosphates by each of nine amino acids (Gly, Lys, Pro, Met, Arg, His, Phe, Trp, Tyr) immobilized by their carboxyl groups, showed no association between binding strength and codon or anticodon assignments (Saxinger and Ponnamperuma 1971). Interactions between free amino acids and poly(A), as measured by the chemical shift of the C_2 and C_8 protons of A (See Fig. 1), are also “not easily reconcilable with the genetic code” (Raszka and Mandel 1972). Further affinity chromatography and NMR experiments on the interaction between amino acids and mono-, di-, and trinucleotides showed that amino acids did selectively interact with specific bases (Saxinger and Ponnamperuma 1974), although the interactions bore no association with the genetic code. Imidazole-activated amino acids esterize with 2’-OH groups of RNA homopolymers with high specificity (Lacey 1975). However, since the two amino acids tested, phenylalanine and glycine, much preferred poly(U) over any other polynucleotide the results do not support the authors’ contention that this mechanism led to the present codon assignments. The dissociation constants of AMP complexes with the methyl esters of amino acids also shows strong selectivity, ranging over an order of magnitude from Trp (120 mM) to Ser (850 mM) (Reuben and Polk 1980). However, neither Trp (UGG) nor Ser (CUN, AGY) have particularly many or few A residues in their codons or anticodons, while the amino acids that do (Lys AAR, Phe UUY) have intermediate dissociation constants (320 and 196 mM respectively). These data did show a strong negative correlation between the association constant ($1/K_D$) and amino acid hydrophobicity, and positive correlations between the dissociation constant and the number of codons assigned to the amino acid and frequency of the amino acid in proteins (Reuben and Polk 1980). Condensation of dipeptides of the form Gly-X in the presence of AMP, CMP, poly(A) and poly(U) was mainly enhanced by the anticodonic nucleotides, where a pattern was apparent (Podder and Basu 1984). Different amino acids differ in their ability to stabilize poly(A)-poly(U) and poly(I)-poly(C) double helices, although the order is the same (except for the following cases: Gly > Ser for A-U but Gly = Ser for I-C; Met, Val > Ile, Leu for AU but Val > Ile, Met, Leu for I – C) (Porschke 1985) and so cannot have contributed to the establishment of the genetic code. Finally, D-ribose adenosine biases esters with L-Phe but

not D-Phe towards the 3'-OH (the pattern is reversed with L-ribose adenosine), indicating that single nucleotides can stereoselectively aminoacylate themselves (Lacey, Wickramasinghe et al. 1993).

Summary: Two comprehensive reviews of these and other data (Lacey and Mullins 1983; Lacey 1992) concluded that the weight of evidence favored specific association between free amino acids and their anticodon nucleotides. Since the various experiments were either equivocal or found correlation between amino acid and nucleotide properties, they indicate that if the genetic code were established by simple interactions between simple molecules (not more complicated than dipeptides or trinucleotides) then the greatest specific binding should be between amino acids and their anticodons.

The various molecular models suggested a wide range of possible interactions, but in most cases several fundamentally incompatible models predicted the same relationships. It would be remarkable indeed if the codon assignments were so robust that any examination of simple transforms of the codons led to the same predictions of modern codon assignments. More likely is that the models have too many degrees of freedom to allow useful predictions, although advances in computer simulation of molecular dynamics may allow statistical assessment of the strength of the predictions of the various models.

All these proposals fundamentally rest on the idea that the genetic code was established very early, perhaps before the synthesis of RNA macromolecules. Indeed, intriguing recent evidence suggests that self-assembly of purine monolayers differentially affects the adsorption of amino acids, and the spacing between residues is consistent with peptide bond distances: it may be that such self-assembly led to the formation of *a* primordial code, although perhaps one very different from the *modern* genetic code (Sowerby and Heckl 1998; Sowerby, Stockwell et al. 2000; Sowerby, Cohn et al. 2001). If this were the case, we might expect RNA binding sites for amino acids to be simple and ubiquitous. It is only in the last decade that a variety of RNA binding sites for amino acids have been discovered, allowing quantitative tests of this and other predictions about primordial RNA/amino acid pairing.

2.1.4 The Codon Correspondence Hypothesis, and Amino Acid-Binding RNA

The codon correspondence hypothesis, which is implicit in any stereochemical theory of the origin of the genetic code, may be stated as follows:

For each amino acid, there is a trinucleotide for which it has greatest affinity. This trinucleotide will be found at the binding sites of RNA molecules that bind amino acids. The association between these trinucleotides and amino acids influenced the form and content of the present genetic code.

These trinucleotides could either be the present codon triplets, the present anticodon triplets, some transformation of these, or some other short RNA structure used in translation today (such as Shimizu's 'C4N' complex at the tRNA acceptor stem (Shimizu 1982)).

The codon-correspondence hypothesis is compatible with establishment of the genetic code either before or during the RNA world. A direct association between trinucleotides and their cognate amino acids would imply an origin of the genetic code prior to complex RNA catalysts, since trinucleotides would likely be randomly synthesized before the directed synthesis of longer oligonucleotides. This might be the case if, for instance, a primitive hypercyclic metabolism relied on trinucleotides to complex undesirable amino acids or to transport desired amino acids to bias the production of short peptides or the composition of longer ones. An association between trinucleotides in the context of a folded RNA tertiary structure and their cognate amino acids would imply an origin of the genetic code in the RNA world, since this would be earliest point in evolution at which long RNA molecules were available. This might be the case if amino acids were originally used as coenzymes for ribozymes (Szathmáry

1993), or to stabilize RNA double helices (Porschke 1985) or to label tRNA-like genomic tags (Maizels and Weiner 1987; Maizels and Weiner 1993).

If there is no association between any trinucleotide and its cognate amino acid, there are several possible explanations. First, no such association might be possible. This would imply that the genetic code evolved in or after a complex RNA world, in which amino acid recognition proceeded through complex and arbitrary interactions. The diversity of RNA molecules that bind arginine (Yarus 1998) shows that efforts to recreate a single, primordial adaptor for each amino acid would be futile. Second, specific associations might exist but be entirely different from the actual codon assignments. This would imply either (a) that these interactions played no role in establishing the genetic code, because more complex and specific binders were available at the time it arose; (b) that these interactions established the original genetic code by direct templating, but the transfer to the present adaptor system caused the original codon assignments to be lost or transformed by some complex function; or (c) that these interactions established the original genetic code, but a long process of codon swapping erased the original assignments as a result of optimization or drift. The latter finding could be used to test the various models of code expansion, since these would predict that the primordial complement of amino acids would bind all and only their cognate primordial trinucleotides. Of the three possibilities listed above, (b) is unlikely unless some transfer mechanism could change all the codons without losing the genetic information. This seems implausible. Both (a) and (c) imply that the genetic code evolved after complex RNA molecules were already available. These hypotheses can only be discriminated by testing the properties of actual RNA molecules.

The obvious place to look for direct associations might be tRNA molecules, but these adaptors have been rendered impotent by evolution. In modern translation, aminoacyl-tRNA synthetases recognize both the tRNA and the amino acid, joining them together: there is no evidence that any sequence on the tRNA directly recognizes the amino acid, and, in fact, the end of the acceptor stem is invariably the same sequence, CCA (see Voet and Voet 1995 for review of translation). Thus it is necessary to look at more exotic examples of RNAs that interact with amino acids.

Most attention in this area has focused on arginine, since arginine binds specifically to two completely distinct classes of natural RNA molecules. The first is probably a molecular coincidence: the guanosine-binding site of self-splicing group I introns also binds arginine, the side-chain of which is similar in structure to the H-bonding face of G (Yarus 1988). However, a conserved Arg codon confers this activity, and the binding site is almost invariably composed of several Arg codons in close juxtaposition (Yarus 1989; Yarus 1991). The second has been extensively studied because of potential medical importance: free arginine can mimic the natural interaction of HIV Tat peptides with TAR RNA (Tao and Frankel 1992). In this case, however, no Arg codons are conserved at the binding site (Yarus 1998).

Natural amino acid-binding RNAs can provide only anecdotal evidence for codon/binding site interactions, because they are almost certainly under strong selection for properties other than maximizing binding to the free amino acids. However, SELEX, a technique for directed molecular evolution (Ellington and Szostak 1990; Robertson and Joyce 1990; Tuerk and Gold 1990), makes it possible to select from large random pools those RNA molecules that perform a desired catalytic or binding function (see Ciesiolka, Illangasekare et al. 1996 for review). This technological advance makes it possible to find out whether RNA molecules that bind to particular amino acids share any characteristic motifs at their binding sites. Aptamers have now been isolated from a variety of amino acids (Table 1), including hydrophobic amino acids such as tryptophan (Famulok and Szostak 1992), valine (Majerfeld and Yarus 1994), phenylalanine/tyrosine (Zinnen and Yarus 1995), isoleucine (Majerfeld and Yarus 1998), tyrosine (Mannironi, Scerch et al. 2000), leucine (I. Majerfeld and M. Yarus, unpublished data), and phenylalanine (M. Illangasekare and M. Yarus, unpublished data), and hydrophilic amino acids such as glutamine (G. Tocchini-Valentini, unpublished data) and citrulline, which is not normally found in proteins (Famulok 1994). However, most research has focused on RNA

aptamers for arginine (Connell, Illangsekare et al. 1993; Yarus 1993; Connell and Yarus 1994; Famulok 1994; Burgstaller, Kochoyan et al. 1995; Geiger, Burgstaller et al. 1996; Tao and Frankel 1996; Yang, Kochoyan et al. 1996), partly because of the analogy to natural amino acid binders. Since structural information is available for many of these sequences, it becomes possible to ask whether particular sequences are overrepresented at binding sites, and, if so, whether these sequences have any relationship to the modern genetic code.

2.1.5 Statistical Evidence for Triplet/Binding Site Associations

Each of the theories of the origin of the genetic code makes specific predictions about the type and likelihood of trinucleotide-amino acid pairings. The various stereochemical theories (see Section 4.4) predict a direct association between amino acids and codons (Pelc and Welton 1966), anticodons (Dunnill 1966; Ralph 1968), codons read 3' → 5' instead of 5' → 3' (Root-Bernstein 1982; Root-Bernstein 1982), a complex of four nucleotides (C4N) formed by the three 5' nucleotides of tRNA with the fourth nucleotide from the 3' end (Shimizu 1982), or a double-stranded complex of the codon and anticodon (Hendry and Whitham 1979; Alberti 1997). The coevolution theories (see Section 4.3) are typically agnostic about which trinucleotide-amino acid pairing established the initial codon assignments, but predict that such pairings, if they exist at all, should correspond to a primordial codon catalog that differs from the present one. Finally, optimization and coevolutionary theories predict no correspondence between trinucleotides and amino acid binding sites.

The selection of RNA molecules (aptamers) that bind amino acid ligands has made these conjectures about the origin of the genetic code testable. If similar amino acids interact favorably with similar RNA sequences, the observed relationships in the genetic code could have a chemical, rather than an adaptive, basis. Because *in vitro* selection searches a large space of possible sequences for optimal or near-optimal "solutions" to particular binding problems, directed evolution may be able to recapitulate primordial interactions between amino acids and short RNA sequences. Since aptamers can be selected to each amino acid, and since the specific nucleotides important to binding can be determined, standard statistical tests for association (such as χ^2 or G) will reveal any consistent interaction between nucleotides participating in binding sites and nucleotides participating in specific short sequences (Knight and Landweber 1998).

As with other aptamer and ribozyme selections, specific sequence motifs recurrent in amino acid aptamers cannot prove that similar interactions actually led to the establishment of present codon assignments. However, the existence of any specific pairings would show that this mechanism could in principle explain the development of the genetic code. If the specific pairings determined from *in vitro* selection actually match the present codon assignments, then it becomes more likely that similar processes took place in primitive cells.

That any codon/binding site associations could survive to the present has been controversial, especially in the absence of a compelling mechanism for transmitting the information to modern translation systems (Ellington, Khrapov et al. 2000) .The association between arginine and its binding sites is exceptionally strong, and has proven remarkably robust to statistical methodology, choice of binding sites, and choice of sequences from selected pools (Knight and Landweber 1998; Ellington, Khrapov et al. 2000; Knight and Landweber 2000; Yarus 2000). In particular, arginine binding sites show strong associations with arginine codons, but not anticodons, codon or anticodon sets for other amino acids, other groups of 4+2 codons incorporating a family box plus a doublet, or other short motifs, and the relationship remains highly significant even if sequences where the selected binding site overlaps the constant regions are excluded from the analysis, the figures are corrected for nucleotide bias at binding sites, or alternative sequences are chosen from the reported pools.

It is less clear whether associations hold for other amino acids, however: arginine may be a statistical fluke, or the fact that it acts as a nucleotide mimic may make it different from other amino acids. Significant associations between Tyr aptamer binding sites and codons have

been reported (Yarus 2000), and Ile aptamers contain conserved Ile codons at their active sites (Majerfeld and Yarus 1998), but it has been unclear whether the associations between these sites and the cognate codons have the same level of robustness as the arginine associations. Data from several other amino acids have only just become available, allowing a more general test for whether the association between binding sites and *codons* is general, or specific to arginine.

Interestingly, only arginine shows more affinity for its cognate codons than for any other codon set (Table 2). Although there is a highly significant association between Tyr binding sites and codons, the association with the Ile codon set is even more highly significant. The other amino acids do not have individually significant codon/site associations (G of 2.7 corresponds to P = 0.05, without correcting for multiple comparisons), although the sign of each of the associations is always positive ($P = (0.5)^6 = 0.016$). However, it is clearly not the case that each aptamer binds its target amino acid using all and only the cognate codons.

It is possible that differences in the fraction of binding and nonbinding amino acids in different experiments could lead to falsely significant results when the data are pooled, due to Simpson's Paradox. Another way of combining the results is to treat each amino acid as a separate observation, and to pool the probabilities using Fisher's method for combining independent tests of a hypothesis (Sokal and Rohlf 1995). This still yields a significant value for both codons ($P = 7 \times 10^{-7}$) and anticodons ($P = 1 \times 10^{-6}$), although the significances are much lower when the single most significant value (Arg or Tyr) is excluded (P = 0.04 and 0.005 for codons and anticodons respectively).

This failure to replicate the codon/site associations found for arginine implied that it would be prudent to reassess the other hypotheses about specific associations: while they may definitely not be true for arginine (Knight and Landweber 1998), they may apply to amino acids in general (Table 3). Individually, only the arginine aptamers showed a significant codon/site association, and only the tyrosine aptamers showed a significant anticodon/site association, when corrected for six multiple comparisons ($P < 0.01$). However, there were highly significant associations overall with *both* codons *and* anticodons, even when the single most influential amino acid was excluded from the analysis ($P < 10^{-4}$ in all cases). There was no significant association for any amino acid, or for the set as a whole, with the codons reversed 3' to 5', indicating that this hypothesis at least can be ruled out.

It is possible that the 21 codon (or anticodon) sets are an unfair comparison class, since they range in size from 1 to 6 codons. A less precise, but perhaps more robust, test is to see whether there is a significant association between the amino acid binding sites and the codon (or anticodon) that contains the cognate doublet: this reflects the intuitively plausible idea that the primitive code may have assigned amino acids only to family boxes. However, the significant associations persist (Table 4). Another possibility is that the association with the anticodons comes from complementary base pairing with functionally important bases in codons. However, this is not the case either: most functionally important residues are in unpaired loops and bulges, and, where an important residue is base-paired, typically only one of the two bases is protected against modification in the presence of ligand. In the Gln and Tyr aptamers, a few paired bases are protected, but there is no case where both members of a pair are protected. For Ile, the only paired residue is a G-U mismatch, which cannot explain a codon/anticodon association. The Arg aptamers contain a total of five base pairs where both members are important; however, there is no correlation between Arg binding sites and anticodons, despite the fact that it is only in these aptamers that base-pairing could have a statistical effect. Thus the results to date are equivocal between codons, anticodons, or some combination of the two.

2.1.6 Conclusions and Future Directions

There are basically four possible roles that the descendants of the original amino acid-binding sites could play: such sites could have evolved into tRNAs, mRNAs, ribosomes, or aminoacyl-

tRNA synthetases. We know that RNA can play all of these roles (Illangasekare, Sanchez et al. 1995; Illangasekare and Yarus 1999; Illangasekare and Yarus 1999; Cech 2000; Nissen, Hansen et al. 2000). If aptamers showed a clear preference for only the codons, or for only the anticodons, of the cognate amino acids, it would be possible to tell which of these fates might have befallen the original binding sites (Knight and Landweber 2000). Alas, the situation appears to be more complicated. Although the association between arginine and its codons is robust, no such associations clearly hold for the other five amino acids for which well-characterized aptamers are now available.

The main difficulty is the limited data available: ideally, with a large sample of independently derived families of aptamer that bind each of the amino acids, it should be possible to test associations between binding sites and individual trinucleotides. It is possible that high-throughput techniques for aptamer isolation will achieve this in the future, but, for the moment, isolating aptamers is a difficult and time-consuming process. Consequently, it may be several years before the picture of site/triplet associations becomes clear. It is definitely premature to conclude from the few aptamers currently available that associations between binding sites between amino acids and their cognate *codons* are of primary importance: except for the case of arginine, which may be unique, there is equally good support for a statistical association with the *anticodons.* Of course, it is possible that two distinct mechanisms for codon assignment operated in the RNA World and survived to the present; however, it will take far more data than are available at present to support such a conclusion robustly.

The fact that binding sites of high affinity are far more complex than single trinucleotides strongly suggests that, if these interactions between triplets and amino acids played a role in determining the modern code, the code must have originated after complex RNA molecules were available to primitive organisms. Given past experience, molecular modeling is unlikely to resolve questions about *which* associations, if any, were of primary importance: the only solution is to find a sufficiently large number of RNA molecules that actually bind amino acids. The potential of this technique is great: it may be possible to discover why amino acids have the actual codon assignments they do, and perhaps why some amino acids were incorporated into the code while others available on the early earth or as metabolic intermediates were excluded. Furthermore, it may be possible to discover the original, stereochemically determined code, and therefore to assess the relative roles of chemistry and selection in shaping modern codon assignments. At this stage, however, many questions still await resolution.

Amino Acid	Kd	Comments	Reference
Arg	400μM	Group I intron: naturally binds G	(Yarus and Majerfield 1992)
Arg	4mM	TAR: Naturally binds Tat peptide in HIV	(Tao and Frankel 1992)
Arg	1mM	3 families selected; no structures available	(Connell, Illangsekare et al. 1993)
Arg	**4mM**	**Selected against GMP binding**	**(Connell and Yarus 1994)**
Arg	**2-4mM**	**Selected by salt elution to mimic TAR**	**(Tao and Frankel 1996)**
Arg	**60μM**	**Derived from citrulline binder by mutagenesis/reselection; NMR structure available**	**(Famulok 1994)**
Arg	**330nM**	**Intensive selection with heat-denaturation; only one sequence structurally characterized, though many selected**	**(Geiger, Burgstaller et al. 1996)**
Val	12mM	No structural data	(Majerfeld and Yarus 1994)
Ile	**200-500μM**	**Only one family survived selection**	**(Majerfeld and Yarus 1998)**
Phe/ Tyr	2-25mM	No structural data	(Zinnen and Yarus 1995)
Trp	18μM	Binds D-Trp-agarose, not free L-Trp; no structural data	(Famulok and Szostak 1992)
Tyr	**35μM**	**Also binds Trp; evolved from L-DOPA binder**	**(Mannironi, Scerch et al. 2000)**
Phe	**<1mM**	**Some clones bind only Phe-agarose**	**Illangasekare & Yarus, unpublished data**
Leu	**~1mM**		**Majerfeld & Yarus, unpublished data**
Gln	**18-20mM**		**Mannironi et al., unpublished data**

Table 1: Natural and Artificial Amino Acid-Binding RNA. Entries in **bold** are those with sufficient structural information about the binding site to test for statistical association between binding sites and triplet motifs. Natural RNA sequences that bind arginine were excluded from the analysis, because they are probably under selection for other properties.

Codons	Arg	Tyr	Ile	Gln	Leu	Phe
Ter	0.05	1.28	*0.13*	-2.17	*2.65*	*11.28*
Ala	0.09	-16.95	-0.16	-6.29	-0.38	-22.23
Cys	-16.97	-0.66	-3.01	0.09	0.04	-2.75
Asp	0.15	3.96	-3.01	*8.91*	-1.08	-0.87
Glu	3.44	-3.17	0.00	-1.34	*1.47*	*7.59*
Phe	-3.38	-2.38	-0.07	*3.54*	-2.00	**1.53**
Gly	0.35	0.25	*2.99*	1.65	0.00	*1.79*
His	-1.04	-1.87	-3.01	-3.48	-0.68	-2.55
Ile	2.86	*9.18*	**0.02**	1.78	-4.60	0.34
Lys	1.34	-14.86	0.00	1.78	0.62	*1.90*
Leu	-19.92	-4.16	*1.58*	-2.75	**0.83**	-4.91
Met	-5.60	3.06	0.00	1.54	-1.35	0.00
Asn	5.46	-0.04	0.00	-1.55	0.01	*2.36*
Pro	0.00	-2.30	-4.82	-2.82	-0.15	-11.73
Gln	0.27	2.30	-3.01	**2.36**	0.62	1.08
Arg	***29.11***	0.24	*0.24*	-0.07	-0.78	-0.15
Ser	-6.07	-4.95	-3.53	-1.91	*5.65*	-7.96
Thr	-0.10	0.57	-0.86	-6.04	*2.61*	-4.79
Val	-0.13	4.45	*5.47*	0.00	*2.82*	*2.41*
Trp	-7.26	0.04	*7.34*	-2.17	0.28	-0.37
Tyr	-3.38	**6.69**	*0.24*	-0.94	-0.12	0.29
Rank	**1**	**2**	**8**	**3**	**6**	**7**

Table 2: Tests for association between amino acid binding sites and their cognate codons. Rows: codon sets for each amino acid; columns: amino acids for which aptamers with known structures have been reported. Bold values (boxed) indicate the cognate codon sets for each amino acid aptamer; values in italics (gray shading) indicate codon sets with at least as strong an association as the actual codon set. All values are the G test for association between codons and binding sites, with the Williams correction; negative values indicate codon sets that are found less frequently at binding sites than would be expected by chance. 'Rank' indicates the rank order of the cognate amino acid's codon set. Arginine is the only amino acid for which the aptamers show a unique association between codons and binding sites. Binding sites for this table and all others are taken from (Yarus 2000) where applicable (Arg, Ile, Tyr), or otherwise from personal communications from the people who isolated the aptamers. See Knight and Landweber 2000 for discussion of the effects of different choices of binding site.

Codons	n	+b+c	+b-c	-b+c	-b-c	G	P
Arg	5	36	16	38	106	29	*6.8E-08*
Tyr	3	12	71	9	179	6.7	9.7E-03
Ile	1	3	12	9	40	0.02	8.9E-01
Gln	3	3	12	9	134	2.4	1.2E-01
Leu	2	16	46	19	78	0.83	3.6E-01
Phe	8	8	63	38	515	1.5	2.2E-01
Total	**22**	**78**	**220**	**122**	**1052**	**44**	***3.7E-11***
Total - Arg	17	42	204	84	946	16	*8.1E-05*

Anticodons	# seq	+b+c	+b-c	-b+c	-b-c	G	P
Arg	5	20	32	37	107	2.9	8.9E-02
Tyr	3	18	65	6	182	22	*3.3E-06*
Ile	1	3	12	9	40	0.02	8.9E-01
Gln	3	0	15	18	125	-3.5	6.2E-02
Leu	2	27	35	23	74	6.7	9.4E-03
Phe	8	10	63	38	515	2.9	8.8E-02
Total	**22**	**78**	**222**	**131**	**1043**	**38**	***6.7E-10***
Total - Tyr	19	60	157	125	861	26.8	*2.3E-07*

Reversed Codons	# seq	+b+c	+b-c	-b+c	-b-c	G	P
Arg	5	16	36	42	102	0.05	8.3E-01
Tyr	3	3	80	6	182	0.03	8.6E-01
Ile	1	4	11	10	39	0.24	6.2E-01
Gln	3	1	14	17	126	-0.38	5.4E-01
Leu	2	12	50	29	68	-2.2	1.4E-01
Phe	8	2	63	38	515	-2.8	9.5E-02
Total	**22**	**38**	**254**	**142**	**1032**	**0.17**	**6.8E-01**

Table 3

Table 3: Test for association between binding sites and the cognate codons, anticodons, and codons reversed 3' to 5'. Column headings: n, number of sequences; +b+c, number of bases both in codons and in binding sites; +b-c, number of bases in binding sites but not in codons; -b+c, number of bases in codons but not in binding sites; -b-c, number of bases neither in codons nor in binding sites; G, the G test for association in a 2 x 2 table, with the Williams correction; P, 2-tailed test for independence with 1 degree of freedom. Values in italics are significant to $P < 0.01$ after correcting for 6 comparisons. There appears to be a strong association between Arg and its codons, and between Tyr and its anticodons, but the other associations are equivocal. Overall, there are significant associations between amino acid binding sites and both their codons *and* anticodons, even when the single most significant association is removed, which is difficult to explain in light of the fact that codons with complementary anticodons typically encode amino acids with very different chemical properties. There is no association between amino acid binding sites and the codons reversed 3' to 5'.

Codon Doublets	# seq	+b+c	+b-c	-b+c	-b-c	G	P
Arg	5	40	64	49	239	18.5	*1.7E-05*
Tyr	3	22	61	20	168	10.2	*1.4E-03*
Ile	1	3	12	9	40	0.02	8.9E-01
Gln	3	3	12	27	116	0.01	9.2E-01
Leu	2	11	113	37	157	-6.5	1.1E-02
Phe	8	14	57	99	454	0.14	7.1E-01
Total	**22**	**93**	**319**	**241**	**1174**	**57.7**	***3.0E-14***
Total - Arg	17	53	255	192	935	26.7	*2.4E-07*
Anticodon Doublets	# seq	+b+c	+b-c	-b+c	-b-c	G	P
Arg	5	25	79	54	234	1.28	2.6E-01
Tyr	3	21	62	18	170	12.5	*4.1E-04*
Ile	1	4	11	10	39	0.24	6.2E-01
Gln	3	3	12	45	98	-0.9	3.5E-01
Leu	2	24	100	31	163	0.59	4.4E-01
Phe	8	23	48	76	477	13.6	*2.3E-04*
Total	**22**	**100**	**312**	**234**	**1181**	**12.1**	***5.0E-04***
Total - Tyr	19	79	250	216	1011	6.62	*9.6E-04*

Table 4: Test for association between binding sites and codon doublets (XYN) or anticodon doublets (NY'X'), where X and Y are specified and N is any base. For example, the codon doublet for Phe is UUN, and the anticodon doublet is NAA. Again, the specific associations hold for both codons and anticodons overall, although few of the results are individually significant. Italics indicate significant values after correction for 6 comparisons.

2.2 Rhyme or Reason: RNA-Arginine Interactions and the Genetic Code

This chapter is included primarily for historical reasons, since the following chapter provides a far more complete analysis of the same data. The original suggestion for my thesis project was to focus on selecting RNA aptamers to homopolymers of single amino acids: the idea was that it would be impossible to get other than anecdotal evidence about codon/amino acid relationships from single amino acid binding sites, but that a homopolymer would present a large, repeating motif so that if codons were involved in binding they would be overrepresented throughout the molecule (and so detailed and laborious chemical mapping of the binding sites would be unnecessary).

Some time after starting this project, I noticed that published amino acid aptamers typically only had a few nucleotides involved in the actual binding site, and many fewer critical nucleotides. I wondered whether it might be possible to test whether nucleotides involved in binding sites were disproportionately likely to be found in codons: at this time, only one amino acid (Arginine) had several different aptamers selected in different labs. I was stunned when a quick χ^2 test revealed that the odds of finding as many codons as observed at binding sites were less than one in 100 000!

This chapter has previously appeared in Chemistry and Biology *as:*

Knight, R. D. and L. F. Landweber (1998). "Rhyme or reason: RNA-arginine interactions and the genetic code." Chem Biol 5(9): R215-20.

This was my first paper in a peer-reviewed journal, and I should take the opportunity to thank Prof. Landweber for her extensive assistance with writing and revision, and also for the suggestion that we use the G test for independence instead of χ^2.

Rhyme or reason: RNA–arginine interactions and the genetic code

Robin D Knight and Laura F Landweber

Theories about the origin of the genetic code require specific recognition between nucleic acids and amino acids at some stage of the code's evolution. A statistical analysis of arginine-binding RNA aptamers now offers the opportunity to test such interactions and provides the strongest support for an intrinsic affinity between any amino acid and its codons.

Address: Department of Ecology and Evolutionary Biology, Princeton University, Princeton, NJ 08544-1003, USA.

Correspondence: Laura F Landweber
E-mail: lfl@princeton.edu

Chemistry & Biology September 1998, **5**:R215–R220
http://biomednet.com/elecref/10745521005R0215

Introduction

The selection of RNA molecules (aptamers) that bind amino-acid ligands has made theories about the origin of the genetic code at last testable. The genetic code assigns similar amino acids to similar codons (Figure 1) [1]. This could be a result of selection to minimize the effect of mutations [2,3] or translation error [4], or of codon concession by metabolic precursors to related derivatives [5,6]. Alternatively, if similar amino acids interact favorably with similar RNA sequences, the observed relationships in the genetic code could have a chemical, rather than an adaptive, basis [7]. If any codon assignments arose from specific binding between amino acids and short RNA motifs, then directed evolution might be able to recapitulate such interactions *in vitro*.

Various authors have suggested that the original amino-acid-binding motifs could have been the actual codons [8] or some transform of them [4], such as the anticodon [9] or codon–anticodon duplexes [10]. Recent support for the codon–amino-acid pairing hypothesis comes from a specific interaction between arginine and its codons; the guanosine-binding site of self-splicing group I introns also binds arginine, and a conserved arginine codon confers this specificity [11].

Another possibility is that the original amino-acid recognition took place at the tRNA acceptor stem rather than the anticodon [12]. Such recognition could occur, for example, by a stereochemical interaction at a 'C4N' (complex of four nucleotides) [13], which consists of the three nucleotides at the 5′ end (assumed to be identical in sequence to the anticodon) together with the 'discriminator base' (the nucleotide immediately preceding the invariant CCA at the 3′ end of the tRNA). This model is consistent with evidence that the acceptor-stem domain and the anticodon domain of tRNA molecules might have independent evolutionary histories [14–18]. The various sites that have been suggested as the primitive binding sites are shown in Figure 2.

Each of the stereochemical hypotheses predicts that specific short RNA motifs will be found at sites that bind amino acids. Consequently, aptamers (nucleic-acid molecules selected to bind specific ligands) [19–21] that recognize amino acids should contain these sequences at their binding sites. RNA aptamers have been isolated to several amino acids, but most research has focused on RNA aptamers that bind arginine [22–27] because free arginine can mimic the natural interaction of HIV Tat peptides with TAR RNA [28].

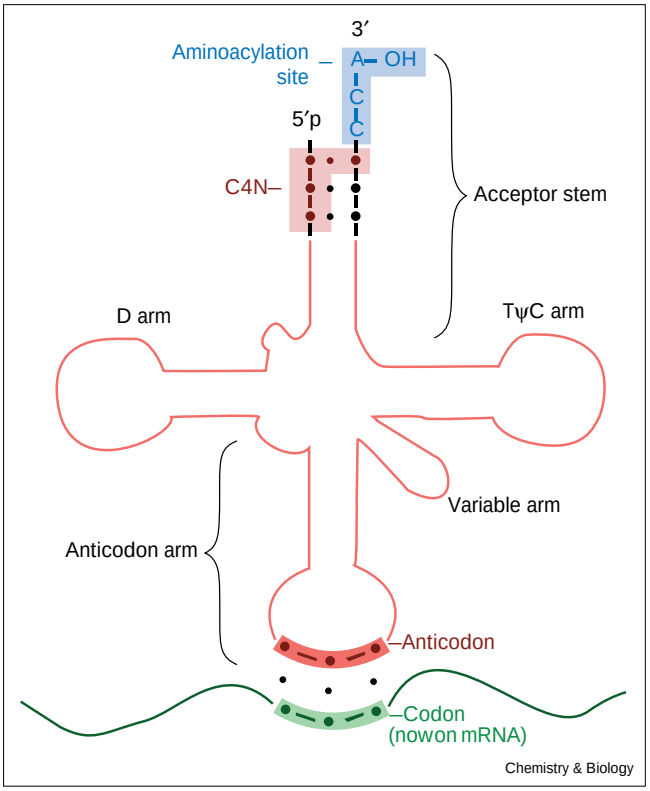
Figure 1

	U		C		A		G	
U	UUU	Phe	UCU	Ser	UAU	Tyr	UGU	Cys
	UUC	Phe	UCC	Ser	UAC	Tyr	UGC	Cys
	UUA	Leu	UCA	Ser	UAA	TER	UGA	TER
	UUG	Leu	UCG	Ser	UAG	TER	UGG	Trp
C	CUU	Leu	CCU	Pro	CAU	His	CGU	Arg
	CUC	Leu	CCC	Pro	CAC	His	CGC	Arg
	CUA	Leu	CCA	Pro	CAA	Gln	CGA	Arg
	CUG	Leu	CCG	Pro	CAG	Gln	CGG	Arg
A	AUU	Ile	ACU	Thr	AAU	Asn	AGU	Ser
	AUC	Ile	ACC	Thr	AAC	Asn	AGC	Ser
	AUA	Ile	ACA	Thr	AAA	Lys	AGA	Arg
	AUG	Met	ACG	Thr	AAG	Lys	AGG	Arg
G	GUU	Val	GCU	Ala	GAU	Asp	GGU	Gly
	GUC	Val	GCC	Ala	GAC	Asp	GGC	Gly
	GUA	Val	GCA	Ala	GAA	Glu	GGA	Gly
	GUG	Val	GCG	Ala	GAG	Glu	GGG	Gly

Chemistry & Biology

The 'universal' genetic code. Similar colors reflect similar sidechain composition: purple, hydrophobic; cyan, aromatic; green, hydroxyl containing; red, acidic; orange, amide; blue, basic; yellow, sulfur containing. Tyrosine is aromatic and contains a hydroxyl group, and so is intermediate between green and cyan; similarly, histidine is aromatic and basic. Intensity of color reflects molecular volume [41]. In general, amino acids with similar properties are linked by single-base mutations.

Figure 2



Structure of modern tRNA and mRNA, showing the codon, anticodon and C4N in their present positions. Note that in contemporary organisms the anticodon never contacts the amino acid directly (although it is sometimes important for recognition by the correct aminoacyl-tRNA synthetase).

As with the group I intron, another anecdotal but compelling example of a specific interaction between an amino acid and its codons comes from an arginine aptamer produced by randomization and reselection from an aptamer for the closely related amino acid citrulline [25]. Arginine differs from citrulline only by one moiety: arginine has an imino group where citrulline has a carbonyl group (Figure 3). The arginine aptamer differs from the citrulline aptamer by precisely three point substitutions, which together create two new arginine codons (Figure 4). Nucleotides in both of these arginine codons surround the amino-acid sidechain, forming a set of hydrogen bonds that interact directly with the amino acid [29].

To test whether there is a statistical association between binding sites and codons, it is necessary to have structural data for several independent sequences that bind the same amino acid. Because structural data are available for five phylogenetically unrelated arginine aptamers selected in four experiments in three different laboratories under different conditions [23,26,27,29], we tested whether the arginine-binding sites of these RNA aptamers contain a statistical excess of any of the predicted motifs (codons or anticodons). This allowed us to test the validity of several stereochemical hypotheses. Because we performed many individual tests on the same data, and because statistical results are often controversial in this field, we chose an arbitrary, low cutoff of $P = 0.01$ for statistical significance in any test.

Statistical analysis of aptamers

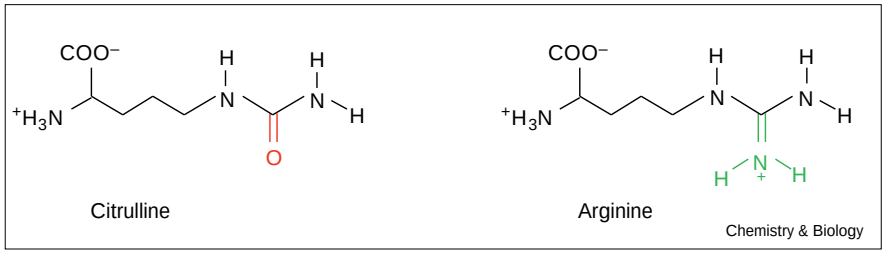
If one or more of the stereochemical hypotheses were true, nucleotides at arginine-binding sites would be expected to participate in arginine codons and/or anticodons. We analyzed 2×2 tables partitioning nucleotides into 'binding site' and 'nonbinding site' classes and into 'motif' and 'nonmotif' classes, depending on whether they are present in the ligand-binding site or in a particular nucleotide-sequence motif, respectively. 'Binding site' nucleotides are either shown by nuclear magnetic resonance (NMR) to form the ligand-binding pocket [29], or are implicated in binding by chemical-modification assays [23,26,27] (Table 1a–e). The latter included both chemical modification protection/deprotection assays (in which nucleotides are protected from chemical modification and/or enzymatic cleavage only when arginine is bound), and assays that measure the extent to which base modification impairs binding. 'Motif' nucleotides participate in the appropriate triplet(s) in any of the three possible reading frames. A statistical association between the two properties would indicate that distribution of the motifs is nonrandom with respect to arginine-binding sites. We used the G test for independence with Williams's correction for continuity [30] to detect interaction between the two variables. Where predictions were directional, the calculated p values were halved (and subtracted from unity when the deviation was in the opposite direction from that predicted) to reflect the fact that deviations in only one direction were relevant.

To ensure that any observed association was due to sequence rather than composition, we also tested triplets in which the first two bases were permuted. Thus, for the arginine codon classes CGN and AGR (N, any nucleotide; R, purine), we tested the non-arginine codons GCN and GAR, which have identical nucleotide compositions but different sequences.

As a further negative control, we performed the same statistical tests on all the RNA aptamers to ligands other than arginine for which published NMR structures were available (Table 1f–j). The aptamers to AMP [31], FMN [32], citrulline [29,33], tobramycin [34] and theophylline [35] would not be expected to show the same binding-motif associations as those for arginine, unless such associations arise from biases in composition that are general to binding sites in all RNA aptamers. NMR determines binding sites more accurately than chemical probing; we

Figure 3

Structures of arginine and citrulline.



therefore restricted our analysis of non-arginine aptamers to those that had been examined using this technique.

Purine bias

Binding sites of all aptamers are strongly purine biased. Although purines make up 54% of arginine and 58% of non-arginine aptamer sequences, they constitute 78% and 72% of their respective binding sites. Because one class of arginine codons, AGR, consists entirely of purines, such a purine bias could produce a spurious association between binding sites and arginine codons. Purine bias should also result in excess AAR, GAR, and GGR purine triplets at binding sites in arginine aptamers, however. We observed no such excess ($P >> 0.01$) in any case.

Only arginine aptamers overrepresent the arginine set

Arginine aptamers contain far more arginine codons at binding sites than expected by chance (Table 2; $G = 20.2$; $P = 3.4 \times 10^{-6}$). Of the 32 nucleotides at binding sites of arginine aptamers, 23 (72%) participate in arginine codons. In contrast, of the 59 nucleotides at binding sites of non-arginine aptamers, only 17 (29%) participate in arginine codons. The probability that a base in a random sequence is in an arginine codon is bounded by the relationship (6 arginine codons/64 total codons) × 3 reading frames = 28%, although simulation of a Markov process using the base frequencies in the arginine aptamers indicates that the true probability is closer to 34%. As expected, aptamers to ligands other than arginine show no association between arginine codons and arginine-binding sites ($G = 0.14$, $P >> 0.01$).

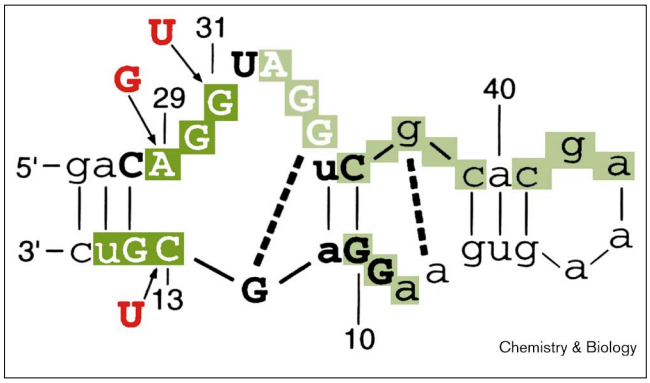
Arginine aptamers overrepresent only arginine codons

If the overrepresentation of codons at arginine-binding sites were due to composition rather than to sequence, then permutations of the codons (i.e. GCN for CGN and GAR for AGR) should be similarly overrepresented. This is not the case (see above). The set of arginine codons showed a much higher association with arginine-binding sites ($G = 20.2$; $P = 3.4 \times 10^{-6}$) than did the codon set of any other amino acid. The next highest match was proline ($G = 4.63$; $P > 0.01$). We also found no association between arginine-binding sites and the arginine anticodons NCG and YCU ($G = 0.19$; $P >> 0.01$; Y, pyrimidine).

No other codon set binds arginine as well as arginine codons

To ensure that the strong association we observed between arginine-binding sites and arginine codons was not a fluke, we tested for possible association between arginine-binding sites and all possible codon sets. We defined a codon set as a 'family box' consisting of all four third-position variants of one codon fixed at the first two positions, together with a 'doublet' consisting of two fixed bases followed by either pyrimidine or either purine. This describes all actual six-codon sets: for instance, serine is UCN + AGY. Out of 480 possible codon sets, the actual arginine set (CGN + AGR) has the highest G of 20.2 ($P = 3.4 \times 10^{-6}$). The next highest set (AGN + CGR), which contains four of six arginine codons, gave a degree of association almost two orders of magnitude less improbable ($G = 13.1$; $P = 1.5 \times 10^{-4}$). The highest association between codons and arginine-binding sites for sets not including arginine codons was for purine-rich GAN + GGY ($G = 5.78$; $P = 0.008$), which is not significant because 378 such codon sets were tested (the probability

Figure 4



Secondary structure of an arginine aptamer derived from a citrulline aptamer by three nucleotide substitutions (arrows); all occur within two new arginine codons (dark green boxes). Note that creation of the first arginine codon requires substitutions at both the first and third positions. Four additional arginine codons are shown as light green boxes. Essential nucleotides are in boldface; nucleotides selected in all isolates in uppercase. Dashed lines indicate noncanonical pairs. Adapted from [29].

Table 1

Sequences and nucleotides forming binding sites of arginine and non-arginine aptamers.

(a) gacAGGuAgGucgcacgaaaguga**a**G**g**aG**C**guc

(b) gggagcucagaauaaacgcucaacccgacagaucggcAaCgCCnuguuuucgacangAgACaccgauccugcaccaaagcuucc

(c) augauAAAccgAugcugggcgAuucuccugaaguaggggaagAguugucauguauggg

(d) gggagaauucccgcggcuggugcgugcaggacGUcGAucgaaucCGccugcaGugcacggcuccc

(e) gggagaauucccgcgcagcGGUcGAaaucgucaugugcacugcuacugcagugcacggcuccc

(f) ggguugGGAAGAAacuguggcacuucggugccAGcaaccc

(g) gacGGUuAgGucgcacgaaaguga**a**G**g**aGUguc

(h) ggcgugu**a**GGAUAugcuucggcaGAAGGacacgcc

(i) ggcacgagGUUUAGCUACAcucgugcc

(j) ggcgaUACCagccgaaaggcccuugGCaGcgcuc

Sequences of aptamers used in this analysis for (a–e) arginine [23,26,27,29], (f) AMP [31], (g) citrulline [29], (h) FMN [32], (i) tobramycin [34] and (j) theophylline [35]. Capital letters denote nucleotides implicated in binding, as determined by NMR or by chemical and enzymatic probing. Nucleotides participating in CGN arginine codons are green; those participating in AGR arginine codons are red.

that at least one trial would give a result this extreme is greater than 0.01).

Monte Carlo simulations

Finally, we tested the appropriateness of the G test, which requires independent observations, because the probability that a nucleotide participates in a codon or binding site is influenced by its surrounding nucleotides. We ran Monte Carlo simulations that randomly generated sequences of the same length as the actual aptamers, and tested whether the observed values of G matched the probabilities obtained from the standard G test. In the weakest condition, the randomized aptamers were constrained to the same base frequencies overall as the actual aptamers and the binding-site positions were randomized. Under these conditions, only 1 of the 100,000 randomized aptamer sets gave a G value higher than the actual set, indicating $P \approx 10^{-5}$. In the most stringent condition, first, nucleotides were assigned with probabilities equal to their actual frequencies in binding sites and elsewhere in the aptamers, to account for compositional bias, and second, binding sites were fixed to their actual positions in the aptamers, to account for spatial correlation within binding sites. Under these conditions, 124 of the arginine aptamers showed positive codon-binding site associations at least as strong as that observed for arginine aptamers. In contrast, 48,985 of the non-arginine aptamers showed associations at least as strong as those observed for non-arginine aptamers. Thus we conclude that only arginine aptamers have significant bias in favor of arginine codons at their binding sites, and that the true probability of association between arginine codons and binding sites might be closer to 1×10^{-3} than 6×10^{-6} (as calculated without accounting for composition bias). Even in the most stringent tests, however, the association between arginine codons and arginine-binding sites remained significant ($P << 0.01$). This analysis also revealed that the true probability of association between AGR codons and binding sites for ligands other than arginine is much greater than 0.01, as 8356 of the 100,000 randomizations of non-arginine codons gave positive associations at least as strong as that observed. (Because of the strong purine bias, AGR codons tended to associate with non-arginine-binding sites [$G = 15.3$; $P = 4.6 \times 10^{-5}$]. GGR codons showed a similar association [$G = 6.47$; $P = 0.005$], when the binding-site nucleotide biases were not taken into account.)

Conclusions

These results show a clear association between both arginine codon classes (CGN and AGR) and regions of RNA molecules that bind arginine. We found no association between arginine-binding sites and arginine anticodons, and no stronger associations between arginine-binding sites and any other possible set of six codons conforming to the 4 + 2 rule of a family box and doublet (see above).

Table 2

Arginine codon/binding-site frequencies for arginine and non-arginine aptamers.

	Codon	nt in Codon*	nt not in Codon*	G	P
Arginine aptamers	Binding	23	9	20.2	3.4×10^{-6}
	Not binding	83	190		
Others	Binding	17	42	0.14	0.35
	Not binding	29	82		

Tests for association between codons and binding sites were directional. *The number of nucleotides involved in arginine codons need not be a multiple of three, because some codons overlap. nt, nucleotide.

Because the C4N hypothesis states that the anticodon forms part of the binding complex, these results provide evidence against both it and the codon–anticodon double-helix hypothesis. If *in vitro* selection protocols mimic an appropriate evolutionary environment, and if aptamer selections are influenced by the same chemical interactions that led to codon assignments, then these results (for the one amino acid for which sufficient data are available) support the hypothesis that amino acids can interact specifically with RNA sequences that contain their cognate codons [36]. If the present-day arginine codons preserve original assignments determined by stereochemical affinity between amino acids and RNA, then their positions in the genetic code could even be considered molecular fossils of an ancient chemical determinism.

Arginine is an unusual amino acid in that the guanidino moiety of its sidechain closely mimics the hydrogen-bonding face of guanine. Furthermore, its positive charge allows it to participate in electrostatic interactions with the phosphate backbone of RNA that are not available to other amino acids. Specific RNA aptamers have been selected to hydrophobic amino acids [37–39], however, and thus interactions other than ionic and hydrogen-bonding must contribute to amino-acid recognition, although there are insufficient structural data to perform statistical tests on these other aptamers. The selection experiments were carried out in three different laboratories and from independent starting pools, so the resulting aptamers should be an unbiased representation of sequence space. Furthermore, attempts to select aptamers against positively charged lysine have been unsuccessful [25], despite the potential for electrostatic interactions similar to those used by arginine aptamers.

Alternatively, it has been suggested that, rather than being primitive, arginine could have been a relatively late addition to the standard repertoire of amino acids [40]. Thus, arginine could have captured those codons for which it has greatest affinity. If so, we conjecture that amino acids incorporated earlier might not show such an association between their codons and binding sites. This hypothesis can only be tested with structural data for multiple aptamers to amino acids besides arginine.

Prospects

These results imply the specific prediction that arginine codons will play prominent roles in other RNA–arginine interactions, and that some of the other 19 amino acids will bind RNA sequences that contain their codons. More general conclusions await the structural analysis of aptamers to other amino acids. For example, Majerfeld and Yarus [38] recently reported that isoleucine aptamers contain isoleucine codons at their isoleucine-binding sites. Because isoleucine has an aliphatic sidechain, electrostatic interactions cannot be involved in this case. If the codon-binding-site relationship holds true for other amino acids, then it becomes likely that this intrinsic affinity limited the set of chemically accessible genetic codes. The application of statistical tests to further RNA–amino-acid interactions will ultimately indicate the relative importance of chance and necessity in determining the present form and content of the genetic code.

Acknowledgements

This research is supported in part by NSF grant MCB-9604377 to L.F.L., a Burroughs Welcome Fund New Investigator in Molecular Parasitology. We thank Jannette Carey, Michael Famulok, Stephen Freeland, Andrew Ellington and members of the Landweber Lab for comments and suggestions.

References

1. Woese, C.R. (1965). On the evolution of the genetic code. *Proc. Natl Acad. Sci. USA*, **54**, 1546-1552.
2. Sonneborn, T.M. (1965). Degeneracy of the genetic code: extent, nature, and genetic implications. In *Evolving Genes and Proteins.* (Bryson, V. & Vogel, H.J, eds), pp. 377-297, Academic Press, New York.
3. Zuckerkandl, E. & Pauling, L. (1965). Evolutionary divergence and convergence in proteins. In *Evolving Genes and Proteins.* (Bryson, V. & Vogel, H.J, eds), pp. 97–166, Academic Press, New York.
4. Woese, C.R. (1967). *The Genetic Code: The Molecular Basis for Genetic Expression.* Harper & Row, New York.
5. Crick, F.H.C. (1968). The origin of the genetic code. *J. Mol. Biol.* **38**, 367-379.
6. Wong, J.T.-F. (1975). A co-evolution theory of the genetic code. *Proc. Natl Acad. Sci. USA*, **72**, 1909-1912.
7. Woese, C.R., Dugre, D.H., Saxinger, W.C. & Dugre, S.A. (1966). The molecular basis for the genetic code. *Proc. Natl Acad. Sci. USA*, **55**, 966-974.
8. Pelc, S.R. & Welton, M.G.E. (1966). Stereochemical relationship between coding triplets and amino acids. *Nature* **209**, 868-872.
9. Dunnill, P. (1966). Triplet nucleotide–amino acid pairing: a stereochemical basis for the division between protein and nonprotein amino acids. *Nature* **210**, 1267-1268.
10. Alberti, S. (1997). The origin of the genetic code and protein synthesis. *J. Mol. Evol.* **45**, 352-358.
11. Yarus, M. (1989). Specificity of arginine binding by the *Tetrahymena* intron. *Biochemistry* **28**, 980-988.
12. Hopfield, J.J. (1978). Origin of the genetic code: a testable hypothesis based on tRNA structure, sequence, and kinetic proofreading. *Proc. Natl Acad. Sci. USA* **75**, 4334-4338.
13. Shimizu, M. (1982). Molecular basis for the genetic code. *J. Mol. Evol.* **18**, 297-303.
14. de Duve, C. (1988). The second genetic code. *Nature* **333**, 117-118.
15. Noller, H.F. (1993), On the origin of the ribosome: coevolution of subdomains of tRNA and rRNA. In *The RNA World*. (Gesteland, R.F. & Atkins, J.F., eds), pp. 137-156, Cold Spring Harbor Laboratory Press, New York.
16. Schimmel, P., Giege, R., Moras, D. & Yokoyama, S. (1993). An operational genetic code for amino acids and possible relationship to genetic code. *Proc. Natl Acad. Sci. USA* **90**, 8763-8768.
17. Maizels, N. & Weiner, A.M. (1994). Phylogeny from function: evidence from the molecular fossil record that tRNA originated in replication, not translation. *Proc. Natl Acad. Sci. USA* **91**, 6729-6734.
18. Dick, T.P. & Schamel, W.W.A. (1995). Molecular evolution of transfer RNA from two precursor hairpins: Implications for the origin of protein synthesis. *J. Mol. Evol.* **41**, 1-9.
19. Ellington, A.D. & Szostak, J.W. (1990). *In vitro* selection of RNA molecules that bind specific ligands. *Nature* **346**, 818-822.
20. Tuerk, C. & Gold, L. (1990). Systematic evolution of ligands by exponential enrichment: RNA ligands to bacteriophage T4 DNA polymerase. *Science* **249**, 505-510.
21. Robertson, D.L. & Joyce, G.F. (1990). Selection *in vitro* of an RNA enzyme that specifically cleaves single-stranded DNA. *Nature* **344**, 467-468.
22. Connell, G.J., Illangsekare, M. & Yarus, M. (1993). Three small ribooligonucleotides with specific arginine sites. *Biochemistry* **32**, 5497-5502.
23. Connell, G.J. & Yarus, M. (1994). RNAs with dual specificity and dual RNAs with similar specificity. *Science* **264**, 1137-1141.

24. Yarus, M. (1993), An RNA-amino acid affinity. In *The RNA World.* (Gesteland, R.F. & Atkins, J.F., eds), pp. 205-217, Cold Spring Harbor Laboratory Press, New York.
25. Famulok, M. (1994). Molecular recognition of amino acids by RNA-aptamers: an L-citrulline binding RNA motif and its evolution into an L-arginine binder. *J. Am. Chem. Soc.* 116, 1698-1706.
26. Tao, J. & Frankel, A.D. (1996). Arginine-binding RNAs resembling TAR identified by *in vitro* selection. *Biochemistry* 35, 2229-2238.
27. Geiger, A., Burgstaller, P., von der Eltz, H., Roeder, A. & Famulok, M. (1996). RNA aptamers that bind L-arginine with sub-micromolar dissociation constants and high enantioselectivity. *Nucleic Acids Res.* 24, 1029-1036.
28. Tao, J. & Frankel, A.D. (1992). Specific binding of arginine to TAR RNA. *Proc. Natl Acad. Sci. USA* 89, 2723-2726.
29. Yang, Y., Kochoyan, M., Burgstaller, P., Westhof, E. & Famulok, F. (1996). Structural basis of ligand discrimination by two related RNA aptamers resolved by NMR spectroscopy. *Science* 272, 1343-1346.
30. Sokal, R.R. & Rohlf, F.J. (1995). *Biometry: the Principles and Practice of Statistics in Biological Research.* (3rd edn). W.H. Freeman and Company, New York.
31. Jiang, F., Kumar, R.A., A, J.R. & Patel, D.J. (1996). Structural basis of RNA folding and recognition in an AMP–RNA aptamer complex. *Nature* 382, 183-186.
32. Fan, P., Suri, A.K., Fiala, R., Live, D. & Patel, D.J. (1996). Molecular recognition in the FMN–RNA aptamer complex. *J. Mol. Biol.* 258, 480-500.
33. Burgstaller, P., Kochoyan, M. & Famulok, M. (1995). Structural probing and damage selection of citrulline- and arginine-specific RNA aptamers identify base positions required for binding. *Nucleic Acids Res.* 23, 4769-4776.
34. Jiang, L., Suri, A.K., Fiala, R. & Patel, D.J. (1997). Saccharide–RNA recognition in an aminoglycoside antibiotic–RNA aptamer complex. *Chem. Biol.* 4, 35-50.
35. Zimmermann, G.R., Shields, T.P., Jenison, R.D., Wick, C.L. & Pardi, A. (1998). A semiconserved residue inhibits complex formation by stabilizing interactions in the free state of a theophylline-binding RNA. *Biochemistry* 37, 9186-9192.
36. Yarus, M. (1998). Amino acids as RNA ligands: a direct-RNA-template theory for the code's origin. *J. Mol. Evol.* 47, 109-117.
37. Majerfeld, I. & Yarus, M. (1994). An RNA pocket for an aliphatic hydrophobe. *Nat. Struct. Biol.* 1, 287-292.
38. Majerfeld, I. & Yarus, M. (1998). Isoleucine:RNA sites with essential coding sequences. *RNA* 4, 471-478.
39. Zinnen, S. & Yarus, M. (1995). An RNA pocket for the planar aromatic sidechains of phenylalanine and tryptophane. *Nucleic Acids Symp. Ser.* 33, 148-151.
40. Jukes, T.H. (1973). Arginine as an evolutionary intruder into protein synthesis. *Biochem. Biophy. Res. Commun.* 53, 709-714.
41. Grantham, R. (1974). Amino acid difference formula to help explain protein evolution. *Science* 185, 862-865.

2.3 Guilt By Association: The Arginine Case Revisited

This chapter was written in response to criticisms that the codon/binding association was an artifact of the particular set of sequences that had happened to be chosen for chemical mapping. In order to test whether the association was robust, I tested what would have happened had other aptamers been chosen from the same set of sequences selected in each lab, and tested alternative sets of binding sites suggested by the main opponent and main proponent of the direct templating hypothesis, these being Andy Ellington and Mike Yarus respectively. I found that the association between codons and amino acid binding sites, at least for arginine, was surprisingly robust to even quite unreasonable assumptions. Subsequent evidence from other amino acid binders(see Chapter 2.1) suggests that arginine is not unique in this respect, although it remains the best example of a codon/amino acid association.

The chapter appeared in the April 2000 issue of RNA*:*

Knight, R. D. and L. F. Landweber (2000). "Guilt by association: the arginine case revisited." RNA 6(4): 499-510.

I wrote this paper while visiting Prof. Yarus's lab late in 1999, in what turned out to be an only slightly successful attempt to select my own aptamers to a protein-coding amino acid, isoleucine, and a prebiotic amino acid not normally found in proteins, norleucine, to compare the strategies that RNA would use to distinguish these structural isomers (a selection for leucine, another isomer, was already in progress) – see Chapter 2.1 for results. Prof. Yarus commented on several drafts of the paper, providing many useful insights, and helped greatly with the section on the various models for the origin of translation.

RNA (2000), *6*:499–510. Cambridge University Press. Printed in the USA.

PERSPECTIVE

Guilt by association: The arginine case revisited

ROBIN D. KNIGHT and LAURA F. LANDWEBER
Department of Ecology and Evolutionary Biology, Princeton University, Princeton, New Jersey 08544-1003, USA

ABSTRACT

If the genetic code arose in an RNA world, present codon assignments may reflect primordial RNA–amino acid affinities. Whether aptamers selected from random pools to bind free amino acids do so using the cognate codons at their binding sites has been controversial. Here we defend and extend our previous analysis of arginine binding sites, and propose a model for the maintenance of codon–amino acid interactions through the evolution of amino acids from ribozyme cofactors into the building blocks of proteins.

Keywords: aptamer; chemical evolution; genetic code; RNA world; SELEX

INTRODUCTION

Inferences about the evolution of core metabolic functions are difficult, because billions of years of evolution separate today's elaborate cellular processes from their chemical origins. The genetic code is especially problematic: not only has it been unclear whether there is an underlying chemical basis linking codons to their cognate amino acids (Gamow, 1954; Woese et al., 1966; Lacey & Mullins, 1983; Yarus & Christian, 1989; Lacey, 1992; Yarus, 1998), but the "universal" genetic code actually has changed in numerous lineages (Osawa, 1995). Similar changes in the genetic code prior to the Last Universal Ancestor, perhaps to add amino acids or to minimize genetic error, could easily have altered beyond recognition any primordial, chemically determined code (Crick, 1968; Wong, 1975; Freeland & Hurst, 1998a; Knight et al., 1999).

In modern translation, the meaning of a particular codon can easily be altered by mutating the tRNA (Osawa & Jukes, 1989; Schultz & Yarus, 1994). This is because the ribosome matches codons with anticodons, but never checks whether tRNAs are correctly charged. The responsibility for linking specific amino acids with specific codons (via the anticodon-bearing tRNAs) lies with a group of enzymes, the aminoacyl-tRNA synthetases, which can employ sophisticated proofreading mechanisms but do not always recognize the anticodon specifically (Ibba & Soll, 1999). Thus the present code is somewhat malleable. In this context, that aptamers selected to bind arginine seem to do so using the canonical set of arginine codons (Knight & Landweber, 1998) is surprising, especially because modern tRNA contains the anticodons instead. If a codon/binding site association implies direct templating of protein synthesis (Yarus, 1998), why would tRNA evolve as an intermediary between template and peptide?

Reprint requests to: Laura F. Landweber, Department of Ecology and Evolutionary Biology, Princeton University, Princeton, New Jersey 08544-1003, USA; e-mail: lfl@princeton.edu.

Are stereochemical associations an artifact?

One possibility is that this codon/binding site association is an artifact. Diverse RNA sequences can perform the same task: in SELEX experiments, dissimilar molecules survive many cycles of harsh selection (Geiger et al., 1996). Few of these sequences are ever further characterized. Consequently, it is possible to choose post-hoc from the same experiments a set of sequences that either does or does not show any particular desired property (Ellington et al., 2000). An association need not be *universal* for prebiotic relevance; however, to influence primordial codon assignments, the association between trinucleotide sequences and binding sites need only occur more often than chance (we assume that the conditions in SELEX recapture to some extent the conditions in the RNA world). Here we test whether the apparent association between arginine codons and arginine binding sites is a statistical quirk of the particular sequences chosen for characterization in each case.

Criteria for choosing aptamers

Ellington et al.'s analysis (Ellington et al., 2000), which suggests only negligible association between arginine

codons and binding sites, differs from our previous analysis (Knight & Landweber, 1998) in both choice of included nucleotides and choice of included sequences. Our choice of sequences was very simple: we included all and only those sequences that (1) had been selected from randomized pools to bind free arginine, and (2) had structural data available. This choice was *a priori* unbiased with respect to whether these particular sequences had greater or lesser codon/binding site associations than others that might alternatively be chosen. We excluded two natural arginine binders, TAR and the group I intron, because they are shaped by selection in organisms for functions other than binding free arginine. In the case of TAR, this is demonstrated by the fact that the K_d for free arginine is halved by substituting a G:C base pair for the natural A:U base pair immediately 3′ of the bulge (Tao & Frankel, 1996). If TAR were really under selection to optimize free arginine binding, nature should have effected this minor change already.

Two of the aptamers, that from Connell and Yarus (1994) and clone 16 from Tao and Frankel (1996), have binding sites that overlap or base pair to the constant region. It can be argued that these sequences should be excluded from the analysis on the grounds that the binding sites were not really selected from random sequence (Ellington et al., 2000). Inclusion or exclusion of these two sequences does not qualitatively affect the results (see below), although the reduced sample size lowers the significance slightly. A separate but related issue is that constant regions might influence the apparent association. This includes primers, and the aptamer used for NMR structure determination (Yang et al., 1996) also had an extra nonbinding-site CGA Arg codon at positions 41–43 not present in Famulok's original aptamer, introduced for convenience to create a GAAA tetraloop (M. Famulok, T. Hermann, and E. Westhof, pers. comm.). However, inclusion or exclusion of these parts of the sequence does not actually affect the results greatly (summarized in Table 3).

Statistical methodology

To assess whether an association exists between binding sites and particular codon sets, we use the G test for independence for a 2×2 contingency table (Sokal & Rohlf, 1995) partitioning nucleotides into classes based on whether or not they are found in codons and whether or not they are found in binding sites. If the two variables are independent, then the probability of observing a particular state in one variable is unaffected by the state of the other variable. Thus, of the nucleotides in binding sites, the fraction of nucleotides both in binding sites and in codons should be proportional to the fraction of nucleotides in codons overall. The same applies to the other three possible combinations of states, so the expected number of occurrences of any particular outcome is the total number of nucleotides multiplied by the fraction of nucleotides in the appropriate row and column.

These expected values are compared against the actual counts in each cell, and a test statistic (G) is computed and placed within a known probability distribution (χ^2 with 1 degree of freedom). This yields a *p* value, which is the probability of finding an equal or greater discrepancy between observed and expected values by chance if the variables really were independent. The test is 2-tailed, but we are interested only in those motifs that give high G values because they occur surprisingly often (rather than surprisingly rarely) at binding sites, making our prediction directional. This can be tested by measuring whether the fraction of nucleotides in codons is greater for nucleotides inside binding sites, in which case we assign G a positive value; otherwise, G is negative. Because the prediction is directional, the *p* value for a given G is halved if G is positive (and therefore the discrepancy is in the direction predicted); otherwise it is halved and subtracted from 1. Thus, we tested whether nucleotides at binding sites were disproportionately likely to be included in arginine codons (or, equivalently, whether nucleotides included in arginine codons were disproportionately likely to be found at binding sites).

In this article, we exhaustively test the other codon sets to ensure that none has as great an association with arginine binding sites as does the Arg set. This goes beyond the original hypothesis (that Arg codons will associate with arginine sites), but raises the possibility that, because of the larger number of comparisons, some will appear to be significant just by chance. We correct for multiple comparisons by finding the probability that at least one codon set will show an association of a particular magnitude: this relation is given by $\alpha' = 1 - (1 - \alpha)^n$, where α is the original probability of a type I error (rejecting the null hypothesis when it is true), α' is the corrected probability, and n is the number of comparisons (21 codon sets, including termination). The probability that at least one amino acid exceeds a significance level α, and that the amino acid that does so is arginine, is given by α'/n.

The method is attractive because it makes relatively few assumptions, the main one being that all nucleotides form an equivalent population. This is reasonable when sequences are selected from a pool that is initially random at all positions, as we can compare the sequences that survive the selection to the set of all sequences of the same length. It could be argued that this method does not take into account biases in nucleotide composition in and out of binding sites (Ellington et al., 2000). This would imply that the composition but not the order of bases in a sequence must systematically affect the likelihood that it will survive the selection process. There is also a causality problem: biased

composition at binding sites influences the probability that codons will appear there, but, conversely, selection for particular codons at binding sites would influence their composition.

Biased sequence composition does not explain the association

There are three ways to address nucleotide bias; we use two of these to show that nucleotide bias cannot explain the association, and explain why the third is less appropriate for this situation. The first is to perform a Monte Carlo simulation comparing the actual sequences with randomly generated sequences with the same nucleotide frequencies inside and outside binding sites. The second is to test whether permuted codons associate with binding sites as well as do the actual codons. The third is to replicate the test for independence separately for each nucleotide.

The first approach compares the actual sequences to the set of sequences with the same composition, directly estimating the probability of a particular level of association. For each replicate, we randomized the order of nucleotides in each sequence, maintaining the number of each type of base inside and outside binding sites. We then calculated G values for each codon set with the randomized binding sites, repeating the process 1,000,000 times per run (Table 1). Using our original set of sequences and binding sites, only 10 random sequence sets show a higher association between Arg codons and arginine binding sites than does the actual sequence set (Fig. 1) (Knight & Landweber, 1998); using Ellington's suggested binding sites and excluding constant regions and the two sequences in which the binding sites overlap the constant regions (see below) 4,699 random sets showed higher association ($p < 0.005$), still more than an order of magnitude fewer than the closest competitor. This shows that compositional bias in aptamers and their binding sites cannot explain the association.

The second approach tests whether permuted codons show as high an association with binding sites as do the actual codons, as would be predicted if composition bias were important. For each of the codon blocks CGN and AGR, there are six possible permutations, leading to 36 combinations of the two codon blocks. Table 2 shows G values for all the possible permutations using the sequences and binding sites from our original article (Knight & Landweber, 1998), clearly showing that the actual Arg codon set has a far greater association with arginine binding sites than does any permuted set, even when those permuted sets still contain a significant subset of Arg codons. Thus, the results cannot be explained by sequence-independent compositional bias of Arg codons.

The third approach, replicating the test for independence on a nucleotide-by-nucleotide basis, assumes that if there is an association between nucleotides that are in codons and nucleotides that are in binding sites then the association should hold for each of U, C, A, and G separately (Ellington et al., 2000). Unfortunately, this approach has both practical and theoretical limitations. Practically, the G test is overly conservative when expected counts in cells are below 5. This can be partially ameliorated by using Fisher's Exact Test, which avoids the poor fit between the G statistic and the χ^2 distribution when cell counts are small, but the fundamental problem is that partitioning the data into four times as many classes increases the chance of accepting the null hypothesis when it is false (a type II error).

Although this test reduces the effective sample sizes by a factor of four, the cost in statistical power would be acceptable if it provided further useful information. However, this is not the case. It is neither necessary nor sufficient that U, C, A, and G at binding sites all associate with codons to show an overall association between codons and binding sites. Considering each nucleotide separately is the wrong level of analysis because it misses the forest for the trees: the key issue is whether *codons*, not the nucleotides that comprise them, associate with binding sites. For instance, U is only found in one of six arginine codons (CGU). Even if U by

TABLE 1. Codon/binding site associations using actual and randomized sets of sequences, using either our original sequences and binding sites (Knight & Landweber, 1998) or Ellington et al.'s (2000) binding sites and excluding constant regions and disputed sequences.

	Knight and Landweber (1998)		Ellington et al. (2000)	
Amino acid	G (corr.)[a]	#G > obs[b]	G (corr.)[a]	#G > obs[b]
Stop	0.28	396414	0.47	311180
A	−0.28	524951	−1.44	801595
C	−9.50	999757	−7.38	990465
D	1.87	99386	0.47	267539
E	0.06	816885	0.47	575553
F	−3.58	959121	−1.15	657947
G	0.21	572598	−0.17	682082
H	−1.32	818412	−2.61	971218
I	−0.01	544318	0.97	135176
K	1.80	370003	5.11	89877
L	−10.54	988548	−5.73	936297
M	−0.26	704474	−1.52	910693
N	0.49	413675	2.24	210977
P	−4.48	918922	−0.64	573271
Q	0.24	420356	2.24	176741
R	19.99	10	8.38	4699
S	−2.10	789413	0.01	307247
T	0.26	294269	2.41	89996
V	1.50	23615	0.00	213132
W	−1.73	829197	−4.14	964768
Y	−1.07	785082	−1.15	851841

[a] Actual G value for each amino acid codon/binding site association.
[b] Number of randomized sequence sets in a sample of 1 million that show a greater association than actually observed.

The slight difference between the G reported here for arginine (20.0) and that reported previously (20.2) is due to exclusion of two undetermined nucleotides, counted as noncodon and nonbinding in our original paper.

FIGURE 1. Five independent classes of arginine aptamers. (**a**) Yang et al. (1996), (**b**) Connell and Yarus (1994), (**c**) Geiger et al. (1996), (**d**) Tao and Frankel (1996) clone 2, (**e**) Tao and Frankel (1996) clone 16. Arginine codons are red (AGR) and green (CGN). Grey highlight indicates nucleotides considered as binding in our original analysis (Knight & Landweber, 1998). Additional nucleotides have been suggested as binding site nucleotides by Ellington et al. (yellow highlight), Yarus (blue highlight), or both (green highlight). Constant regions not randomized in the selection are underlined. Also shown are other aptamers with known NMR structures: (**f**) AMP, Jiang et al. (1996), (**g**) citrulline, Yang et al. (1996), (**h**) FMN, Fan et al. (1996), (**i**) tobramycin aptamer 1, Jiang et al. (1997), (**j**) Theophylline, Zimmermann et al. (1998), (**k**) neomycin B, Jiang et al. (1999), (**l**) tobramycin aptamer 2, Jiang & Patel (1998).

itself does not show any codon/binding site association, CGU codons may well do so due to associations of CG doublets generally: evolution is blind to whether the association applies to each base individually, as long as it applies to the codon as a whole. It is possible to construct sequences for which there is a codon/binding site association for each of the four nucleotides but not for nucleotides overall, and vice versa (data not shown). This is an example of Simpson's Paradox, which demonstrates the importance of examining phenomena at the correct level.

TABLE 2. Codon/binding site associations for permuted arginine codons, using the original Knight and Landweber (1998) binding sites.

	AGR	ARG	RAG	RGA	GRA	GAR
CGN	***20.0***	***12.6***	***6.9***	***6.4***	*1.7*	*4.1*
CNG	*0.7*	−0.1	−0.4	−0.3	−1.8	−0.4
NCG	***8.7***	4.2	0.6	1.2	0.0	0.2
NGC	*0.8*	0.0	−1.3	−1.0	−3.0	−1.6
GNC	***10.5***	**6.8**	4.0	2.7	1.3	3.5
GCN	*2.9*	0.7	−0.2	−0.1	−1.1	−0.3

All values are corrected G values for the set of codons given by the row and column, that is, the first entry is CGN + AGR. Significant G values are shown in bold; the greatest association in each row and in each column is shown in italics. Note that the actual Arg codon set (first cell) shows by far the greatest association, and that the sets containing one of the Arg blocks (CGN, the first row, or AGR, the first column) always show stronger associations than do comparable sets that lack one of these blocks.

Testing Arginine/Arg codon associations

Given the original set of 5 sequences (Fig. 1), the differences in choice of binding sites proposed by Ellington et al. (2000) and by Yarus (2000) affect the exact level of significance of the associations, but do not qualitatively change the result that the arginine codon/binding site association is highly significant, and that no other codon set associates with arginine sites as well as the arginine codon set when the original sequences are considered (Table 3). The Ala and Glu codon sets are never significantly overrepresented though they have the same composition as the two blocks of Arg codons, indicating that nucleotide bias cannot explain this finding. G63 in the Geiger et al. (1996) aptamer is described as "weakly protected," though this protection is unconvincing (M. Yarus, unpubl. data). Inclusion or exclusion of this nucleotide affects the results only minimally. Yarus (2000) discusses the concordance between different methods for determining binding sites, and finds that, in general, each method implicates the same nucleotides. Therefore, although chemical modification and nuclease mapping do not provide incontrovertible struc-

TABLE 3. Codon/binding site associations using different sets of sequences and binding sites.

				Ellington									
	Yarus			Figure 4			Figure 6				Knight & Landweber		
Amino acid	all	C	CD	all	C	CD	all	N	CD	CD-N	all	C	CD
Stop	0.33	−0.02	0.21	1.77	0.55	0.47	**8.35**	5.30	1.99	0.84	0.28	−0.01	0.00
A	−1.58	−0.46	−2.59	−1.73	−0.15	−1.44	−1.02	−0.37	−0.07	0.00	−0.28	0.02	−0.42
C	−4.83	−19.12	−12.32	−6.39	−12.72	−7.38	−0.78	−0.39	−3.75	−2.98	−9.50	−10.37	−5.34
D	1.55	0.27	0.38	1.77	0.50	0.47	1.52	0.84	0.26	0.14	1.87	0.49	0.49
E	1.38	3.44	1.29	1.14	2.37	0.47	1.88	0.46	1.71	0.16	0.06	0.22	−0.42
F	−2.38	−1.71	−2.00	−0.95	−1.09	−1.15	−2.45	−2.08	−1.03	−0.80	−3.58	−0.86	−0.80
G	0.01	0.35	−0.13	0.04	0.30	−0.17	1.04	1.03	0.26	0.14	0.21	0.40	−0.05
H	−4.72	−4.85	−4.43	−2.27	−2.32	−2.61	−8.02	−6.87	−6.18	−4.94	−1.32	−1.49	−1.86
I	2.37	2.07	6.29	0.31	0.03	0.97	−0.22	−0.03	−0.12	0.00	−0.01	−0.01	0.62
K	0.31	0.86	0.47	3.67	5.33	5.11	5.59	1.97	5.91	2.21	1.80	2.34	2.21
L	−9.08	−12.70	−9.59	−13.49	−8.44	−5.73	−0.68	−0.19	−1.83	−0.90	−10.54	−6.87	−4.14
M	−1.79	−4.85	−4.16	−0.64	−2.32	−1.52	−0.35	−0.18	−1.20	−0.65	−0.26	−1.49	−0.65
N	0.88	4.56	6.38	0.64	2.75	2.24	0.40	0.64	2.50	2.92	0.49	1.80	0.95
P	−1.64	−0.94	0.00	−6.67	−2.67	−0.64	−11.20	−9.60	−3.75	−2.98	−4.48	−1.75	−0.18
Q	−0.10	0.11	0.76	0.31	1.05	2.24	0.98	0.19	**7.78**	2.92	0.24	0.45	0.95
R	**26.19**	**32.44**	**24.98**	**11.88**	**13.14**	**8.38**	5.03	**5.79**	6.53	**7.46**	**19.99**	**19.47**	**15.18**
S	−2.85	−3.40	−1.04	−1.43	−0.69	0.01	−0.28	0.00	0.14	0.84	−2.10	−0.90	0.00
T	0.00	1.21	2.07	0.13	1.36	2.41	0.00	−0.07	−0.01	−0.19	0.26	1.24	1.56
V	2.72	−0.17	−0.57	3.26	−0.02	0.00	4.49	4.34	0.80	0.84	1.50	−0.13	−0.04
W	−3.46	−5.95	−6.96	−2.27	−3.92	−4.14	−1.53	−1.29	−2.36	−1.86	−1.73	−3.17	−2.98
Y	−2.19	−1.71	−2.00	−1.41	−1.09	−1.15	−2.45	−2.08	−3.75	−2.98	−1.07	−0.86	−0.80

Yarus: set of sequences and nucleotides from Yarus (2000). Ellington: set of sequences and binding sites from Ellington et al. (2000), Figure 4 [alternative binding sites for sequences analyzed in Knight and Landweber (1998)] and Figure 6 (alternative sequences to those characterized). Knight and Landweber: sequences and binding sites analyzed in Knight and Landweber (1998). C: constant regions excluded. D: disputed sequences (Connell and Yarus (1994), and Tao and Frankel (1996) clone 16) excluded. N: nucleotides implicated by chemical modification but shown by NMR not to participate in binding counted as nonbinding. In all cases, the arginine codon set shows the greatest association, except in the set from Ellington Figure 6. However, failure of arginine codons to show the greatest association in this set is an artifact of including bases shown by NMR not to be in the binding site (compare columns marked "all" with adjacent columns marked "N"). Significance cutoffs (1-tailed test corrected for 21 multiple comparisons): $p = 0.05$, G = 7.92; $p = 0.01$, G = 10.91; $p = 0.001$, G = 15.22. Note: these are the probabilities that at least one codon set would show a greater or equal association; divide by 21 for the probability that the particular codon set showing the association is arginine.

tural information, we can be fairly confident that the method is useful for this type of analysis.

Arg codon/site association is not an artifact of aptamer sampling

There remains the question of whether this codon/binding site result is general, or a statistical fluke resulting from the particular choice of sequences for characterization. Although this choice was presumably random with respect to whether a codon/binding site association would eventually be observed from each sequence, it is possible that the chosen sequences were not representative of the entire population of arginine binders. Indeed, Ellington et al. (2000) choose a different set of sequences from the same experiments (their Fig. 6), make the assumption that they have the same structures, and show that the resulting G value is greatly reduced (see Fig. 2 for alignments). Therefore the question is whether the set Ellington et al. (2000) chose for analysis is more representative of the data as a whole than the set we used, because—while the arginine codon/binding site association is still significant—this choice affects the apparent improbability of the association by several orders of magnitude.

To test what would be expected to occur on average, had other sequences been chosen, we assumed (as Ellington et al. do) that the alternative published sequences share the same binding site nucleotides as the sequence actually characterized. This gives a choice of 2 sequences from Connell and Yarus (1994), 3 sequences for Geiger et al. (1996) (the remaining sequences have no obvious similarity, so the binding sites cannot be inferred), 12 sequences for Yang et al. (1996) that were previously selected by Famulok (Famulok, 1994), 22 sequences for TAR-like clone 16 and 19 sequences for clone 2 from Tao and Frankel (1996) (Fig. 2), for a total of 30,096 possible combinations taking one from each family (15,048 if the Connell and Yarus sequence is excluded; 684 if the Tao and Frankel clone 16 is also excluded).

We tested every possible combination for codon/binding site association for each of the 21 codon sets,

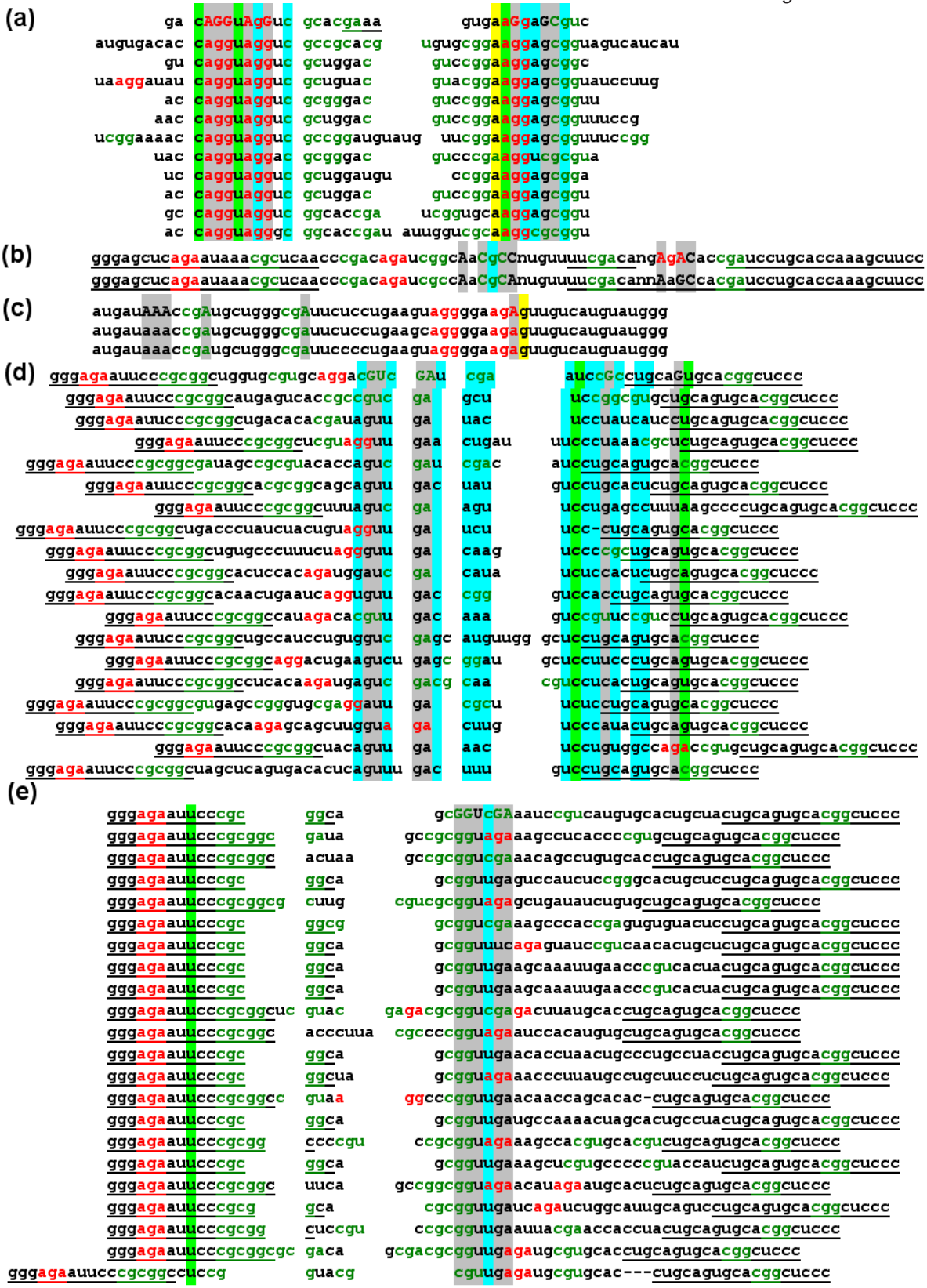


FIGURE 2. Alternative aptamers from each experiment. First aptamer in each set, numbering, and coloring of binding sites are as in Figure 1.

with or without the disputed sequences and constant regions and using either Ellington et al.'s, Yarus's, or our original choice of binding site nucleotides. Using our original set of sequences and binding sites (Fig. 3A), the mean G value for Arg codon/site associations is 8.21, indicating that on average we would still expect to see a significant association (the significance cutoff for $p < 0.05$ after correcting for 21 multiple comparisons is

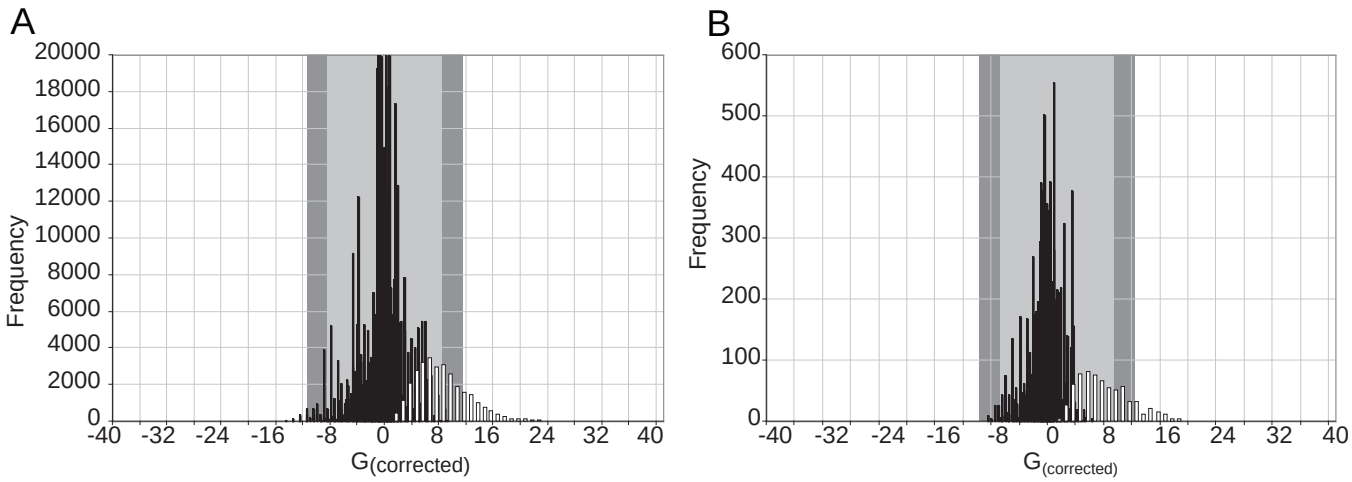


FIGURE 3. Distributions of codon/binding site G values for all codon sets. The light grey region indicates $p < 0.05$ after correcting for 21 comparisons; the dark grey region indicates $0.05 < p < 0.01$. The distribution for the Arg codon set (white) clearly differs from all other codon sets. **A**: Sequences and binding sites as in Knight and Landweber (1998). **B**: Binding sites as for **A**, but excluding constant regions and sequences where the constant regions overlap the binding sites.

7.92). No other codon set comes close, the nearest being the set of termination codons (G = 3.98). Although this mean value is much lower than the value of 20.0 for the actual sequence set, it is still impressive given the tenuous assignment of nonconserved nucleotides to binding sites based only on their position in the molecule. If we exclude constant regions and the two disputed sequences (Fig. 3B) the average G value for Arg increases to 8.86, indicating that these potentially problematic regions do not materially affect the result (the mean is 8.39 if we include the disputed sequences but exclude the constant regions). For this set, the next highest mean G value is 2.32, for Lys.

Ellington et al.'s choice of binding sites decreases the mean G value somewhat (mean of 5.52 for Arg if the disputed sequences are included; 4.68 without them), while Yarus's choice of binding sites increases it (mean of 9.76 with the disputed sequences; 8.36 without them). Note that for a single comparison (i.e., testing the specific prediction that Arg codons associate with arginine binding sites) a G value of 4.68 (with 1 degree of freedom) still corresponds to a 1-tailed *p* value of 0.015. Thus even if we picked a set of sequences at random and assumed that the binding sites in uncharacterized sequences were the same as those in characterized sequences, we would still expect to see the association.

Two data sets individually include sufficient sequences for meaningful analyses: Tao and Frankel's (1996) two classes of arginine binders and Famulok's (1994) arginine and citrulline binders. Because the sequences within a set are not independent, the magnitude of the G value is greatly inflated and is itself not meaningful. However, the relative G values are illustrative: the arginine codon set has by far the greatest codon/binding site association in Famulok's arginine, but not his citrulline, aptamers (Fig. 4A). In contrast, arginine does not show particularly strong codon/binding site associations in either of the classes of Tao and Frankel's aptamers (Fig. 4B), and had these been the only aptamers available, the general codon/binding site association for arginine aptamers would not have been apparent. This underscores the importance of analyzing multiple, independent data sets.

Other small-ligand sites do not overrepresent Arg codons

We also tested the set of aptamers to small-molecule ligands other than arginine for which NMR data are now available [Fig. 1(f–l)]. Comparing the codon/binding site associations for the nonarginine set to that for the set of aptamers using the most inclusive set of binding sites (Yarus's) (Fig. 4C) shows that the association between arginine codons and binding sites is not general to small-molecule ligands, but applies only and strikingly to arginine.

Furthermore, the association between codons and binding sites appears to go beyond arginine. Both isoleucine and tyrosine, the only other two amino acids for which strong, specific RNA binders are known, have conserved cognate codons at their binding sites. When the data from these selections are pooled, the evidence against the null hypothesis of no codon/binding site association is even more overwhelming (Yarus, 2000). Interestingly, the two DNA aptamers to arginine for which structures have been determined by NMR (Lin et al., 1998; Robertson et al., 2000) extend this codon/binding site association beyond RNA. For these two sequences, tgacc<u>agggcaaACGgtAGG</u>Tgagtggtca and <u>cg</u>accAA<u>CGT</u>nn<u>CGC</u>CTggtc<u>g</u> (binding sites capitalized; Arg codons underlined), the G value for Arg

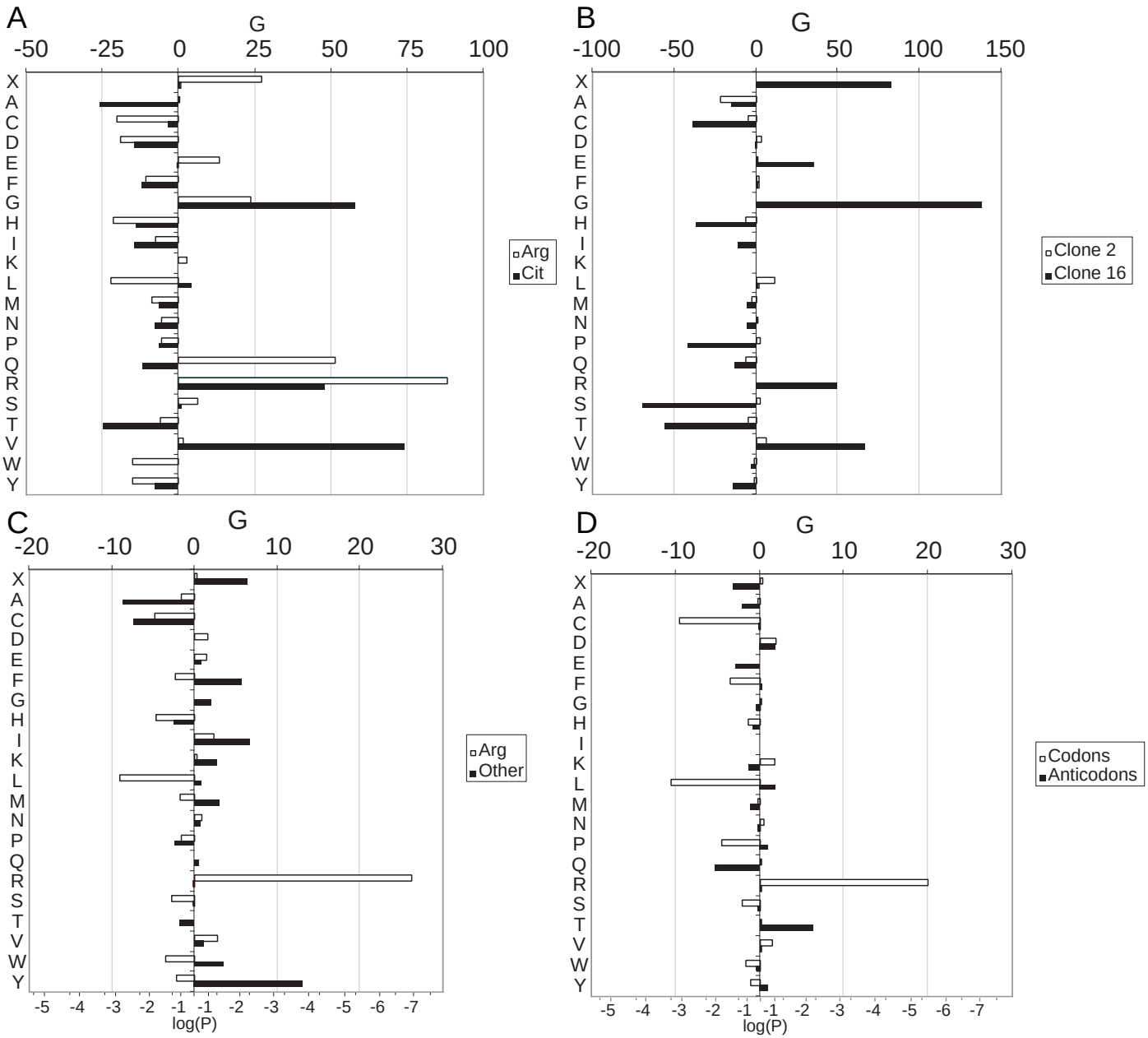


FIGURE 4. Codon/binding site association values for (**A**) Famulok (1994) arginine (white) and citrulline (black) binders; (**B**) Tao and Frankel (1996) arginine binders in the families of Clone 2 (white) and Clone 16 (black); (**C**) the set of arginine aptamers using Yarus's (2000) binding sites (white) and the set of non-arginine aptamers for which NMR structures are available (black); and (**D**) codons and anticodons using the original set of aptamers and binding sites. Note that the arginine codon set has by far the highest association with arginine sites in both **B** and **C**, but comes in third for citrulline sites and third-to-last for non-arginine sites generally. The Tao and Frankel (1996) set, taken alone, does not show Arg codon/binding site associations, showing the importance of considering multiple independent sites. There is no association between the set of arginine anticodons and arginine binding sites. Bottom scale in **C** and **D** gives the log of the probability that a particular codon set would show as great or greater an association, calculated as follows: the probability that *any* of *n* codon sets shows a given level of association *p* just by chance is $p' = 1 - ((1 - p)^{21})$. The probability that a *particular* codon set (e.g., the Arg set) shows such an association is p'/n. Large ticks are 10^x; small ticks are 5×10^x.

codon/binding site association is 8.82 ($p = 0.0015$), and is much higher than that for any other codon set. This implies that the information lies among the bases and that the choice of backbone itself is not critical, which should assure those who doubt the plausibility of prebiotic RNA synthesis (Lazcano & Miller, 1996).

DISCUSSION

We have shown that the statistical association between codons and binding sites cannot easily be explained away by the choice of particular sequences for analysis. In the absence of a plausible model, however,

these individual observations of fact have no context. The model must explain why associations between amino acids and their codons should be preserved from the RNA world, maintaining informational continuity.

There are two pathways that could link motifs in primordial binding sites to anticodons in modern tRNAs: the anticodons could be either the descendants of the sites themselves or the descendants of sequences that recognized those sites by complementary base pairing. However, although there is an association between arginine codons and their binding sites, there is clearly no such association with the anticodons (Fig. 4D). We compare four models, Szathmáry's CCH (Coding Coenzyme Handle) (Szathmáry, 1993, 1999), Yarus's DRT (Direct RNA Templating) (Yarus, 1998), Ellington's oligopeptide ligation model (Ellington et al., 2000), and a modified form of CCH in the light of this evidence.

DRT (Fig. 5A) proposes that translation evolved from amino acid binding sites similar to those found in modern SELEX experiments, each of which binds a single amino acid using multiple codons. Accumulation of several of these sites in a single molecule might promote oligopeptide formation, perhaps by bringing the amino and carboxyl groups of successive residues into proximity. Although these sites would have been selected to recognize specific amino acids, they would also recognize (by base pairing) specific RNA sequences to which they were complementary. Evolution of a *trans*-aminoacylating activity by other ribozymes could transform such complementary sequences into adaptors like modern tRNAs, accepting amino acids and pairing with their binding sites to read out the message. Freed from the constraint of having to recognize the amino acids themselves, the sites on the message would become vestigial, eventually withering to single codons.

The main objection to DRT is that it requires a discontinuity at the point at which adaptors take over from direct templating. Furthermore, it requires that each residue in a peptide be encoded by a large RNA site, but the evolvability of such a system may be limited depending on how specificities are connected in sequence space. For instance, Famulok was able to select arginine binders that differed from citrulline binders by only three bases, but was unable to select any lysine, glutamine, or albizziin (an analog of citrulline) binders from the same pool (Famulok, 1994). There are also potential reading frame difficulties in shifting from many bases per amino acid to only three bases per amino acid. However, DRT does provide selective pressure for improving peptide synthesis from the very origin of the genetic code, and the principle of continuity may have preserved the templating rules established in such an early incarnation (Knight et al., 1999). It also correctly predicts the parity of the relationship between binding sites and tRNA.

In CCH (Fig. 5B), the genetic code arises before peptide synthesis, and the original function of amino acids is to act as cofactors for ribozymes. The genetic code could have linked amino acids to particular oligonucleotides, which could then base pair to ribozymes without an extensive and specific amino acid binding site. Later, RNA molecules would specialize into coding sequences and catalysts. However, known amino acid binding sites are much larger than a base triplet, and there is no evidence for direct binding between amino acids and trinucleotides in solution. Furthermore, if the coding coenzyme handles were the original amino acid binding sites and evolved into tRNA, the conserved motif in tRNA (the "anticodon") should be the same as the motif found at binding sites, which is not the case.

Ellington et al. (2000, see their Fig. 7) suggest an alternative model, in which the first peptide bond formation was between oligopeptides such as poly(Lys, Arg), which could have acted as stabilizing counterions to RNA. Ribozymes catalyzing this condensation could evolve sequence specificity, promoting the formation of particular desirable peptides. Evolution of the binding sites to recognize free amino acids would allow the construction of these peptides once the prebiotic supplies were depleted. These amino acid binding sites later became tRNAs. Positively charged amino acids, much less polymers thereof, are unlikely to have existed in appreciable quantities in prebiotic settings (Miller, 1987; Weber & Miller, 1981), although perhaps selection pressure existed for forming other peptides. This oligopeptide ligation model also shares with CCH the drawback that it predicts the wrong parity for the genetic code: if tRNAs evolved from amino acid binding sites, then those sites should have more anticodons than expected (even if paired with the complementary codons). This is not the case. Ellington et al. (2000) suggest that the genetic code began with a single amino acid and then expanded through tRNA diversification, and so perhaps the amino acid whose binding sites contain more anticodons than expected is yet to be found.

We suggest the following modified version of CCH (Fig. 5C): RNA sequences that were able to bind particular amino acids conferred some selective advantage, such as resistance to degradation, charge stabilization, or increased catalytic activity. These RNA sequences were statistically likely to overrepresent the sequences that became their cognate codons at the binding sites for these amino acids. Later, they evolved catalytic activity such that they were able to aminoacylate either themselves or other molecules; several aminoacylating ribozymes have recently been isolated (Illangasekare et al., 1995; Lohse & Szostak, 1996; Welch et al., 1997; Illangasekare & Yarus, 1999). There may have been some noncoded or multi-ribozyme peptide synthesis at this point; nonribosomal peptide synthesis survives to this day, albeit in a very different form (von Dohren et al., 1999). As riboorganisms came to depend heavily on amino acid metabolism, it would be useful to have a class of molecules that acted as amino acid carriers. Evolution of the aminoacylating ribozymes to act *in trans*

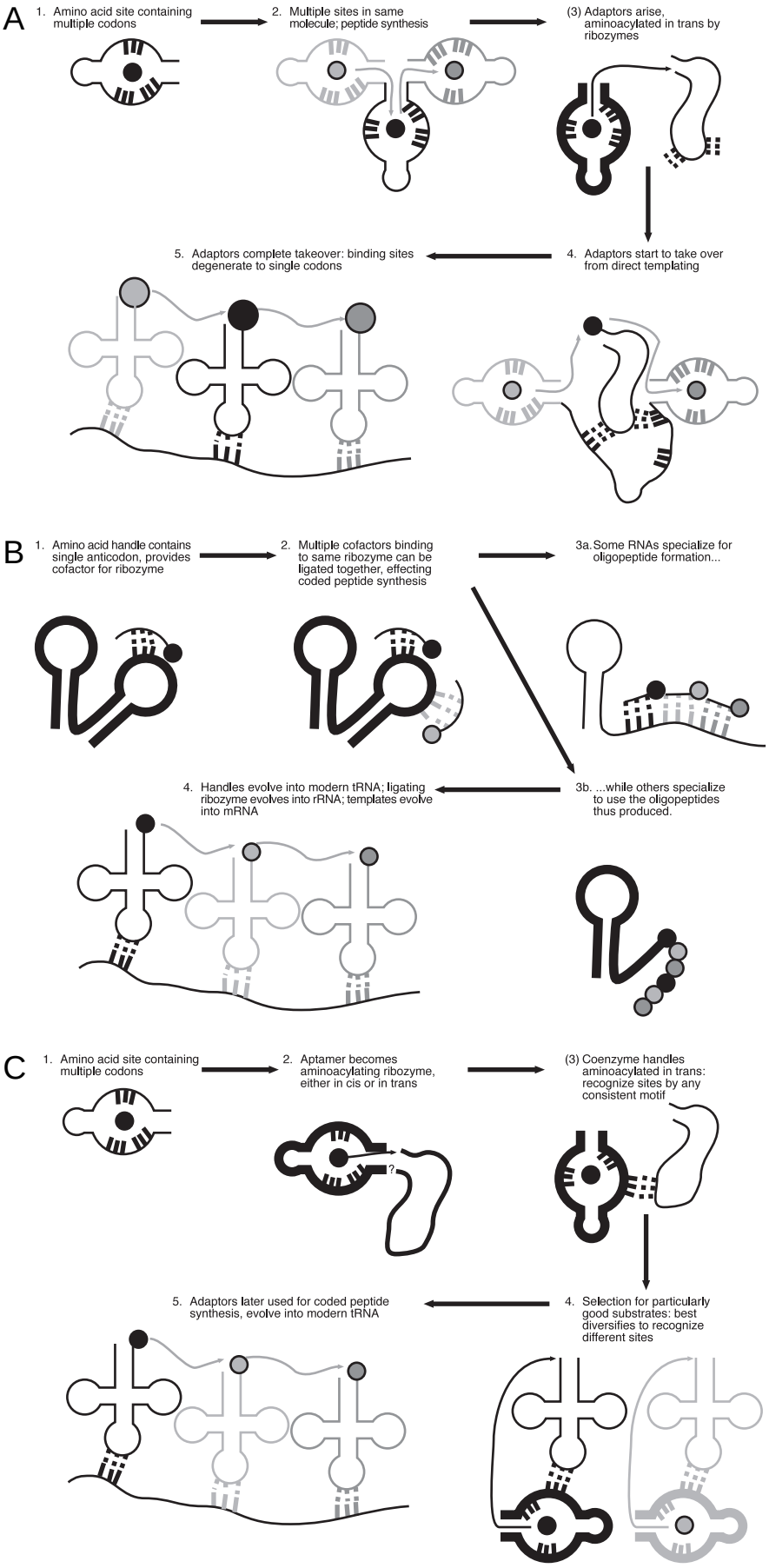


FIGURE 5. Four models of the origin of the genetic code. In DRT (**A**), multiple amino acid sites form the primordial mRNA templates; *trans*-aminoacylating ribozymes become modern rRNA; and amino acid acceptors/adaptors become modern tRNA. The original selection is for directed peptide synthesis. In CCH (**B**), oligonucleotide handles that correspond to particular amino acids become the modern tRNA; some RNA molecules specialize to use these cofactors in catalysis, becoming ribozymes that are later displaced by protein enzymes, whereas others specialize for coding, later becoming mRNA. The original selection is for noncovalent attachment of amino acids to ribozymes. In Ellington et al.'s model (2000, see Fig. 7 in their paper), peptide-ligating ribozymes evolve to use free amino acids, the aptamers evolving into modern tRNA and mRNA. The original selection is for ligation of positively charged oligopeptides to stabilize RNA structure. In our modified version of CCH (**C**), amino acid aptamers evolve into ribozyme aminoacyl-tRNA synthetases, later displaced by protein versions; the aminoacylation substrates evolve into modern tRNA, and mRNA is a later invention. The original selection is, as in CCH, for utilization of amino acids as ribozyme cofactors. DRT and modified CCH are consistent with our observation that aptamers bind arginine using its codons rather than its anticodons.

would greatly increase the turnover of aminoacylated RNA, as a single catalyst could act on many substrate molecules per unit time.

This carrier, the ur-tRNA, would be selected to be a particularly good substrate for aminoacylation. It would also need to somehow recognize the charging ribozyme. Because the charging ribozyme would overrepresent codons at the amino acid binding site (the one place other than the general catalytic site that would have to remain constant), ur-tRNAs with the corresponding anticodons would be, on average, more likely to pair with ribozymes charging particular amino acids. This would be especially important if nonhomologous sites charging the same amino acid had nothing in common except the overrepresentation of codons, as seems to be the case with actual aptamers from different selections. If particular sequences (i.e., the ur-tRNA) made especially good aminoacylation substrates, then they could duplicate and diverge (maintaining monophyly as Ellington et al. propose) not as individual catalysts, but as acceptors that recognized particular preexisting catalysts by the sequences most likely to be overrepresented in their amino acid-specific domains. In essence, if the amino acid is thought of as a key for the lock formed by the binding site, then the anticodon acts on average as a second key for recognition of the same site if the site is composed primarily of codons.

Like DRT, this model has the advantages of explaining why tRNAs act only as acceptors of amino acids, and why they contain the complements of the trinucleotides that are overrepresented at binding sites rather than those sequences themselves. Unlike DRT, CCH defers coded peptide synthesis until after the evolution of tRNA-like molecules, eliminating discontinuities in reading frame size and in insertion of an extra component into an early but already developed translation apparatus. This model still has several difficulties: it requires that the acceptor and the amino acid bind the site at the same motif, that there would be selection for aminoacylation of an intermediate amino acid carrier rather than of the final target for the amino acid, and that selection would somehow favor amino acid binding in the first place. We propose it not as a definitive pathway for the evolution of the genetic code, but rather as a starting point for further elaboration.

CONCLUSION

The association between amino acids and their cognate codons, although apparent for the amino acids for which sufficient data are presently available, need not be universal, especially because natural selection would have favored assignments of similar amino acids to related codons (Ardell, 1998; Freeland & Hurst, 1998a, 1998b). Some amino acids, such as tryptophan, glutamine, and asparagine, may have entered the code relatively late, after the present tRNA/aminoacyl-tRNA synthetase/ribosome system arose (Wong, 1975). Consequently, those amino acids will probably not be recognized by their cognate codons, and other, earlier amino acids may have taken on those codons as part of their primordial set. However, we can now tentatively predict that future amino acid selections will reveal the primordial codons, if any, assigned to those amino acids in the RNA world. One thing is certain: new selection experiments are greatly needed to generate more data and fresh hypotheses about RNA–amino acid interactions. It will be particularly interesting to see whether some amino acids common in prebiotic syntheses but not used in translation, such as norleucine, norvaline, and pipecolic acid (Wong & Bronskill, 1979), were excluded from the code because no consistent codon motif could recognize them. We hope that data from SELEX experiments will complement the conclusions from other analyses to provide a coherent account of when and why particular amino acids entered the canonical genetic code, as its mystery finally yields to experimentation.

ACKNOWLEDGMENTS

We thank Mike Yarus, Andy Ellington, and Steve Freeland for comments and discussion. This research supported in part by National Science Foundation grant MCB-9604377.

REFERENCES

Ardell DH. 1998. On error minimization in a sequential origin of the standard genetic code. *J Mol Evol 47*:1–13.

Connell GJ, Yarus M. 1994. RNAs with dual specificity and dual RNAs with Similar specificity. *Science 264*:1137–1141.

Crick FHC. 1968. The origin of the genetic code. *J Mol Biol 38*:367–379.

Ellington AD, Khrapov M, Shaw CA. 2000. The scene of a frozen accident. RNA 6:485–498.

Famulok M. 1994. Molecular recognition of amino acids by RNA-aptamers: An L-citrulline binding RNA motif and its evolution into an L-arginine binder. *J Am Chem Soc 116*:1698–1706.

Fan P, Suri AK, Fiala R, Live D, Patel DJ. 1996. Molecular recognition in the FMN-RNA aptamer complex. *J Mol Biol 258*:480–500.

Freeland SJ, Hurst LD. 1998a. The genetic code is one in a million. *J Mol Evol 47*:238–248.

Freeland SJ, Hurst LD. 1998b. Load minimization of the code: History does not explain the pattern. *Proc Roy Soc Lond B 265*:1–9.

Gamow G. 1954. Possible mathematical relation between deoxyribonucleic acid and protein. *Kgl Dansk Videnskab Selskab Biol Medd 22*:1–13.

Geiger A, Burgstaller P, von der Eltz H, Roeder A, Famulok M. 1996. RNA aptamers that bind L-arginine with sub-micromolar dissociation constants and high enantioselectivity. *Nucleic Acids Res 24*:1029–1036.

Ibba M, Soll D. 1999. Quality control mechanisms during translation. *Science 286*:1893–1897.

Illangasekare M, Sanchez G, Nickles T, Yarus M. 1995. Aminoacyl-RNA synthesis catalyzed by an RNA. *Science 267*:643–647.

Illangasekare M, Yarus M. 1999. Specific, rapid synthesis of Phe-RNA by RNA. *Proc Natl Acad Sci USA 96*:5470–5475.

Jiang F, Kumar RA, Patel DJ. 1996. Structural basis of RNA folding and recognition in an AMP{RNA aptamer complex. *Nature 382*:183–186.

Jiang L, Majumdar A, Hu W, Jaishree TJ, Xu W, Patel DJ. 1999. Saccharide-RNA recognition in a complex formed between neomycin B and an RNA aptamer. *Structure Fold Des 7*:817–827.

Jiang L, Patel DJ. 1998. Solution structure of the tobramycin-RNA aptamer complex. *Nat Struct Biol 5*:769–774.
Jiang L, Suri AK, Fiala R, Patel DJ. 1997. Saccharide-RNA recognition in an aminoglycoside antibiotic-RNA aptamer complex. *Chem Biol 4*:35–50.
Knight RD, Freeland SJ, Landweber LF. 1999. Selection, history and chemistry: The three faces of the genetic code. *Trends Biochem Sci 24*:241–247.
Knight RD, Landweber LF. 1998. Rhyme or reason: RNA-arginine interactions and the genetic code. *Chem Biol 5*:R215–R220.
Lacey JC Jr. 1992. Experimental studies on the origin of the genetic code and the process of protein synthesis: A review update. *Orig Life Evol Biosph 22*:243–275.
Lacey JC Jr, Mullins DW Jr. 1983. Experimental studies related to the origin of the genetic code and the process of protein synthesis—A review. *Orig Life 13*:3–42.
Lazcano A, Miller SL. 1996. The origin and early evolution of life: Prebiotic chemistry, the pre-RNA world, and time. *Cell 85*:793–798.
Lin CH, Wang W, Jones RA, Patel DJ. 1998. Formation of an amino-acid-binding pocket through adaptive zippering-up of a large DNA hairpin loop. *Chem Biol 5*:555–572.
Lohse PA, Szostak JW. 1996. Ribozyme-catalysed amino-acid transfer reactions. *Nature 381*:442–444.
Miller SL. 1987. Which organic compounds could have occurred on the prebiotic Earth? *Cold Spring Harb Symp Quant Biol LII*:17–27.
Osawa S. 1995. *Evolution of the genetic code.* Oxford: Oxford University Press.
Osawa S, Jukes TH. 1989. Codon reassignment (codon capture) in evolution. *J Mol Evol 28*:271–278.
Robertson SA, Harada K, Frankel AD, Wemmer DE. 2000. Structure determination and binding kinetics of a DNA aptamer-arginamide complex. *Biochemistry 39*:946–954.
Schultz DW, Yarus M. 1994. Transfer RNA mutation and the malleability of the genetic code. *J Mol Biol 235*:1377–1380.
Sokal RR, Rohlf FJ. 1995. *Biometry: The principles and practice of statistics in biological research*, 3rd ed. New York: W.H. Freeman and Company.
Szathmáry E. 1993. Coding coenzyme handles: A hypothesis for the origin of the genetic code. *Proc Natl Acad Sci USA 90*:9916–9920.
Szathmáry E. 1999. The origin of the genetic code: Amino acids as cofactors in an RNA world. *Trends Genet 15*:223–229.
Tao J, Frankel AD. 1996. Arginine-binding RNAs resembling TAR identified by in vitro selection. *Biochemistry 35*:2229–2238.
von Dohren H, Dieckmann R, Pavela-Vrancic M. 1999. The nonribosomal code. *Chem Biol 6*:R273–R279.
Weber AL, Miller SL. 1981. Reasons for the occurrence of the twenty coded protein amino acids. *J Mol Evol 17*:273–284.
Welch M, Majerfeld I, Yarus M. 1997. 23S rRNA similarity from selection for peptidyl transferase mimicry. *Biochemistry 36*:6614–6623.
Woese CR, Dugre DH, Dugre SA, Kondo M, Saxinger WC. 1966. On the fundamental nature and evolution of the genetic code. *Cold Spring Harb Symp Quant Biol 31*:723–736.
Wong JT-F. 1975. A co-evolution theory of the genetic code. *Proc Natl Acad Sci USA 72*:1909–1912.
Wong JT-F, Bronskill PM. 1979. Inadequacy of prebiotic synthesis as origin of proteinaceous amino acids. *J Mol Evol 13*:115–125.
Yang Y, Kochoyan M, Burgstaller P, Westhof E, Famulok F. 1996. Structural basis of ligand discrimination by two related RNA aptamers resolved by NMR spectroscopy. *Science 272*:1343–1346.
Yarus M. 1998. Amino acids as RNA ligands: A Direct-RNA-Template theory for the code's origin. *J Mol Evol 47*:109–117.
Yarus M. 2000. RNA–ligand chemistry: A testable source for the genetic code. *RNA 6*:475–484.
Yarus M, Christian EL. 1989. Genetic code origins. *Nature 342*:349–350.
Zimmermann GR, Shields TP, Jenison RD, Wick CL, Pardi A. 1998. A semiconserved residue inhibits complex formation by stabilizing interactions in the free state of a theophylline-binding RNA. *Biochemistry 37*:9186–9192.

2.4 The Adaptive Evolution of the Genetic Code

This chapter summarizes the evidence that the genetic code was chosen by natural selection over the vast range of possible alternatives. Here I briefly review evidence that the code has changed in recently-diverging lineages (see Section 3 for far more detail on this), and outline the evidence for and against the idea that various aspects of the code represent adaptations, including the choice of bases and amino acids, the pattern of degeneracy in the genetic code table, and the assignments of particular amino acids to particular codon blocks. I also review the evidence that the code expanded from an earlier form, leaving traces of its history in the modern code table. Interestingly, even if metabolic pathways did restrict which codons could be assigned to which amino acids, the code still appears nearly optimal in terms of error minimization. The next chapter provides some experimental tests of some of the issues raised in this chapter.

2.4.1 Abstract

The genetic code is structured so that similar amino acids are assigned similar codons. That these patterns exist is uncontroversial, but their origin and significance have remained obscure until recently. Statistical comparisons of the actual code to the distribution of possible alternatives clearly show that vanishingly few codes minimize genetic errors (from replication and translation) better than does the actual code. The most plausible explanation is that the code has been selected for fidelity over alternatives, although stereochemical constraints may have influenced at least some codon assignments (see chapter 2.1).

The code can and has evolved, as shown by variants in both mitochondrial and nuclear lineages. Where there is variation, there can be adaptation. Are the variant genetic codes superior to the code found in most organisms, adapted to different environments, or neutral changes? At least some of the processes leading to the evolution of the canonical code may have been entirely different from those leading to modern variants, especially since tRNAs and aminoacyl-tRNA synthetases now intervene between codon and amino acid, and since it is likely that amino acids were added to a limited initial set. The pathway by which the code expanded from a simpler form is still highly controversial, but we can still ask whether there is evidence that selection chose particular aspects of the coding system over plausible alternatives.

Here we review the evidence that various aspects of the genetic code, including its composition, its degeneracy, and the assignments of particular codons to particular amino acids, are in some sense optimal. We also examine several specific proposals about how the code evolved prior to its fixation in the LUCA. We conclude that although the code appears nearly optimal, there are still many interesting and unsolved questions about its early evolution.

2.4.2 Introduction

Perhaps the most fundamental divide between theories of code evolution separates those that postulate a direct reason for every codon assignment from the beginning (stereochemical and mathematical theories) from those that assume that the code can and has changed (adaptive and coevolutionary theories). This chapter focuses on the latter, all of which fundamentally assume that the code has changed for some *reason*, perhaps merely to include new amino acids that increase the catalytic diversity of encoded proteins, but perhaps in subtle ways such as to minimize the likelihood of genetic errors during replication and translation. If we accept that the code is only one of a vast number of possibilities, it becomes possible to ask why we have this code rather than its alternatives: in other words, what is it good for?

Not every feature of an organism is an adaptation. Some features are determined by the laws of physics (Thompson 1917); others arise as side-effects of other adaptive choices (Gould and Lewontin 1979). In order to demonstrate that a trait t is 'an adaptation for' a property p, it is necessary to show (a) that variation in t actually does cause variation in p, and (b) that the fitness advantage attributable to heritable variation in p led to the fixation of t in a population in which it was originally polymorphic (Neander 1991). Thus, the claim that t is an adaptation for p is strong, but often testable.

The idea that the structure of the genetic code is an adaptation (for example, because it minimizes the average difference in the chemical properties of amino acids substituted by a single-base misreading) has been challenged primarily on the second point, that the code actually was selected over alternatives. The first point, that variation in the code structure could change the frequency of genetic errors, is uncontroversial: the pattern of codon/amino acid relationships could easily be structured in way that clusters similar amino acids together to minimize the effects of substitution, or that disperses them through the code table to maximize the likelihood of detecting a substitution, so that a single-base mutation would be disproportionately likely to introduce a small or large change in amino acid property. It was rapidly observed that some features of amino acids, such as polarity, really do cluster together in the code (Woese 1965), but do such patterns require an explanation or could they have arisen by chance (or as a side-effect of some other process)?

Resistance to the idea that the genetic code is adapted to minimize errors comes from three general intuitions:

1. The actual genetic code is the only possible genetic code, so is not an adaptation because it has not been selected over alternatives (Gamow 1954).
2. Variation in the genetic code is so destructive that, even were a variant code better at minimizing new errors, the errors introduced by the change in the code would be so great as to prevent its fixation. Thus the genetic code has had no variants with which it could compete on an equal basis (Crick 1968).
3. Genetic codes that minimize errors better than the actual genetic code can be invented; therefore, there is no error minimization to be explained as an adaptation (Wong 1980).

The first two of these intuitions were demolished by the demonstration that the code actually has changed (Barrell, Bankier et al. 1979), although it was some time before this consequence was fully appreciated (Osawa 1995, Knight, Freeland et al. 1999; Knight, Freeland et al. 2001), and so the 'universal' code found in the last common ancestor of life is not the only possible code. The third, the extent to which the code structure is optimized, and whether what was probably a multidimensional optimization process can be recaptured using a single, linear measure, is at the center of current debate in the field. This chapter reviews the many, varied claims about adaptive patterns in the code, and highlights some of the main conceptual and methodological differences underlying the wildly different estimates of code optimality that can be found in the literature.

2.4.3 How Can the Code Change?

The 'frozen accident' theory of the genetic code states that, as soon as the code became good enough that cells relied on its proteins, no further change would be possible (even if it were far from optimal) (Crick 1968). This theory was partly motivated by the observation that the code was unchanged in organisms as diverse as *E. coli*, various bacteriophages, yeast and humans. However, we now know that the code is not universal and in fact can change, because it actually has changed in a variety of extant lineages (Knight, Freeland et al. 2001). All of these codes are recent variations on the 'universal' or 'canonical' code, which is still found in most organisms. Still, if the code is not universal, we need to explain (a) why it was

not one of the known variant codes that gave rise to all extant life instead of the 'canonical' code, and (b) what the possible range of genetic codes might look like.

There are a few general patterns in variant codes. First, 4-codon blocks with the same initial doublet (e.g. CGN) can be either split or unsplit. In the canonical code, most split blocks are 2/2 between the pyrimidines and the purines (e.g. GAY Asp and GAR Glu); however, the AUN block is split 3/1 between Ile and Met. In variant codes, the main form of change seems to be variation back and forth between this type of 2/2 and 3/1 split, which can be explained by variation in chemical modification at the 'wobble' base at the first position of the tRNA anticodon (Muramatsu, Nishikawa et al. 1988; Senger, Auxilien et al. 1997; Knight, Freeland et al. 2001).

The second type of reassignment is block reassignment. For example, yeast mitochondria assign the CUN block, normally Leu, to Thr (Bonitz, Berlani et al. 1980); the AGR codon block has been reassigned from Arg to Ser, Gly and Stop in the metazoa (reviewed in Osawa, Ohama et al. 1989). This type of reassignment is caused by duplication and mutation of tRNAs. Although most reassignments are compatible with known mechanisms of coding ambiguity (Schultz and Yarus 1994; Schultz and Yarus 1996), where a tRNA expands its range to read additional codons that are also read by another tRNA, the identity of some changes (such as the CUN block reassignment) cannot be reconciled with this mechanism. Consequently, it is prudent to assume that the only restriction is that *all* the codons for a particular amino acid cannot be reassigned or made ambiguous, since this would effectively remove an amino acid from the code and would probably be deleterious.

These patterns describe the types of substitution that have occurred, but not the evolutionary forces driving them. The most widely believed hypothesis, 'Codon Capture', proposes that changes in directional mutation causes certain codons to disappear from the genome; changes in the assignment of these codons (for instance, by mutation of tRNAs) can therefore occur with selective neutrality, and be fixed by selection once the direction of mutation changes and there is new pressure to translate the newly abundant codons (even if the meaning is altered) (Jukes, Osawa et al. 1987; Osawa and Jukes 1988; Osawa and Jukes 1989; Osawa, Jukes et al. 1992). Although this model does not adequately describe all codon reassignments (Knight, Landweber et al. In Press), successive rounds of AT- and GC-pressure could potentially exchange the meaning of any two arbitrary codons (Szathmáry 1991). Thus the modern translation system of tRNAs, aminoacyl-tRNA synthetases (aaRS), release factors and ribosomes can in principle support adaptive optimization of codon assignments.

The forces acting on genetic codes prior to the last common ancestor of extant life may have been quite different, however. It is likely that the genetic code evolved from an earlier form, encoding fewer amino acids: this process of code expansion does not seem to be continuing in modern organisms, although selection for strains of bacteria that can use nonstandard amino acids has in some cases led to stable incorporation (Wong 1983; Budisa, Minks et al. 1998; Budisa, Minks et al. 1999). The proteins involved in the translation system, such as the aminoacyl-tRNA synthetases, cannot have preceded protein synthesis itself, although the recent artificial selection of ribozymes that catalyze these reactions (Illangasekare, Sanchez et al. 1995; Illangasekare and Yarus 1999; Illangasekare and Yarus 1999; Lee, Bessho et al. 2000) support the idea that the protein synthetases may have usurped the role from earlier catalysts made of RNA (Nagel and Doolittle 1995; Wetzel 1995).

There is some evidence that the code is still expanding. Many amino acids, such as hydroxyproline and phosphoserine, are not incorporated during translation but are produced later by enzymes acting on the nascent polypeptides (sometimes reversibly). Selenocysteine, which is incorporated during translation (Zinoni, Birkmann et al. 1987) in certain species by a special tRNA (Leinfelder, Zehelein et al. 1988) at UGA stop codons in certain sequence contexts (a hairpin recognition element upstream, and a 4^{th} base context that is recognized only weakly by release factors)(Zinoni, Heider et al. 1990; Tate, Poole et al. 1996; Tate,

Mansell et al. 1999), may be an example of a recently added amino acid. The selenocysteine tRNA is originally recognized by seryl-tRNA synthetase and charged with Ser, after which a specific enzyme, selenocysteine synthase, recognizes the charged tRNA and converts the amino acid to selenocysteine (Forchhammer, Boesmiller et al. 1991; Commans and Bock 1999; Lenhard, Orellana et al. 1999). This phenomenon has parallels with Asn and Gln, which, in some species, are mischarged by aspartyl- and glutamyl-tRNA synthetase respectively, and converted to the amide on the tRNA (Schön, Kannangara et al. 1988) (see Wong 1975; Di Giulio 1993 for reviews relating this fact to code evolution; for more recent references on the function and phylogenetic distribution of these enzymes see Becker and Kern 1998; Tumbula, Becker et al. 2000). Thus it is possible that tRNA-dependent modification is an early stage in cotranslational incorporation of amino acids, and that codon assignments are influenced by the order in which amino acids were added (Dillon 1973; Wong 1975; Wong 1981; Di Giulio 1989; Miseta 1989; Taylor and Coates 1989; Di Giulio 1991; Di Giulio 1997; Di Giulio 1998).

Thus the genetic code *can* change, and in fact *has* changed. We can therefore ask:

1. Has the form and/or content of the code been shaped by natural selection?
2. What is the appropriate class of codes against which to compare the canonical genetic code, and in what respects (if any) does it differ from a randomly selected code from this set?
3. If the code is optimized for some property, has this process of optimization erased historical information that might otherwise have been present in the code's structure?

2.4.4 Early Models

Order in the genetic code was noted as soon as the codon assignments were first discovered. In particular, similar amino acids are assigned to codons connected by single-base substitutions, and the second-position base is associated with hydrophobicity (Speyer, Lengyel et al. 1963; Pelc 1965; Sonneborn 1965; Volkenstein 1965; Woese 1965; Zuckerkandl and Pauling 1965; Epstein 1966; Fitch 1966; Goldberg and Wittes 1966). Although stereochemical models (Dunnill 1966; Pelc and Welton 1966; Woese, Dugre et al. 1966; Woese, Dugre et al. 1966) were also initially popular, most authors immediately assumed that this order required an adaptive explanation: that the code had been optimized to reduce errors in replication and/or translation. These two sources of error have subtly different consequences: in both, transitions between the two purines or the two pyrimidines are more frequent than transversions from purine to pyrimidine or vice versa, but translation alone introduces the concept of reading frame-dependent error.

The Lethal Mutation model (Sonneborn 1965; Zuckerkandl and Pauling 1965) suggested that the genetic code had adapted to minimize the effects of point mutations, while the Translation Error model (Woese 1965; Woese 1967) suggested that the primary source of error was during translation. Woese noted that, although the 3rd and 1st position bases were optimized (in the sense that point substitutions tended to substitute amino acids with similar hydrophobicity), but 2nd position changes tended to be nonconservative, indicating a reading frame-dependent effect; this was consistent with the relative frequency of streptomycin-induced misreadings in model polypeptides (Davies, Gilbert et al. 1964). However, some caution is warranted: the pattern of misreading may be due to the peculiarities of codon/anticodon pairings in modern tRNA and mRNA, and may not accurately reflect the situation when coding was established (Szathmáry 1991). Whether this was also true at the time during which modern coding was established remains an open question.

In a remarkably prescient paper, Alff-Steinberger explicitly tested the average error made by point substitutions at different positions using Monte Carlo simulations, comparing the actual code to 200 alternative codes made by shuffling the amino acid properties among the 20 blocks of synonyms found in the canonical code, and comparing the average size of the errors induced by single-base misreadings at each position. He found that almost no random codes

minimized changes in polarity better than did the canonical code: the 3rd-position base was most highly optimized relative to random codes, followed by the1st position base, and there was no evidence for optimization in the 2nd-position base, consistent with the relative effects of translation error (Alff-Steinberger 1969). This clear evidence that the standard code minimized errors better than random codes was ignored for over 20 years; however, we have been unable to reproduce the quantitative results of this study.

2.4.5 Is the Choice of Components Optimal?

The modern code links L-alpha-amino carboxylic acids to codons made of nucleic acids based on D-ribose and purines and pyrimidines with a particular hydrogen-bonding pattern, but we need not take these particulars for granted. Could the set of amino acids and bases, and even the peptide and nucleic acid backbones, have been selected from a range of possibilities as the most stable, least energetically costly, or most catalytically active solutions? Although investigations along these lines are necessarily speculative (especially since no organisms with alternative coding systems survive to the present), several interesting possibilities have been suggested. Since there are hundreds of different amino acids produced in cells (Voet and Voet 1995), perhaps a couple of dozen plausible nucleic acid backbones, and perhaps half a dozen plausible alternatives to peptide backbones, as well as the possibility of codes with more than 20 amino acids and more than 4 bases, the contrast class of possible codes that this section considers is extremely large ($> 10^{100}$).

The idea that RNA preceded DNA is well-established in the literature as the RNA World hypothesis (Gilbert 1986), and the difficulty of ribonucleotide reduction suggests that DNA arose only after proteins were already being used as sophisticated catalysts (Freeland, Knight et al. 1999). However, D-ribose is not the only possible backbone that can support complementary base pairing. Analogs of RNA using sugars with fewer (Joyce, Schwartz et al. 1987; Weber 1987; Weber 1989) and more (see Eschenmoser 1999 for detailed review) carbons can be synthesized; ribose backbones do not even allow more stable base pairs than do other pentoses, which might suggest that the pairing strength has been 'optimized' rather than maximized (Eschenmoser 1999), or, alternatively, might suggest that pairing strength was not of primary importance. No sugar is stable under mainstream predictions of prebiotic conditions (Larralde, Robertson et al. 1995), and it is possible that the first backbones were based on alternative chemistries. PNA, peptide nucleic acid (Nielsen, Egholm et al. 1991), has the nucleotide bases bonded to a peptide backbone; it forms stable base pairs (even in heteroduplex with DNA or RNA)(Hanvey, Peffer et al. 1992), and can be plausibly synthesized prebiotically (Nelson, Levy et al. 2000). It is possible that an early system based on PNA was displaced by RNA (Nielsen 1993), perhaps because the charged backbone reduces aggregation and because the 2'-OH group allows a wider range of catalytic activity (Joyce, Schwartz et al. 1987; Joyce and Orgel 1993). However, such suggestions remain speculative.

Similarly, the four standard bases are not the only possibilities. A variety of bases with alternative hydrogen bond donor and acceptor patterns that support new complementary base pairs have been synthesized (Piccirilli, Krauch et al. 1990; Wu, Ogawa et al. 2000); some of these can even be incorporated by standard polymerases (Piccirilli, Krauch et al. 1990; Lutz, Horlacher et al. 1998; Ogawa, Wu et al. 2000). However, many possible base pairs are *not* stable because of increased tautomerism (Piccirilli, Krauch et al. 1990). Although adenine is easily produced by HCN polymerization (Oró and Kimball 1961), many nonstandard purines are produced under similar conditions (Levy and Miller 1999), perhaps suggesting that adenine was preferable for some reason. None of the standard bases are stable under prebiotic conditions, however, suggesting that either life arose rapidly or alternative bases were originally used (Levy and Miller 1998).

Is the number of bases and/or amino acids adaptive? Szathmary has calculated that 4 bases are better than 2 or 6 based on estimates of the likelihood of arbitrary catalysis (measuring functional diversity) and the actual pairing energies of the standard bases and Piccirilli et al.'s

nonstandard bases (Szathmary 1991; Szathmary 1992). Thus the genetic alphabet may be a tradeoff between catalytic ability and replicative fidelity.

Similarly, the protein side-chains need not be carried by a peptide backbone. One possible alternative is thioesters resulting from the condensation of thiols with carboxylic acids, which may have been common in thermal vents (de Duve 1995). Other possibilities include beta-amino acids, hydroxy acids, amides produced by diamino and dicarboxylic acid monomers, and esters. The relative stability of the amide backbone, along with the conformational rigidity enforced by the alpha-amino linkage, may account for the usage of alpha-amino acids in proteins (Weber and Miller 1981).

The choice of amino acids in the code may also be adaptive. Many functional groups, such as halides, carbonyls, phosphates, and sulfonates, are not represented in the standard amino acids, although amino acids containing these groups can be synthesized (and some, such as citrulline, are even common in cells). There is only partial overlap between the amino acids used in the code and those available by prebiotic synthesis, suggesting that some were invented later as metabolism grew more complex (Wong and Bronskill 1979), and perhaps that some primordial amino acids were eliminated from the code because they were, overall, less useful in proteins. The most extensive investigation into the set of coding and noncoding amino acids is that of Weber and Miller (Weber and Miller 1981), who rule out several prebiotically plausible amino acids on structural grounds (for instance, ornithine, which is an analog of lysine but one methylene group shorter, is unstable to cyclization by lactam formation). However, they are unable to account for the absence of several amino acids common in prebiotic synthesis, such as norleucine, norvaline, pipecolic acid, and alpha-aminobutyric acid; similarly, complex amino acids such as Trp, Arg, and His are ‘justified’ by assuming that their functional groups are necessary for catalytic activity, and that these are the simplest amino acids that contain those functional groups.

An alternative perspective (Budisa, Minks et al. 1998; Budisa, Minks et al. 1999) suggests that amino acids were chosen from metabolic pathways as those that were more useful for protein synthesis than as intermediates. In particular, this model suggests that some amino acids were excluded from the code because they were incompatible with accurate translation: for instance, in modern cells, norleucine is toxic because it is misincorporated for Met. However, this seems backwards: it is more likely that some amino acids are misincorporated in modern cells *because* the cells have not been exposed to high levels of them, especially since aminoacyl-tRNA synthetases can discriminate between amino acids as similar as leucine and isoleucine with near-perfect accuracy. The argument that amino acids that are important as metabolic intermediates cannot be incorporated into the code because such diversion would be costly is also unconvincing: while it is true that amino acids such as homoserine and pipecolic acid are used only as intermediates, many amino acids that are found in the code, such as Asp, Arg, Glu, Asn, and Gln, are central in extant metabolism.

Consequently, there is clear reason to believe that some amino acids (such as ornithine), bases (such as bromouracil, which has such a high rate of tautomerism that it is a powerful mutagens), and backbones (such as PNA, which is uncharged and therefore tends to aggregate, and which has no backbone hydroxyls to participate in catalysis) were not used for genetic coding. However, the evidence too sketchy to conclude that the choices made in the actual coding system represent an evolutionary optimum.

2.4.6 Is the Pattern of Codons Optimal?

The genetic code is degenerate, in the sense that most amino acids are assigned more than one codon. There are two senses in which this pattern of degeneracy could be adaptive. The first sense ignores the fact that different amino acids have different properties, and merely treats them as different symbols. We can ask whether the symbols are arranged to minimize (or maximize!) the chance of substituting a different symbol when misreading a single base. The second sense assumes that there is a metabolic or functional reason that proteins have a

particular distribution of amino acids, and asks whether the code is optimized to reflect this optimum frequency by varying the number of codons assigned to each amino acid. The appropriate contrast class for this section is all codes that assign at least one codon to each of the 20 amino acids and Stop (not preserving the degeneracy of the actual code). To a first approximation, assuming each codon is assigned to an amino acid independently, this is (21^{64} – 20^{64}) or about 10^{84} possible codes. It is obvious that the actual code does not look like a random code picked from this class, since different codons for the same amino acid nearly always start with the same first- and second-position base, which would not be expected by chance (Dillon 1973). Still, it is unclear whether this pattern, which has the effect of reducing missense substitutions, was selected for the purpose rather than being an artifact of the way the translation apparatus works.

Chemical explanations for the pattern of degeneracy, which take into account rules of base pairing in RNA, have been remarkably successful. Crick's Wobble Hypothesis (Crick 1966), which suggests that 3rd-position blocks are split between purine-ending and pyrimidine-ending codons (and not, for instance, between A+C and G+U-ending codons) because G recognizes both C and U at the 3rd position (and, similarly, U pairs with both A and G), is borne out by detailed studies of pairing in modern tRNAs. However, the conformation of the 'wobble base' in the tRNA, which allows this ambiguous misreading, may be a derived state and hence itself an adaptation (Szathmáry 1991). Similarly, it is possible to predict which codon doublets will be split: those where the first two positions are G or C always form a family box, while those where the first two positions are A or U are always split between two or more amino acids (Lagerkvist 1978; Lagerkvist 1980; Lagerkvist 1981). No variant codes deviate from this rule (Knight, Freeland et al. 2001), which may reflect the different bond strengths of AU and CG base pairs. Thus the overall pattern of degeneracy may reflect chemical constraints, although the small number of known variants makes such conclusions highly uncertain.

Mathematical explanations for degeneracy, which test whether the genetic code matches one of several formally optimal codes, have been far less successful. For example, the optimal way of encoding a probability distribution of states (in as terms of minimizing message length) is in a Huffman code, in which the most likely outcomes are assigned shorter symbols. Although the use of tRNA adaptors might in principle allow this, all actual codons are the same length (except for stop codons, which have a strong 4th-base context effect, but which are read by protein release factors rather than by tRNA (Tate, Poole et al. 1996)). Similarly, the genetic code might be a Baudot code, in which adjacent symbols are generated by sliding a reading frame across a cycle of bits (Cullman and Labouygues 1983), or a Gray code, an example of a minimum change code in which binary encodings of objects are arranged in a ring such that changes in progressively less significant digits of the encoding lead to substitution of increasingly similar objects (Swanson 1984). However, the evidence that the genetic code is an optimal code in either of these formal senses is unconvincing: it is highly unclear that amino acid properties should be measured as a ring rather than on a linear scale, and the claim that 'subcodes' for particular amino acids are optimal (Cullman and Labouygues 1983; Figureau and Pouzet 1984; Cullman and Labouygues 1987; Figureau 1987; Figureau 1989) reduces to nothing more than the claim that codons for a single amino acid are, in general, adjacent: this can be adequately explained by the simple fact that they are usually read by the same tRNA (or by overlapping sets of tRNAs), which cannot discriminate among them.

It is possible that a code that is a formal optimum for error minimization would not be ideal for evolution, which relies on mutations as the raw material for adaptation. An analysis that looked at the likelihood that single-base mutations would substitute amino acids differing in polarity, combined with the average shortest path length for the interconversion of pairs of amino acids, suggested that the code represents "tradeoff between robustness and flexibility" (Maeshiro and Kimura 1998), although, as always, it is difficult to tell a tradeoff between two dubious adaptations from what would be expected by chance (Gould and Lewontin 1979). Similarly, codon assignments may be adapted as a tradeoff between misreading frequency and speed of translation (Klump 1993), but in the absence of comparable data from random codes (and a

compelling reason why the tradeoff would produce the observed values) it is impossible to assess the claim.

It has been suggested that the combination of triplet codons and 20-21 symbols is a 'hardware optimum' minimizing the number of components required for translation (Soto and Toha 1985), but this appears to be numerology based on the fact that e^3 is close to 20. In any case, the genetic code does *not* maximize the entropy of codon assignments (required for optimization in this sense), since different amino acids are assigned grossly inequitable numbers of codons. Additionally, the number of tRNAs actually used varies from species to species, but should be a universal, low number were the code really optimized to minimize the number of components required for translation: while this effect is extreme in mitochondria, there is no evidence that it has influenced code evolution even there (Knight, Landweber et al. In Press). The frequency with which particular codons and amino acids are used varies widely between species, although much of the variance can be accounted for by changes in base composition. In particular, over 80% the variance in frequency in even as large and chemically active amino acid as arginine is explained by genome GC content (Knight, Freeland et al. 2001). Thus, although the number of codons assigned to each amino acid does correlate with overall abundance in proteins (Mackay 1967), it is most likely that this is because neutral mutation leads proteins to reflect the code rather than the reverse (King and Jukes 1969; Ota and Kimura 1971). Similarly, although smaller and less complex amino acids tend to be assigned more codons (Dufton 1983; Dufton 1997), the range of amino acid usage in different species makes it unlikely that this codon assignment is an adaptation to minimize the metabolic cost of making proteins.

There have been several group-theoretic explanations for the code's structure based on symmetry breaking (Antillon and Ortega-Blake 1985; Hornos and Hornos 1993; Bashford, Tsohantjis et al. 1998), but these share the flawed assumption that variant codes diverged while the code was still partially ambiguous (rather than from an already complete canonical code). There is also no reason why symmetry at this level would be biologically relevant; worse, it is possible to invent an algebra that recaptures *any* dichotomous classification, and so these techniques can only describe, rather than explain, the code structure.

Thus there is no evidence that the number of codons assigned to each amino acid is adaptive, or that the pattern of degeneracy makes the code formally optimal in terms of its error-resistance properties: in fact, the code has enough redundancy to be used as an error-detecting code (Swanson 1984), but this capacity is not used. All variant codes change the pattern of redundancy, so if a particular pattern *were* adaptive it would be possible for it to be fixed by natural selection. However, the most convincing explanations for the code's redundancy to date have been chemical: Crick's wobble hypothesis explains why 3rd positions are split between purines and pyrimidines, while Lagerkvist's observations on degeneracy and doublet GC content explain which codon blocks are split. These chemical explanations do not exclude an adaptive explanation, since they may be the modern mechanisms by which the adaptation is carried out. However, the initially plausible adaptive explanations thus far proposed become far less compelling on closer scrutiny.

2.4.7 Evidence for Adaptive Codon Assignments

It is unclear that the pattern of degeneracy itself is adaptive, but the arrangement of amino acids among codon blocks appears rather non-random. In particular, the second-position base correlates with hydrophobicity (Woese 1965), and amino acids with related functional groups such as acidic or basic or aromatic side-chains tend to be connected by point substitutions (Pelc 1965). Thus we can ask whether the actual code minimizes the effect of single-base errors better than do random codes that permute the amino acid properties across the 20 codon blocks, for a total of 20! or 2.4×10^{18} possible codes.

Every study that has compared the actual code to randomly generated codes from this set has found that very few codes are better than the actual code at minimizing the difference in

polarity, but not differences in other attributes of amino acids, such as size and composition. This is measured by calculating the difference in the property between the old and new amino acid for each possible substitution, and calculating the overall effect of point substitutions at the first, second, and third codon positions, and for the code overall. For example, using Polar Requirement as a measure, a change from UUU Phe to CUU Leu involves a change in Polar Requirement from 5.0 to 4.9, contributing a difference of 0.1 to the sum of differences in 1st-position substitutions; conversely, a change from UUU Phe to GUU Val involves a change from 5.0 to 5.6, contributing a difference of 0.6 to the sum of differences in 1st-position substitutions (Fig. 1). The calculations can be repeated for randomly generated codes to obtain a distribution against which the actual code can be compared.

Results from this type of analysis provide compelling evidence that the code is far better at minimizing errors in polarity than would be expected by chance. Alff-Steinberger reported that the code minimized errors in molecular weight, polar requirement, number of dissociating groups, pK of carboxyl group, isoelectric point, and alpha-helix-forming ability, with the 3rd position base most highly optimized and the 2nd position base not optimized at all (Alff-Steinberger 1969), although we are unable to replicate any of the numerical results in this paper and find that, of these measures, only differences in polar requirement are actually minimized by the code's structure (unpublished work). Haig and Hurst found that the code was highly optimized for polar requirement (1 in 10 000), somewhat less optimized for hydropathy (1 in 1000), and not optimized for molecular volume and isoelectric point (Haig and Hurst 1991; Haig and Hurst 1999). This is interesting, because polar requirement (Woese, Dugre et al. 1966) is a measure of the hydrophobicity of free amino acids, while hydropathy (Kyte and Doolittle 1982) is a measure of the hydrophobicity of the side-chains alone (but is unsuitable for detailed estimates of code optimality because many values were arbitrarily adjusted relative to the experimental data: see their pp. 109-110!). Freeland and Hurst found that, with polar requirement, the code appears increasingly optimal when the ratio of transitions to transversions increases, and found only one better variant in a sample of 1 million codes when weighting for both transition bias and the relative frequency of experimentally observed mistranslation at the three positions (Freeland and Hurst 1998). Similar results are obtained when the modular power (the power to which the difference between pairs of amino acids is raised, e.g. squared deviations represent a modular power of 2) is varied, or the PAM74-100 matrix (Benner, Cohen et al. 1994), a measure of the actual frequency of amino acid substitutions in distantly related proteins, is used (Ardell 1998; Freeland, Knight et al. 2000). In fact, about 100-fold fewer better codes are found with PAM74-100 than with polar requirement (Freeland, Knight et al. 2000), perhaps suggesting that the code is better adapted to the functional properties of amino acids within proteins than to the properties of the free amino acids. This point is critically important for determining when the code was most recently optimized.

A second approach is to test which amino acid properties correlate with specific types of substitution matrix, which can either be generated directly from the genetic code (which is arranged such that some amino acids are easier to interconvert than others) or derived from observed substitutions in proteins (which, at short mutational distances, largely reflect the code's structure). Studies based on these methods, although less clear-cut than those that test directly what fraction of codes minimize changes in particular properties under point substitution better than does the actual code, consistently show that the main property implicated in substitutability is hydrophobicity or measures tightly correlated with it (Wolfenden, Andersson et al. 1981; Sitaramam 1989; Szathmáry and Zintzaras 1992; Joshi, Korde et al. 1993; Benner, Cohen et al. 1994; Tomii and Kanehisa 1996; Koshi and Goldstein 1997; Xia and Li 1998). Depending on the study, the most tightly correlated hydrophobicity scale may be one that is measured directly as a chemical property of the amino acids or side-chains (Woese, Dugre et al. 1966; Wolfenden, Andersson et al. 1981; Kyte and Doolittle 1982; Fauchere and Pliska 1983; Radzicka, Young et al. 1993; Rose and Wolfenden 1993), or indirectly as relative abundance in the interiors/exteriors of proteins (Robson and Suzuki 1976; Levitt 1978; Wertz and Scheraga 1978; Nakashima, Nishikawa et al. 1990). However, other

potentially important factors (such as the size of the side-chains) do not turn out to be nearly as tightly correlated with the code structure. The converse of this approach has also been tried, with similar results: in this case, random amino acid indices were generated, the ones that best matched the genetic code substitution matrix were selected, and these scales were correlated with measures of polarity (Knight, Freeland et al. 1999)Similar results have been obtained by multivariate (Sjöström and Wold 1985) and neural network (Jiménez-Montaño 1994; Tolstrup, Toftgard et al. 1994) analyses that partition codons into classes based on multidimensional analysis of amino acid properties, and find that the second-position base defines classes linked to various hydrophobicity measures (or principal components thereof). Thus it should not be controversial that the code minimizes changes in hydrophobicity better than does almost every random permutation of amino acid properties among codon blocks.

Despite this overwhelming evidence, a small but vocal group of researchers still doubt that the code has been optimized to minimize the effects of genetic errors. This arises from two fundamental misunderstandings. The first misunderstanding is to assume that, if the code has been optimized by natural selection, that it must be the best of all possible codes at minimizing the distance function. Consequently, better codes found by powerful computer search algorithms (or calculated from first principles), but which do not resemble the standard code, have been presented as prima facie evidence that the code is not optimal, and therefore cannot have been optimized (Wong 1980; Di Giulio 1989; Goldman 1993; Di Giulio, Capobianco et al. 1994; Judson and Haydon 1999; Di Giulio 2000; Di Giulio 2000).

These analyses fail to take into account two important facts. First, the average effect of amino acid changes in proteins is unlikely to be perfectly recaptured by a single linear scale of physical properties, and so a code that minimizes a single one of these properties will not necessarily look anything like the actual code (Freeland, Knight et al. 2000). Second, although search algorithms can sample billions of different codes, evolution is unlikely to have had similar opportunity given the extreme cost of changing an already functional code, and so we might either expect the code to be trapped at a local, rather than global, optimum, or that the code is a case of asymptotic adaptation and has not yet reached perfection. Thus, the fact that the code is not the best of all possible codes on a particular hydrophobicity scale does not mean that it has not evolved to minimize changes in hydrophobicity under point misreading, any more than the fact that the vertebrate retina is wired backwards means that the eye is not adapted for vision.

The second misunderstanding is to assume that the important property for measuring the extent of code optimization is not the fraction of codes to which the actual code is superior, but rather the distance still separating the actual code from the Panglossian ideal (Wong 1980; Di Giulio 1989; Di Giulio, Capobianco et al. 1994; Di Giulio 1998; Di Giulio 2000; Di Giulio 2000).Figure 2 shows the difference between these two models, using an actual distribution of random codes derived from the PAM74-100 matrix (Freeland, Knight et al. 2000). The distribution is roughly Gaussian: better codes are rarer (and confer lesser relative advantage) as the code becomes more highly optimized. No-one has studied the accessibility of better codes as optimization proceeds, but it is likely that better codes get exponentially rarer at the tails of the distributions and so it would be necessary to cross huge fitness valleys to find a code that is better than a near-optimal one. However, the fact that such a small fraction of codes are better at minimizing errors than the actual code strongly suggests that selection has played a role in determining its structure (Freeland, Knight et al. 2000).

Another interesting observation is that the amino acids are arranged so as to give a smooth fitness landscape: in other words, the first- and second-position bases have a roughly consistent, additive effect on several amino acid properties, including polarity (Aita, Urata et al. 2000). This has the effect of allowing ‘fine-tuning’ of amino acid properties by single-base mutational events. Interestingly, we find that relative codon and amino acid usage in different species can be largely explained by a mutation-selection balance on individual bases, and that the rate of change under directional mutation of each base at each position is highly correlated

with the average effect of changing that base (Knight, Freeland et al. 2001). It is possible that this result holds because of the structure of the code, rather than in spite of it.

The remainder of this section highlights some methodological details and issues, some of which will allow easier comparison of the various estimates of code optimality in the literature, and some of which suggest future directions of research.

Stop codons: Termination codons have been ignored since the beginning (Alff-Steinberger 1969): the reason for this is that the error produced by substituting one amino acid for another is well-defined, but errors induced by termination are not. All studies reviewed here ignore changes to/from stop codons, by not counting these as valid changes (with the result that codons adjacent to stop codons effectively have fewer mutational neighbors). Although intuition suggests that changes to stop codons are always deleterious, it is not clear that this should be as true for translation error as it is for mutations. The C-termini of proteins often contain signals that prevent degradation, and it may be better to terminate and destroy an incorrect protein early in synthesis than to have either an inactive copy (or, worse, a copy that catalyzes the wrong reaction, such as hydrolyzing ATP without making a useful product) littering the cell. Since the code seems to be adapted against translation error rather than mutation error, more research is needed to determine whether a translational change to Stop is, on average, more or less costly than an average missense change.

Measures of Amino Acid Properties: Polar Requirement is a measure of the hydrophobicity of free amino acids, as measured chromatographically in mixtures of a water/pyridine solvent system (Woese, Dugre et al. 1966). This system was chosen because pyridine somewhat resembles a pyrimidine base; note that the published values are actually based on 2,4-dimethylpyridine, which blocks interactions with the ring nitrogen. All studies that have used this measure have found the code to be close to optimal (best of 10 000 or better) (Haig and Hurst 1991; Szathmáry and Zintzaras 1992; Ardell 1998; Freeland and Hurst 1998; Freeland and Hurst 1998; Haig and Hurst 1999; Freeland, Knight et al. 2000). Hydropathy (Kyte and Doolittle 1982) is a measure of the distribution of analogs of the side-chains between water or organic phases and the vapor phase, but the values were ‘manually adjusted’ in several cases and so this should not be used as a quantitative index for studies of code optimality. The code looks about 10-fold worse using this measure (Haig and Hurst 1991), although it is 78% correlated with polar requirement. The PAM74-100 matrix is a substitution matrix generated by counting actual pairwise substitutions in highly-diverged protein families (Benner, Cohen et al. 1994), and is markedly different from substitution matrices from more closely related amino acids; it is 85% correlated with the matrix similarly derived from low-distance proteins, but only 54% correlated with the matrix that would be expected from random divergence based on the genetic code structure. Although PAM74-100 is highly correlated with some polarity measures, it is only 50% correlated with a distance matrix constructed from polar requirement, which implies that it reflects polarity in a rather different way. Since PAM74-100 measures actual substitutions in proteins, it is reassuring that it makes the code appear even more optimal than does polar requirement (Ardell 1998; Freeland, Knight et al. 2000). However, the significant correlation between PAM74-100 and the genetic code matrix is cause for concern (Di Giulio 2001), and randomized matrices as different from the genetic code matrix as is PAM74-100 should be tested to ensure that these results do not merely reflect the code structure. Other measures of amino acid properties are available: whether the code is most highly optimal with respect to properties of the free amino acids, the side-chains, or statistical properties of amino acids in modern properties may be an important clue to deciding when the code adapted to the cellular environment.

Mutation: The pattern of mutation is markedly biased in different organisms (Sueoka 1961; Muto and Osawa 1987; Sueoka 1988). Perhaps the best-known bias is between transitions and transversions (Epstein 1967), which may have a direct chemical rationale (Topal and Fresco 1976) which may also apply to mistranslation (Topal and Fresco 1976). When transitions are weighted more heavily than transversions (to reflect the observed fact that they occur more frequently in organisms), the code appears increasingly optimal as the transition

bias increases (Ardell 1998; Freeland and Hurst 1998; Freeland, Knight et al. 2000). However, the effects of other biases in the frequency of particular types of mutation have not yet been systematically investigated. These may be especially important in mitochondria, where only transition pairs are correlated, and all other pairs of nucleotides are uncorrelated in frequency (Knight, Landweber et al. In Press). In particular, the known variant codes all have error values very similar to (but slightly worse than) the standard code, and thus could be slightly deleterious mutants (Freeland, Knight et al. 2000). However, it is also possible that these codes are specific adaptations to the altered distribution of mutations and/or translation errors in mitochondria. It would also be interesting to test whether the code is better optimized to mutational biases prevalent in single- or double-stranded RNA than to the DNA used at present.

Thus the code appears nearly optimal with respect to changes in hydrophobicity, but not other parameters, over wide range of parameter space. Although it is attractive to assume that this is the result of natural selection for error minimization, alternative explanations cannot yet be ruled out: for example, stereochemical principles could assign similar amino acids to similar codons (see chapter 2.1). However, the view that the code is not highly ordered is no longer tenable.

2.4.8 Metabolic Restrictions on Codon Assignments?

The most serious challenge to the idea that the code has been optimized for error minimization is the idea that it contains a strong historical signature: surely, the argument goes, if codon assignments had been shuffled to reduce error, these nonadaptive patterns would have been erased (Wong 1976; Wong 1980; Di Giulio 1989; Di Giulio 1989; Di Giulio 1991; Di Giulio 1998; Di Giulio 1999; Di Giulio 2000; Di Giulio and Medugno 2000)? Thus ‘coevolution’ between codon assignments and the amino acid complement has been viewed as a major alternative to adaptive models: coevolution accepts that the code did change, and that codon assignments were rearranged, but only to insert new amino acids as metabolism invented them and *not* to optimize the code structure.

However, there is no necessary contradiction between history and adaptation. Adaptive traits typically reveal traces of their past: the bat’s wing is no less adapted for flying, nor the human hand for grasping, because they are immediately recognizable as pentadactyl limbs. More fundamentally, however, historical patterns in the code can only cause difficulties for the adaptive hypothesis if they actually exist.

Although it seems likely that the genetic code evolved from an earlier form with fewer codons and amino acids (Crick 1968), there is no agreement in the literature about how this occurred. For instance, the first codons have been proposed to be RRY (Crick 1968), RNY (Eigen and Schuster 1979; Shepherd 1981), GNN (Dillon 1973; Taylor and Coates 1989; Davis 1999), all-(A,U) (Jiménez-Sánchez 1995), all-(G,C) (Lehman and Jukes 1988; Hartman 1995), all-purine (Baumann and Oró 1993), NYN vs. NRN (Fitch 1966; Fitch and Upper 1987), GCU alone (Trifonov and Bettecken 1997), etc. The fundamental difficulty is that the genetic code consists of a small, highly connected set of elements, and it is easy to see patterns that are only weakly supported by evidence. Recently, Trifonov has compiled from the literature 40 conjectures about the pathway of code evolution, averaged the rank orders together, and called the result a ‘consensus temporal order’ of the amino acids (Trifonov 2000). Trifonov is encouraged by the generally low pairwise correlations between estimates (only about 20 of the 720 pairwise correlations are above 0.5), suggesting that this lack of relatedness implies that the measures are truly independent. However, the consensus rank order is 88% correlated with the rank abundance of amino acids in proteins (which itself is almost completely determined by the number of codons (King and Jukes 1969)), and 80% correlated with molecular weight (the two measures are themselves 73% correlated). This suggests that different authors intuitively agree that larger amino acids, which tend to have fewer codons, were probably late entries into the code, but disagree on every other point!

The idea that metabolically related amino acids are clustered together in the code is not new (Pelc 1965). However, perhaps the best-known hypothesis to explain this is that of Wong (Wong 1975), which had been cited over 120 times as of early 2001. According to this model, new amino acids were formed by tRNA-dependent modification, much as Gln and Asn are formed from Glu and Asp in some bacteria, and thus took over a subset of codons from their metabolic precursors. Specifically, Wong identified 8 precursor-product pairs: Ser → Trp Ser → Cys, Val → Leu, Thr → Ile, Gln → His, Phe → Tyr, Glu → Gln, and Asp → Asn. Using the hypergeometric distribution, he estimated the probability that, for each of these pairs of amino acids, the number of codons of the product that were connected to at least one codon of the precursor would be as great as those actually observed. He then combined these probabilities using Fisher's method, to get an overall probability of 0.0002 of observing at least as much overlap between products and precursors by chance (Wong 1975).

Unfortunately, there are good reasons to disbelieve this result. The first is that even randomly generated codes contain many 'product-precursor pairs' by chance, especially if known pathways from *E. coli* are used instead of Wong's original pairs (Amirnovin 1997; Amirnovin and Miller 1999). Although the 'codon correlation score' used in this study to assess the position of the actual code relative to random codes has been criticized (Di Giulio 1999; Di Giulio and Medugno 2000), the fact remains that it is difficult to show that patterns involving small groups of codons are real. There are more fundamental problems with Wong's analysis, however. First, some of the alleged product-precursor pairs involve running metabolic pathways backwards: the conversion from Thr to Met would have to proceed via the common intermediate homoserine, but getting to this state from Thr would require reversal of two steps normally coupled to ATP hydrolysis. Second, Wong assumes that all codons can change independently, yet no base at the wobble position of the anticodon in tRNA can distinguish NNU from NNC. Thus there are really only 48 codon blocks that can change independently, not 64, which means that adjacent NNY block represent only one event and not two. When these problems are corrected, the probability of observing as many adjacent product-precursor pairs by chance alone rises to 16.8%. When additional pathways observed in *E. coli* but not considered by Wong are added, the probability rises to 62% (Ronneberg, Landweber et al. 2000). Thus this particular pathway of code evolution has no statistical support.

A more convincing biosynthetic pattern in the code is that amino acids with the same 1^{st}-position base tend to be metabolically related (Dillon 1973). Specifically, amino acids with A at the first position are derived from Asp; those with C at the first position are derived from Glu; those with U at the first position are derived from intermediates in glycolysis; and those with G at the first position are both at the heads of metabolic pathways and are plausibly prebiotic (Miseta 1989; Taylor and Coates 1989). Could amino acids have been constrained such that their metabolic pathway determined their 1^{st}-position base? If so, could the patterns of similarities in the code be explained on the structural grounds that biosynthetically related amino acids are likely to have similar properties?

These questions can be directly addressed by comparing the actual code to random codes generated by partitioning amino acids to classes related by 1^{st}-position base in the actual code, and randomizing only within each of these classes. This reduces the number of possible codes from 20! to $(5!)^4$ or about 2×10^9 possible codes, a factor of about 10^9. These sets are so different in size that they are effectively independent: even if the code is 1 in a million, every single one of the biosynthetically constrained set could be better than the actual code! However, this is not the case. Although the distributions of constrained and unconstrained codes are significantly different, the differences are rather small, and the code still appears better than nearly every constrained code, for both polar requirement and PAM74-100 and over a wide range of transition/transversion biases (Freeland and Hurst 1998; Freeland, Knight et al. 2000). Consequently, even if history really did constrain the first-position base absolutely, error-minimizing optimization would still be required to explain the values of the second- and third-position bases assigned to the codons for each amino acid.

It may be possible to reconstruct the order in which some amino acids were added from phylogenies of modern components of the translation apparatus. In particular, the tRNAs and aminoacyl-tRNA synthetases form families that clearly show common descent. However, research along these lines has not provided much encouragement to date. tRNA phylogenies are notoriously incongruent (Fitch and Upper 1987; Eigen, Lindemann et al. 1989; Dick and Schamel 1995; Nicholas and McClain 1995; Saks and Sampson 1995; Rodin, Rodin et al. 1996; Ribas de Pouplana, Turner et al. 1998; Chaley, Korotkov et al. 1999), in part because tRNAs can switch isoacceptor classes by as little as a single point mutation (Saks, Sampson et al. 1998). Although tRNAs may well date back to the RNA world (Eigen and Winkler-Oswatitsch 1981; Maizels and Weiner 1987), and thus pre-date the protein aminoacyl-tRNA synthetases (Nagel and Doolittle 1995; Wetzel 1995; Ribas de Pouplana, Turner et al. 1998), they probably do not contain enough information to reliably reconstruct the pattern of duplication and divergence that led to the incorporation of new amino acids. The aminoacyl-tRNA synthetases (aaRS), which fall into two distinct classes, provide better prospects for reconstruction (Nagel and Doolittle 1991; Nagel and Doolittle 1995). Interestingly, GlnRS is derived from eukaryotic GluRS (Lamour, Quevillon et al. 1994; Rogers and Soll 1995; Siatecka, Rozek et al. 1998; Brown and Doolittle 1999); AsnRS is probably derived from eukaryotic/archaeal AspRS (Gatti and Tzagoloff 1991; Cusack 1993; Shiba, Motegi et al. 1998); and TyrRS may be paraphyletic to TrpRS (Ribas de Pouplana, Frugier et al. 1996; Philippe and Forterre 1999, but see Brown, Robb et al. 1997). If amino acids were added to the code on the timescale during which the aaRS duplicated and diverged, ancestral sequences at earlier nodes should not contain later amino acids. We have preliminary evidence that ancestral sites in these three pairs of synthetases systematically underrepresent the later amino acids (Knight, unpublished data), but whether this effect is due to differential ancestry or differential conservation (D. Brooks, personal communication) remains to be established.

In summary, the code probably evolved from a simpler form, although the exact pathway is still unknown. However, even if the code contains traces of its evolutionary history (such as the association between first-position base and metabolic pathway), they cannot explain why it seems to minimize errors to the extent that it does.

2.4.9 Conclusions

The existence of variant genetic codes proves that the genetic code is not fixed, but rather can and does change. It therefore seems reasonable to assume that the code has changed between the time tRNAs were invented and the last common ancestor of extant life, and that at least some of this change has been adaptive, especially in light of the fact that almost no codes minimize genetic errors (in terms of the difference in polarity between the intended and accidentally inserted amino acid) than does the actual code.

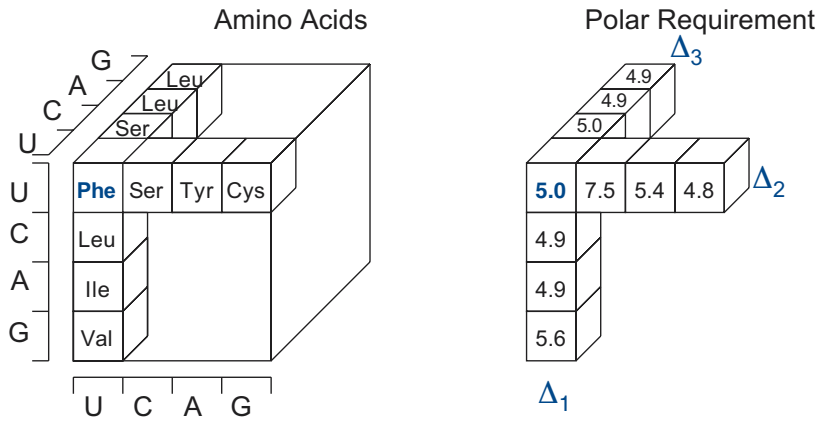
Although there is still debate about exactly how well the code minimized errors in polarity relative to the possible alternatives, most current disagreement stems from inappropriate distance estimates: it is no use comparing the actual code to the best of all possible codes if there is no pathway by which the optimum could be reached, and it is unclear that any single measure of polarity will recapture the actual effect of substitutions in proteins sufficiently accurately that we would expect the code to be adapted to it alone. It is necessary to take into account the frequency distribution of possible codes, rather than to naively assume that optimization will be linear throughout the whole range.

Whether the choice of components of the genetic code was optimized by natural selection is less clear. There are good chemical reasons why certain bases and amino acids were not used, but the presence and absence of others is still a mystery. Simple chemical rules based on GC-content explain the pattern of degeneracy in codon blocks, although it is unclear whether these are the actual causes of the degeneracy or merely proximal mechanisms by which adaptive rules based on translation speed and/or accuracy are enforced. Different

species use markedly different amino acid compositions, which are highly correlated with their overall genome nucleotide composition, and codons are used at highly unequal frequencies, so it seems unlikely that the number of codons assigned to each amino acid gives a unique optimum amino acid composition for proteins. However, this entire area needs attention.

Another big question that remains is the evolutionary pathway of amino acid addition. In particular, was the set of amino acids chosen to fit both stereochemical and selective constraints? It is likely that the wealth of sequence data, and, in particular, sequences of the aminoacyl-tRNA synthetases, will provide a definitive answer within the next decade. However, the finding that the genetic code appears highly optimal relative to other possible codes should no longer be controversial: we can now focus on precisely which properties were optimized, and when in evolution this optimization occurred.

Calculating the 'error value' Δ of a code



$$\Delta_1 \text{ for Codon UUU} = \frac{(5.0 - 4.9)^2 + (5.0 - 4.9)^2 + (5.0 - 5.6)^2}{3}$$

Figure 1: Calculating the error value of a code. Assign a value (in this case, Polar Requirement) to each amino acid, or to each pair of amino acids, to generate a distance matrix showing the magnitude of effect of each type of substitution. Then average the effects of all the substitutions over the whole code, optionally weighting for different types of mutation, or using a different modular power (e.g. linear instead of squared deviations). In this example, the distance for the UUU to CUU Phe to Leu transition is |5.0 – 4.9| = 0.1, which contributes to the first-position transition error. Distance matrices can be constructed from a linear measure, such as polar requirement, or from substitution matrices calculated from observed substitutions in proteins.

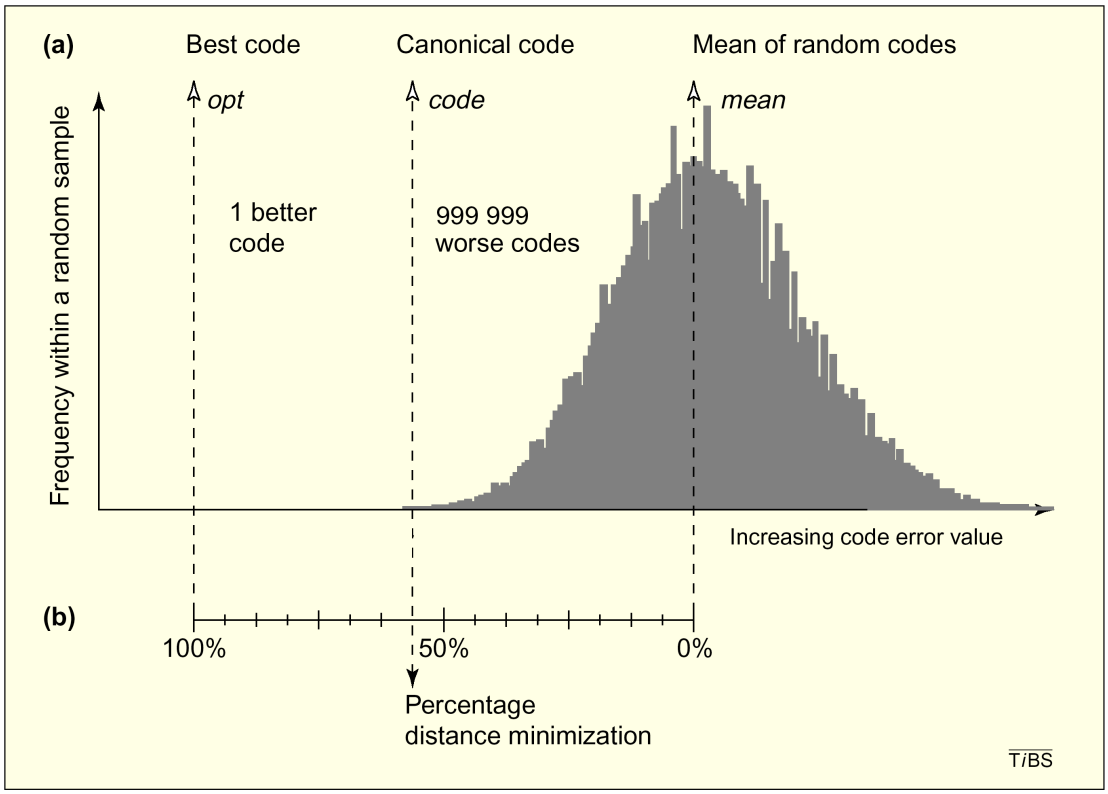


Figure 2: A comparison of methods. (a) The statistical (sampling) approach provides a direct estimate of the probability that a code as good as that used by Nature would evolve by chance; (b) the engineering approach measures code optimality on a linear scale. The figure illustrates the very different values that resulted from using the two methods for the same set of random codes.

2.5 When Did the Genetic Code Adapt to its Environment?

The previous chapter reviews evidence that the genetic code is arranged so that certain types of errors, especially those involving insertion of amino acids with different polarities, are minimized. Interestingly, this is true for two very different measures of amino acid similarity which are not themselves highly correlated. This provides the opportunity to test whether the code is adapted to an RNA world (clusters similar free amino acids, perhaps used as cofactors), to a simple protein world (should cluster amino acids with similar side-chain properties), or to modern proteins (should cluster amino acids with similar conformational preferences). Here I test the apparent optimality of the code to several different measures of amino acid similarity, including a re-measurement of Polar Requirement (with explicitly determined experimental error) using modern thin-layer chromatography instead of paper chromatography. I also test whether the code structures of mitochondrial variants are adapted to their peculiar environment.

2.5.1 Abstract

The genetic code appears close to optimal for error minimization, whether the distance from amino acids is measured *a priori* on chemical grounds, such as polar requirement, or *a posterori* by actual substitution frequencies in proteins. Here we ask whether it is possible to test *when* the code was most recently adapted: when similar codons are assigned to similar amino acids, are the amino acids most similar as free zwitterions, as side-chains, or in the context of evolved proteins? We update polar requirement with thin-layer chromatography to determine whether specific interactions with the ring nitrogen increase or decrease the apparent similarity of adjacent amino acids, and compare the apparent optimality of the canonical and recent variant codes using a variety of measures taken from the AAIndex database of amino acid properties. We find that the code appears most optimal using measures based on the free amino acids, then on properties of the side-chains in simple contexts, and that most properties derived from modern contexts do not show any optimization: this suggests that the optimization may have occurred early in code evolution, perhaps even before the evolution of protein synthesis itself. We also show that the use of substitution matrices to assess code optimality is circular, because these substitution matrices are sufficiently correlated with the code at long substitution distances to make the code appear optimal simply because amino acids that are clustered together in the code are more likely to substitute for one another under mutation. We find little evidence that mitochondrial codes are, in general, specific adaptations for error minimization in the high-transition mitochondrial environment.

2.5.2 Introduction

The genetic code is organized in a way that minimizes the difference in polarity introduced by genetic errors (mutation and mistranslation). This order in the code was noted almost as soon as the codon assignments were elucidated (Speyer, Lengyel et al. 1963; Pelc 1965; Sonneborn 1965; Volkenstein 1965; Woese 1965; Epstein 1966; Fitch 1966; Goldberg and Wittes 1966). In particular, the genetic code minimizes the difference in polarity between the intended and misincorporated amino acid far better than would be expected by chance. This result is robust, as measured by a variety of techniques. First, the code can be compared to a sample of random genetic codes with the same structure (Alff-Steinberger 1969; Haig and Hurst 1991; Ardell 1998; Freeland and Hurst 1998; Freeland and Hurst 1998; Haig and Hurst 1999; Freeland, Knight et al. 2000). Second, the code-dependent likelihood of substitutions can be correlated with the difference in polarity between the original and new amino acid (Wolfenden, Andersson et al. 1981; Sitaramam 1989; Szathmáry and Zintzaras 1992; Joshi, Korde et al. 1993; Benner, Cohen et al. 1994; Tomii and Kanehisa 1996; Koshi and Goldstein

1997; Trinquier and Sanejouand 1998; Xia and Li 1998). Third, the features most related to codon assignment can be determined by dimensionality reduction techniques (Jiménez-Montaño 1994; Tolstrup, Toftgard et al. 1994; Trinquier and Sanejouand 1998).

However, this result that the code is organized in a way that minimizes the effect of genetic error is still open to interpretation. In particular:

1. Is the apparent optimality of the code an artifact of specific measures used? In particular, the polar requirement measure of hydrophobicity (Woese, Dugre et al. 1966; Woese, Dugre et al. 1966; Woese 1973) was specifically chosen because it grouped 'related' amino acids together in the code, and the PAM74-100 substitution matrix (Benner, Cohen et al. 1994), which measures accepted substitutions in highly-diverged protein families, is itself correlated with the genetic code matrix (Di Giulio 2001). Thus the two measures that have given the highest estimates of code optimality (Haig and Hurst 1991; Ardell 1998; Freeland and Hurst 1998; Freeland and Hurst 1998; Haig and Hurst 1999; Freeland, Knight et al. 2000) may actually be circular, rather than functioning independently as *a priori* and *a posteriori* measures of amino acid similarity.

2. All recent comparisons of the actual genetic code with randomized genetic codes have shuffled the amino acids across the codon blocks, which keep the same structure as in the standard genetic code (Ardell 1998; Freeland and Hurst 1998; Freeland and Hurst 1998; Freeland, Knight et al. 2000). However, this may be an unfair comparison, since the actual code is compared against codes in which very unusual amino acids, such as Trp, are assigned far more codons than in the standard code. It is also possible that the number of codons assigned to each amino acid has functional significance. Does the code's apparent optimality change when it is compared only against codes with similar degeneracy for each amino acid?

3. Selection is not the only explanation for the order in the genetic code: it is also possible that similar amino acids were assigned to similar codons because of biosynthetic relatedness (Dillon 1973; Wong 1975; Wong 1981; Di Giulio 1989; Miseta 1989; Taylor and Coates 1989; Di Giulio 1991; Di Giulio 1997; Di Giulio 1998; Di Giulio 1999; Di Giulio and Medugno 2000) or stereochemical constraints (Woese 1965; Woese, Dugre et al. 1966; Woese, Dugre et al. 1966; Yarus and Christian 1989; Yarus 1991; Knight and Landweber 1998; Yarus 1998; Knight and Landweber 2000; Yarus 2000; for review see Knight, Freeland et al. 1999). The many measures of amino acid similarity now available may allow us to test some of these hypotheses. If the genetic code were established in the RNA world by stereochemistry, or for binding free amino acids as cofactors (Szathmáry 1993; Szathmáry 1999), the code should be most conservative for the hydrophobicity of free amino acids. If the code evolved later, when simple peptides were being produced, side-chain property should be more important. If the code was optimized for making modern proteins, protein-based parameters (e.g. sheet and helix preferences) should be most important. Finally, if the recent variant genetic codes (see Knight, Freeland et al. 2001 for review) are still evolving adaptively, variant genetic codes should in general appear more optimal than the standard code, and the mitochondrial codes should appear more optimal in the high-transition mitochondrial environment.

2.5.3 Materials and Methods

We measured polar requirement as described(Woese, Dugre et al. 1966; Woese, Dugre et al. 1966), but using thin-layer chromatography on cellulose instead of paper chromatography and with pyridine instead of 2,6-dimethylpyridine. Pyridine/water mixtures ranged from 40% to 75% mol fraction water. We used the 20 standard protein amino acids and 12 non-protein amino acids: alpha-aminobutyric acid, Citrulline, Homoarginine, Homoglutamine, Homophenylalanine, Homoserine, Norleucine, Norvaline, Ornithine, Pipecolic Acid, Phosphoserine, Hydroxyproline (all reagents purchased from Sigma). Solvent was left in chamber to equilibrate for at least 30 min before runs. We spotted 0.5 uL of 0.1M amino acid solution onto 20 x 20 cm cellulose TLC plates (Sigma catalog number Z12,287-4). Amino

acids with lower solubility (Cys, Tyr, Phe, Trp, Homo-Phe) required multiple spots for sufficient concentration for visualization with ninhydrin (0.4% wt/vol in ethanol): up to 10 spots in the case of Tyr and Homo-Phe. We did three separate experiments, each covering approximately the same range of pyridine/water mixtures. For the first series of runs, we spotted protein amino acids on one plate and the nonprotein amino acids on another, but later used one plate for all amino acids. We ran all plates in duplicate (2 x protein and 2 x nonprotein; later, 2 plates) in the same chamber.

Amino acid properties were taken from AAIndex (Kawashima and Kanehisa 2000), which was downloaded from ftp://ftp.genome.ad.jp/db/genomenet/aaindex/aaindex1. This database contains 434 linear measures of amino acid properties, derived by a wide variety of techniques. We wrote an ISO C program to import the database into Microsoft Excel for easy determination of correlation between measures. The xls format database, with correction of several typographic errors that made import difficult, is available from the authors by request.

We used GenView and Gencode (Freeland, Knight et al. 2000, Ronneberg et al. Forthcoming in Bioinformatics) for testing the optimality of the genetic code against random samples of genetic codes generated by permuting the amino acid properties across codon blocks. Miscellaneous statistical analyses and graphing were performed in Microsoft Excel.

2.5.4 Results/Discussion

Thin-layer chromatography: We did three separate runs, in August 1999, November 2000 and December 2000. The relevant parameter is RF, which is the slope of the log mobility relative to the solvent front as a function of mol fraction water. Although the slopes were very stable within runs (nearly all r^2>0.9; many > 0.95), they were not always consistent between runs: in particular, the summer run is quite different for the hydrophilic amino acids (giving much greater mobilities) (Fig. 1). Thus we averaged the slopes for the three runs to get a consensus rank order, but the absolute magnitudes may not be reproducible. It is possible that differences in temperature or humidity contributed to these differences.

The cysteine sample was oxidized in the December run, a problem that has affected previous measures (Woese, Dugre et al. 1966). We added 0.1M DTT to prevent the problem from recurring. This erroneous result was clearly identifiable because of the anomalously high polar requirement (the slope was near that of Asp).We discarded this result and averaged the remaining two slopes. For all other measures, we averaged all three slopes, combining the standard errors as the square root of the average of the squared individual standard errors.

Our results (Table 1) are highly correlated with Woese's results (r = 0.95) (Fig. 2). However, despite this high correlation, we find from 10 to 40 times as many better codes with the new measurements as with polar requirement (Fig. 3). This result is surprising, since pyridine should be a better mimic of interactions with specific bases than the 2,6-dimethylpyridine used for polar requirement: the methyl groups block the ring nitrogen, preventing the amino acid from interacting with the heterocyclic atom and making the measure more strictly one of polarity (Woese, Dugre et al. 1966). We expected that, if the base/codon associations were established early in evolution, pyridine would make the code appear more optimal.

PAM Matrix Comparisons: Polar requirement measures the properties of the free amino acids, and can be considered an *a priori* measure of amino acid similarity. To discover the conditions to which the code is adapted, it would be better to get a measure of similarity in actual proteins. One such method is to construct a substitution matrix showing which mutations have been accepted into real proteins, giving an *a posteriori* measure (Freeland, Knight et al. 2000). However, there is a risk of contamination by the code structure: amino acids that can be interconverted by single point mutations will replace each other more frequently. Substitution matrices for proteins at different levels of divergence, and for the code itself, are available (Benner, Cohen et al. 1994). At high divergences, interconversions

between dissimilar amino acids that are connected by point mutations (such as Arg/Trp) are less frequent, suggesting that physical properties overcome the effects of the code.

The PAM74-100 matrix is highly correlated with some measures of hydrophobicity, such as Fauchere's hydrophobicity measure ($r^2 = 0.45$) (Fauchere and Pliska 1983; Tomii and Kanehisa 1996), but the 190-element distance matrix derived from polar requirement is only 50% correlated w/ PAM74-100 ($r^2 = 0.25$). Thus it is plausible that PAM74-100 provides a measure of hydrophobicity in distantly diverged proteins that is closer to the property that the genetic code actually optimizes, and presents an independent criterion for assessing code optimality. However, the PAM74-100 matrix is also significantly correlated with a substitution matrix derived from the genetic code itself (Benner, Cohen et al. 1994; Di Giulio 2001). Although this correlation is weak, and explains less than 30% of the variance (Fig. 4), it is possible that the apparent optimality of the code with PAM74-100 derives solely from the fact that it (partly) reflects the structure of the code itself (Di Giulio 2001).

In order to test the circularity of the PAM matrix, we generated random PAM matrices by starting with the genetic code matrix as described (Benner, Cohen et al. 1994), adding a Gaussian deviate to each element, scaling the mean and variance of the elements to the mean and variance of the elements in PAM74-100, and testing the apparent optimality of the code with these matrices (see Fig. 4 for example). We find that randomly generated substitution matrices that are as weakly correlated with the code matrix as is PAM74-100 make the code look just as optimal. In fact, even very weakly correlated matrices 'contaminated' with the code matrix make the code appear far better than random codes (Table 2). In fact, no codes better than the actual code were found in a sample of 1 million, using either PAM74-100 or either of two random matrices at the same correlation distances (although the pattern of errors differed markedly, with PAM74-100 concentrating the errors at the second position, while the random matrices distributed them across all three positions). Consequently, the apparent optimality of the code with PAM74-100 can be explained entirely in terms of its correlation with the code's structure, and this measure should not be used for future analyses of code optimization.

Different Measures of Amino Acid Similarity: Although there is a consensus result that the genetic code is optimal to some type of polarity, the exact type may allow us to elucidate when the code was optimized. Counterintuitively, the properties of the side-chains are context-dependent: although glycine is very soluble as a free amino acid, since its character is dominated by the polar amino and carboxyl groups, polyglycine is insoluble because its character is dominated by the polyamide bond. Measurements of amino acid properties fall into several categories, which may be relevant to code evolution (Table 3). Redundant measures were weeded from some categories: for example, we did not test all of the compositional indices in the database, especially since the sample we tested showed that these properties are clearly not important in structuring the code.

For each measure in Table 3, we generated a sample of 100 000 random codes, using a modular power of 2, and a transition bias of 2 (roughly consistent with that found in nuclear lineages), and counted the number of random codes that appeared better than the standard code. All measures were normalized to a mean of 0 and a standard deviation of 1; this did not affect the apparent optimality of well-characterized measures, such as polar requirement (37 vs. 38 better codes found in a sample of 1 million). We also constructed averaged indices for each category, which were the normalized averages of the normalized index scores.

Interestingly, indices that made the code look highly optimal were scattered through the categories. However, the code appeared increasingly optimal with earlier scales: all of the scales derived from free amino acids made it look highly optimal, most scales derived from amino acid side-chains made it look at least moderately optimal, and few of the scales derived from modern proteins showed any evidence of optimization at all (Table 3). In particular, there is no evidence that the code is optimized relative to indices that track the frequency of different

amino acids in different components of proteins (e.g. helix vs. sheet preferences, or location in transmembrane domains), and even solvent accessibility only appears weakly optimized.

Of the side-chain properties, those based on analogs of the side-chain that lack the amino and carboxyl group may be considered dubious. In particular, these model Gly as free hydrogen, Ser as methanol, etc., and look at the partitioning of these often volatile compounds between phases (Wolfenden, Andersson et al. 1981). If we exclude these measures (WOLR810101, BULH740101, LAWE840101, RADA880101–05, VHEG790101) from the analysis, the new average only produces 48 better codes in a sample of 100 000 (instead of 1573 for the combined measure). This still does not beat the measures derived from the free amino acids, although it becomes much better than any of the measures derived from modern proteins.

The apparent optimality of the different measures was not explained by their correlation with the most optimal measure, Grantham's polarity (Grantham 1974). Although the general trend was that the most optimal measures were more highly correlated, even rather strongly associated measures ($r^2 > 0.8$) could lack optimization completely (Fig. 5). The fact that some very weakly correlated measures, such as Trifonov's consensus scale of the amino acids (Trifonov 2000), made the code look highly optimal suggests that optimization may have proceeded along several distinct dimensions. However, the result that the code is optimized with respect to polarity is remarkably robust to the measure of polarity used (except for properties derived from modern proteins).

Optimality of Variant Codes: Although most attention has focused on the adaptive evolution of the canonical code, modern variant codes may still be evolving by natural selection. The transition/transversion ratio is much higher in mitochondria than in nuclei (Brown and Simpson 1982), so we might expect that variant codes would appear more optimal than the standard code at higher transition biases. We previously found that all variant codes looked less optimal than the standard code using no transition/transversion bias and polar requirement as the measurement (Freeland, Knight et al. 2000). Here, we know that all variant codes are recent variants of the canonical code, and so measures of free amino acids such as polar requirement are not appropriate. Similarly, measures derived from proteins may not be appropriate since mitochondrial proteins are a small and unusual subset, especially in metazoa. Instead, we use Fauchere's measure of side-chain hydrophobicity based on the doubly amidated amino acids (Fauchere and Pliska 1983), which should provide a good estimate of the average substitution effect in relatively simple structural contexts. We used a modular power of 2 (squared deviations) for consistency with the results from different measures tested against the standard code.

All variant codes do worse than the standard code at transition biases above 1 using these parameters, except for the yeast mitochondrial code, which does much better at all transition biases (Fig. 6). Codes found only in nuclear lineages (the two ciliate nuclear genetic codes and the variant yeast nuclear genetic code) do not behave differently from the mitochondrial codes. Interestingly, the variant code with UGA = Trp, which has evolved independently several times in both nuclear and mitochondrial lineages (Knight, Freeland et al. 2001), is unequivocally worse than any other variant code under all conditions. This suggests that this frequently recurring variant is not driven by selection for error minimization.

The comparison of the standard code with other codes where the number of codons assigned to each amino acid might be unfair: better codes might assign the most extreme amino acids to blocks with fewest codons. This concern can be addressed by restricting the possible reassignments such that 'family boxes' of four codons can only interchange with other family boxes, and blocks of one, two, or three amino acids can interchange freely (the fact that these smaller groups of codons are very labile in different variant codes suggests that they form a single class: for instance, Cys and Ile have 2 or 3 codons in different variants, and Met and Trp have 1 or 2). However, this is not the case: for the standard code, and for most variants, this restricted set actually makes the standard code (and its known variants) look better. For example, the number of codes better than the standard code drops from 750 to 265 in 1

million better (using Fauchere's measure of side-chain hydrophobicity, a transition bias of 2, and a modular power of 2). Thus, the apparent optimality of the code and of known variants cannot be explained by the number of amino acids assigned to each codon.

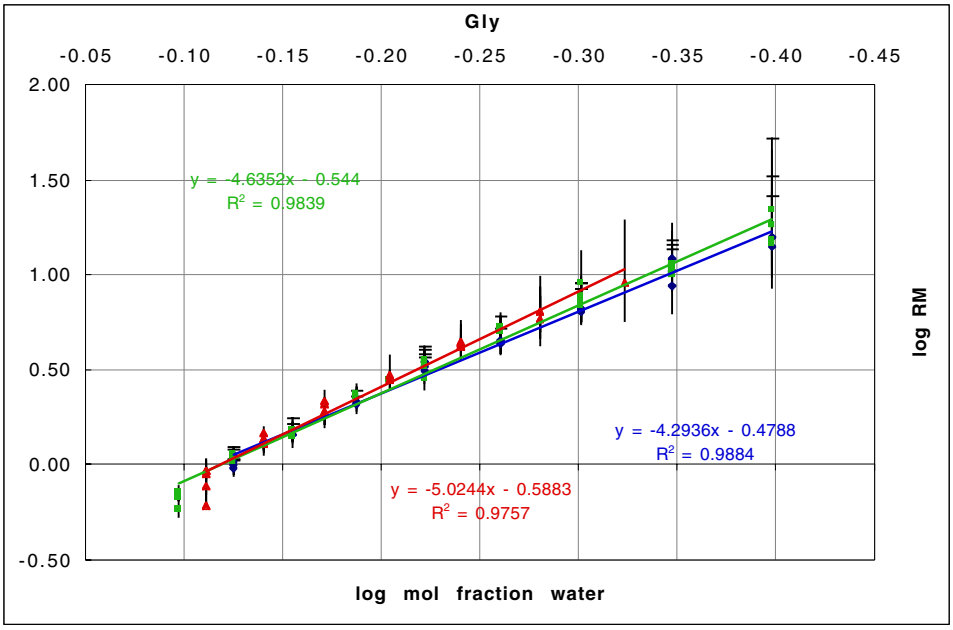
2.5.5 Conclusions

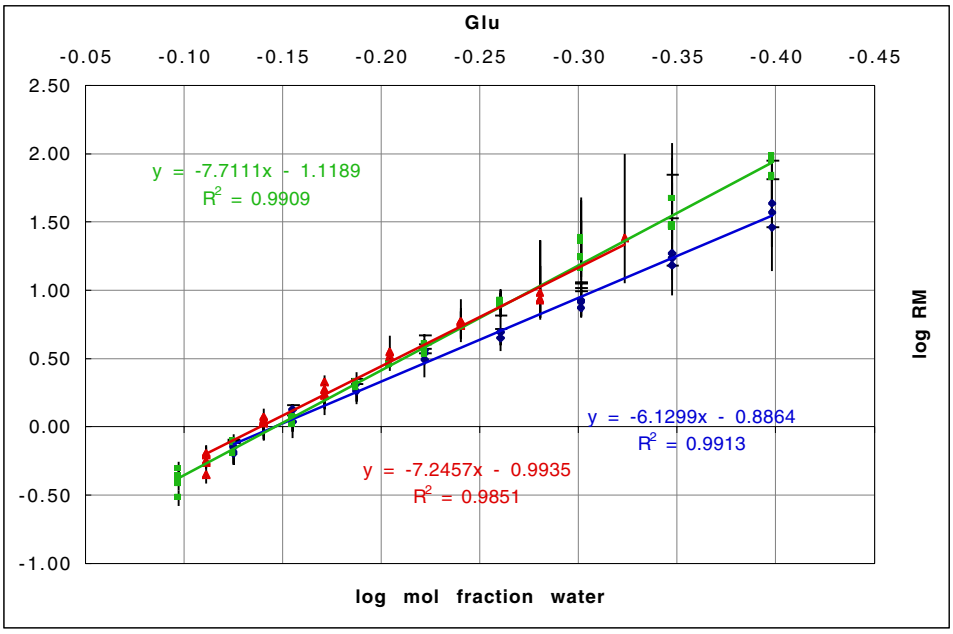
The genetic code appears more optimal with properties derived from the free amino acids or the side-chains than with properties derived from modern proteins. This suggests that the code is not adapted to modern proteins, but instead was fixed at an earlier stage of evolution. Although the general trend is clearly in this direction, there is a lot of scatter: one of the measures that makes the code look most optimal is the 'long-range non-bonded energy per atom' (Oobatake and Ooi 1977; Sitaramam 1989), a measure derived from a sample of modern proteins. It is difficult to see how this property could be more important than, for instance, size or charge. We can rule out the possibility that the code's apparent optimality with several different measures merely reflects the fact that the measures are partly correlated with each other. However, more precise measurements of the behaviors in model solvent systems of the free amino acids, and of their side-chains in simple compounds, are required to discriminate between the hypotheses that the code is optimized to conserve the properties of the free amino acids vs. the side-chains in simple peptides.

The PAM74-100 matrix, or any other substitution matrix derived from proteins, should not be used for assessing code optimality. Even a very small component derived from the code's structure, which affects which amino acids can interconvert easily, can have a huge effect on the apparent distance minimization of the code. Despite this, the result that the code is organized to minimize differences in polarity is remarkably robust, both to the exact measure used and over a wide range of parameter space (Freeland, Knight et al. 2000). The fact that the code appears optimal with several measures that are not themselves highly correlated suggests that some multidimensional model of optimization is required.

Finally, we did not find that variant codes are in general better at minimizing errors than the standard code, although the property measured has a substantial effect on the ranking of particular variants. It is possible that a property other than Fauchere's measure or polar requirement would better recapture the salient features of amino acid similarity in the set of mitochondrial proteins. It is also possible that a model incorporating the full 12-element substitution matrix, rather than just the transition/transversion bias, would reveal these codes to be adaptive. It is also possible that taking into account translational accuracy would make mitochondrial codes appear more optimal: unfortunately, these data are unavailable. However, the available evidence does not exclude the possibility that these variant codes are simply neutral (or mildly deleterious) mutants of the standard code.

Figure 1: Raw data for two amino acids, Gly (1a) and Glu (1b). The summer run (blue) is significantly different from the others for the very polar amino acid Glu, but not for Gly (which has a nonpolar side-chain). Error bars represent the top and bottom of the visible area of each spot; the central point is taken as the darkest point in the spot as measured by visual inspection.





	Set 1			Set 2			Set 3			Summary		Polar
aa	slope	r^2	SE	slope	r^2	SE	slope	r^2	SE	**Slope**	**SE**	Req.
Ala	-3.27	0.98	0.11	-3.81	0.98	0.13	-3.77	0.97	0.12	**3.62**	**0.12**	7.00
Arg	-5.17	0.97	0.18	-4.43	0.86	0.37	-4.29	0.94	0.19	**4.63**	**0.26**	9.10
Asn	-4.27	0.99	0.10	-4.79	0.97	0.17	-4.62	0.97	0.14	**4.56**	**0.14**	10.00
Asp	-6.16	1.00	0.09	-8.87	0.98	0.26	-8.46	0.99	0.18	**7.83**	**0.19**	13.00
Cys	-3.04	0.81	0.32	-2.30	0.68	0.33	*-7.15*	*0.98*	*0.16*	**2.67**	**0.35**	4.80
Gln	-4.14	0.99	0.09	-4.51	0.98	0.13	-4.41	0.98	0.12	**4.35**	**0.12**	8.60
Glu	-6.13	0.99	0.12	-7.25	0.99	0.19	-7.71	0.99	0.13	**7.03**	**0.15**	12.50
Gly	-4.29	0.99	0.10	-5.02	0.98	0.17	-4.64	0.98	0.10	**4.65**	**0.13**	7.90
His	-4.05	0.98	0.11	-4.71	0.98	0.15	-4.56	0.98	0.12	**4.44**	**0.13**	8.40
Ile	-2.82	0.95	0.14	-3.03	0.92	0.18	-2.87	0.93	0.14	**2.91**	**0.16**	4.90
Leu	-3.00	0.94	0.16	-3.36	0.92	0.20	-2.87	0.91	0.16	**3.07**	**0.18**	4.90
Lys	-4.57	0.95	0.23	-4.70	0.87	0.39	-4.66	0.96	0.17	**4.64**	**0.28**	10.10
Met	-3.15	0.96	0.14	-3.25	0.93	0.18	-3.14	0.93	0.17	**3.18**	**0.16**	5.30
Phe	-3.05	0.94	0.16	-2.89	0.93	0.17	-2.93	0.92	0.16	**2.96**	**0.16**	5.00
Pro	-3.24	0.95	0.16	-3.63	0.96	0.15	-3.22	0.94	0.15	**3.37**	**0.15**	6.60
Ser	-3.44	0.97	0.14	-3.70	0.97	0.14	-3.70	0.97	0.13	**3.61**	**0.13**	7.50
Thr	-3.22	0.96	0.14	-3.21	0.97	0.12	-3.36	0.96	0.13	**3.26**	**0.13**	6.60
Trp	-2.67	0.93	0.15	-2.65	0.91	0.17	-2.82	0.90	0.18	**2.71**	**0.17**	5.20
Tyr	-3.18	0.91	0.21	-3.23	0.90	0.25	-3.34	0.88	0.28	**3.25**	**0.25**	5.40
Val	-2.68	0.93	0.16	-3.07	0.94	0.16	-2.98	0.92	0.16	**2.91**	**0.16**	5.60
ABA	-3.02	0.96	0.14	-3.24	0.94	0.17	-3.20	0.95	0.13	**3.16**	**0.15**	
Cit	-4.63	0.99	0.12	-5.05	0.99	0.12	-4.88	0.99	0.10	**4.85**	**0.12**	
hR	-4.58	0.94	0.25	-5.60	0.93	0.32	-4.38	0.95	0.19	**4.86**	**0.26**	
hQ	-4.03	0.98	0.11	-4.43	0.98	0.13	-4.30	0.97	0.12	**4.26**	**0.12**	
hF	-4.44	0.97	0.33	-4.91	0.81	0.60	-3.92	0.94	0.20	**4.43**	**0.41**	
hS	-3.28	0.97	0.12	-3.65	0.98	0.12	-3.41	0.95	0.14	**3.45**	**0.12**	
nL	-3.52	0.94	0.18	-4.16	0.94	0.22	-3.35	0.91	0.21	**3.67**	**0.21**	
nV	-3.09	0.93	0.19	-3.67	0.93	0.21	-3.25	0.93	0.17	**3.34**	**0.19**	
Orn	-4.81	0.95	0.25	-4.51	0.80	0.50	-4.66	0.91	0.26	**4.66**	**0.35**	
hP	-2.93	0.93	0.17	-3.33	0.95	0.16	-3.17	0.94	0.15	**3.14**	**0.16**	
pS	-4.91	0.89	0.38	-4.16	0.42	1.03	-12.58	0.79	2.90	**7.22**	**1.79**	
Hyp	-3.30	0.94	0.35	-3.17	0.94	0.18	-3.78	0.79	0.35	**3.42**	**0.31**	

Table 1: Results for re-measurement of polar requirement using thin-layer chromatography. Abbreviations: ABA, alpha-aminobutyric acid; Cit, citrulline; hX, homo-X (pipecolic acid is hP, or homoproline); nX, nor-X (e.g. nL is norleucine); Orn, ornithine; pS, O-phosphoserine; Hyp, hydroxyproline. Figures in italics for Cys are from the oxidized form, and were excluded from the analysis. Polar req. is Woese's polar requirement (Woese 1973) for comparison. The correlation between our averaged slopes and polar requirement is 0.95.

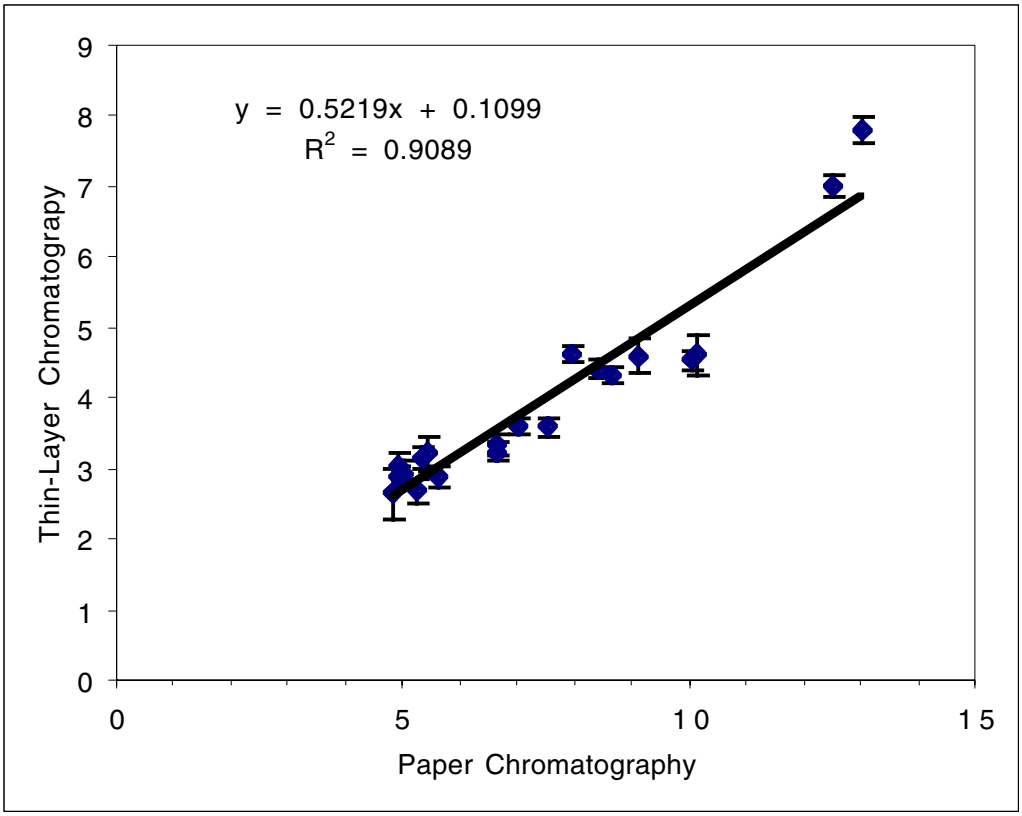


Figure 2: comparison of our results with those derived from paper chromatography (Woese 1973). Error bars show 1 SE. The ordering of many of the amino acids changes between the two measures, and many are unresolved in both cases. The major differences are the acidic amino acids Asp and Glu (disproportionately higher in our measurements), and Lys and Asn (disproportionately lower).

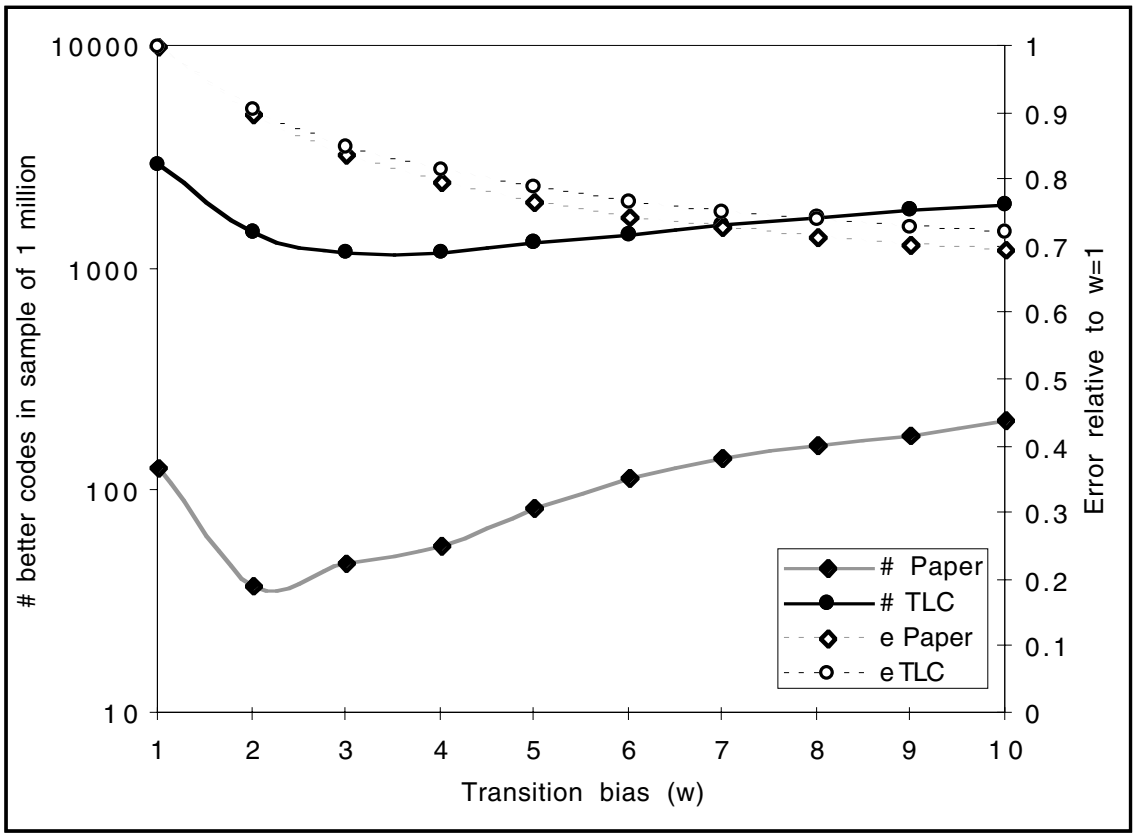


Figure 3: Comparison of paper and thin-layer chromatography results and code optimality. The transition/transversion ratio was varies from 1 to 10 in order to show the effects of increasing transition bias (the modular power was held constant at 2). Solid symbols and lines show the number of better codes than the actual code in a sample of 1 million random codes (left-hand axis); hollow symbols with dashed lines show the ratio of the error value of the actual code to its error value at w = 1 (unbiased mutation). Although the behavior of the code is similar with paper chromatography (diamonds/gray lines) and TLC (circles/black lines), about 20 times as many better codes are found with our new measurements.

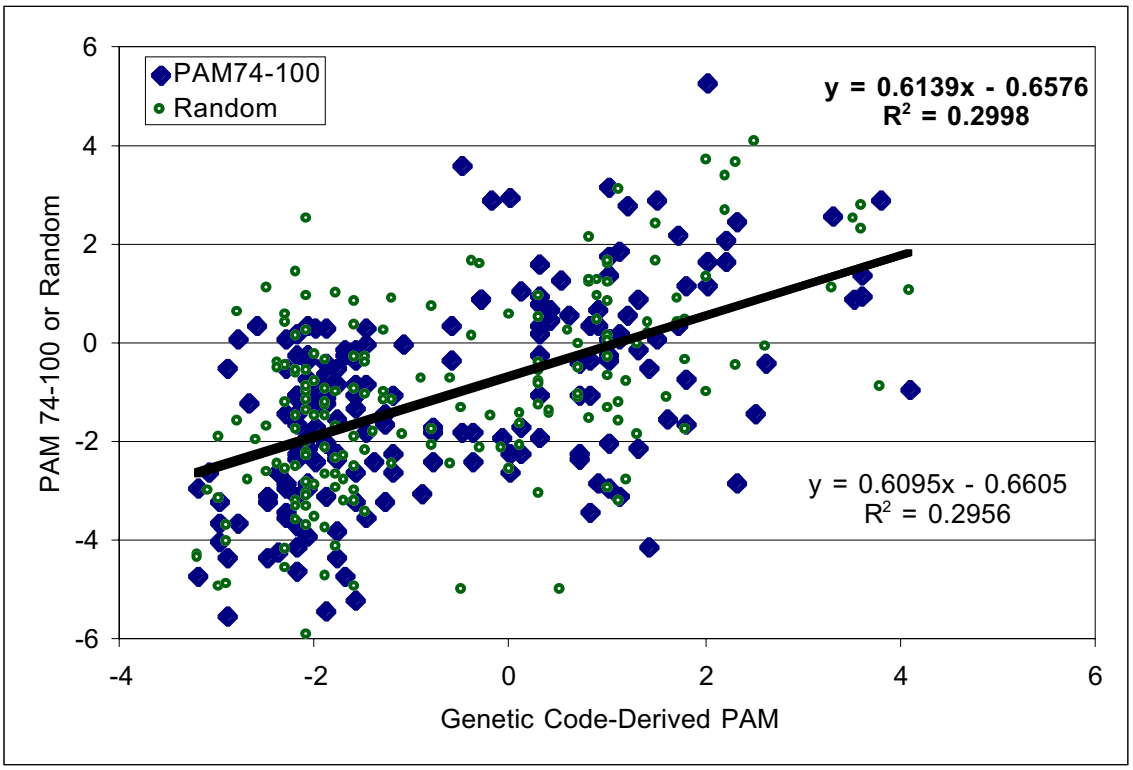


Figure 4: Comparison of the 190 off-diagonal elements of the PAM matrix derived from the genetic code (x axis) with the PAM74-100 matrix (diamonds, bold font) and a random matrix (circles, light font) with almost identical variance, slope and correlation coefficient derived by adding a Gaussian error term to the genetic code PAM. R^2 (PAM74-100,random) = 0.08.

		nat	max	min	num better	mean	var
PAM74-100	**MS[0]**	**0.7845**	**1.2735**	**0.7845**	**0**	**1.05**	**0.05**
	MS[1]	0.8768	1.7705	0.8207	16	1.34	0.11
	MS[2]	1.2622	1.7768	0.9284	69786	1.41	0.09
	MS[3]	0.2155	0.6951	0.1646	247	0.40	0.06
Code	**MS[0]**	**0.5866**	**1.0761**	**0.5866**	**0**	**0.93**	**0.03**
	MS[1]	0.737	1.481	0.737	0	1.20	0.08
	MS[2]	0.8707	1.5091	0.8707	0	1.25	0.07
	MS[3]	0.1537	0.5192	0.1537	0	0.36	0.05
Rnd 30% (a)	**MS[0]**	**0.7911**	**1.3024**	**0.7911**	**0**	**1.05**	**0.05**
	MS[1]	0.9879	1.9184	0.8851	117	1.34	0.11
	MS[2]	1.149	1.9045	1.0006	1728	1.41	0.1
	MS[3]	0.2387	0.7623	0.1719	1903	0.40	0.07
Rnd 30% (b)	**MS[0]**	**0.762**	**1.3131**	**0.762**	**0**	**1.05**	**0.05**
	MS[1]	0.9376	1.8692	0.8681	7	1.34	0.11
	MS[2]	1.1238	1.8815	0.9742	734	1.41	0.1
	MS[3]	0.2265	0.7402	0.1793	699	0.40	0.07
Rnd 12%	**MS[0]**	**0.8459**	**1.2935**	**0.843**	**1**	**1.05**	**0.05**
	MS[1]	1.0139	1.8622	0.8639	452	1.34	0.11
	MS[2]	1.2145	1.8553	1.0019	16514	1.41	0.09
	MS[3]	0.3114	0.7576	0.1714	86077	0.40	0.07
Rnd 1%	**MS[0]**	**0.9838**	**1.3154**	**0.837**	**93703**	**1.05**	**0.05**
	MS[1]	1.2325	1.9641	0.887	160200	1.35	0.12
	MS[2]	1.2752	1.9346	1.0259	80711	1.41	0.1
	MS[3]	0.4466	0.7803	0.1857	752343	0.40	0.07

Table 2: Comparison of actual substitution matrices with randomly derived ones. Row headings: PAM74-100, Benner et al's PAM 74-100 matrix; Code, Benner et al's genetic code-derived matrix; Rnd x%, matrix generated from the code matrix + error term, in which the code matrix explains x% of the variance in the least-squares regression line (i.e. Rnd 30% indicates a random matrix at the same distance as PAM74-100). Two separate matrices at 30% were generated and tested. MS[0], values for the genetic code as a whole; MS[x], values for only changes involving position x. Column headings: nat, the error value for the natural (canonical) genetic code; max, the error value for the best code found; min, the error value for the worst code found; num better, the number of better codes found in a sample of 1 million; mean, the mean error value of the entire sample; var, the variance of code error values in the sample. These error values reflect a modular power of 2 (i.e. squared errors) and a transition/transversion bias of 2.

Property	Measure	AAIndex #/ref	#better
Prebiotic:	Paper chromatography in water/2,6-		
Free aa	dimethylpyridine system	Woese et al. 1966	5
	Thin-layer chromatography with water/pyridine solvent	This study	127
	Grantham's Polarity	GRAR740102	0
	RF value in high salt chromatography	WEBA780101	379
	AVERAGE — FREE AA		**0**
	Free energies of transfer of AcWl-X-LL		
Early Peptides:	peptides from bilayer interface to		
Side-chains	water	WIMW960101	4
	Average of partition coefficients of side-chain analogs for several solvent systems	WOLR810101	69947
	Water/octanol partition coefficient for side-chain in double amide	FAUJ830101	84
	Partition coefficient of amides in TLC system	PLIV810101	291
	Effect of side-chain on retention coefficient in TFA	BROC820101	6597
	Effect of side-chain on retention coefficient in HFBA	BROC820102	2090
	Effect of side-chain on retention coefficient in HPLC, pH7.4	MEEJ800101	66
	Effect of side-chain on retention coefficient in HPLC, pH2.1	MEEJ800102	587
	Effect of side-chain on retention coefficient in NaClO4	MEEJ810101	151
	Effect of side-chain on retention coefficient in NaH2PO4	MEEJ810102	362
	Transfer free energy to surface	BULH740101	624
	Transfer free energy, CHP/water	LAWE840101	12212
	Transfer free energy from chx to wat	RADA880101	58528
	Transfer free energy from oct to wat	RADA880102	4348
	Transfer free energy from vap to chx	RADA880103	3868
	Transfer free energy from chx to oct	RADA880104	72117
	Transfer free energy from vap to oct	RADA880105	69996
	Transfer free energy to lipophilic phase	VHEG790101	60227
	AVERAGE — PEPTIDES		**1573**
Modern Proteins:			
Solvent Accessibility	Accessible surface area in proteins	RADA880106	45780
	Accessible surface area in proteins	JANJ780101	16769
	Proportion of residues 100% buried	CHOC760104	4174
	Proportion of residues 95% buried	CHOC760103	1400
	Membrane-buried preference parameters	ARGP820103	3209
	AVERAGE — SOLVENT ACCESS.		**4607**

Compositions	Membrane domain of multi-spanning proteins	NAKH920108	12736
	Membrane domain of single-spanning proteins	NAKH920105	54367
	Sheet propensity	KANM800102	12730
	Helix propensity	KANM800101	47764
	Beta-strand indices for alpha/beta-proteins	GEIM800107	9807
	Beta-strand indices for beta-proteins	GEIM800106	6860
	Conformational preference for all beta-strands	LIFS790101	4578
	Energy transfer from out to in(95%buried)	RADA880107	55661
	Normalized frequency of alpha-helix	CHOP780201	25062
	Normalized frequency of beta-sheet	CHOP780202	4588
	Normalized frequency of beta-turn	CHOP780203	15683
	Relative frequency in alpha-helix	PRAM900102	59929
	Relative frequency in beta-sheet	PRAM900103	13013
	Relative frequency in reverse-turn	PRAM900104	40342
	Surrounding hydrophobicity in alpha-helix	PONP800104	69373
	Surrounding hydrophobicity in beta-sheet	PONP800105	17215
	AVERAGE — COMPOSITIONS		**8692**
Synthesis Cost	AA composition	NAKH900101	9448
	AA composition	DAYM780101	1190
	AA composition	JUKT750101	5783
	Heat capacity (Hutchens, 1970)	HUTJ700101	3145
	Absolute entropy (Hutchens, 1970)	HUTJ700102	36799
	Sequence frequency (Jungck, 1978)	JUNJ780101	3611
	AVERAGE — SYNTHESIS COST		**36210**
Side-chains	Average non-bonded energy per atom	OOBM770101	998
	Surrounding hydrophobicity	MANP780101	3465
	Long range non-bonded energy per atom	OOBM770103	2
	Side chain hydropathy, corrected for solvation	ROSM880102	1700
	Short and medium range non-bonded energy per atom (Oobatake-Ooi, 1977)	OOBM770102	78309
	AVERAGE — SIDE-CHAINS		**6473**
	AVERAGE – ALL MODERN PROTEINS		**67424**

Consensus Scales	Consensus normalized hydrophobicity scale	EISD840101	41047
	Hydropathy index	KYTJ820101	390
	Hydrophobicity	JOND750101	9128
	Hydrophobicity	PRAM900101	18499
	Hydrophobicity	ZIMJ680101	15718
	Hydrophobicity factor	GOLD730101	8522
	Hydrophobicity index	ARGP820101	8903
	Mean polarity	RADA880108	108
	Normalized average hydrophobicity scales	CIDH920105	163
	Principal property value z1	WOLS870101	27
	AVERAGE — CONSENSUS SCALES		**5959**
	Trifonov's Consensus Scale	Trifonov 2000	18

Table 3: Relationship between amino acid indices, stage of evolution, and apparent code optimality. All indices were scaled to a mean of 0 and a standard deviation of 1, and a sample of 100 000 codes was generated for each index. 'Average' values refer to the arithmetic mean of the relevant indices: the elements of this vector were scaled to a mean of 0 and a standard deviation of 1, and used as input as per the individual indices. Divide by 100 000 to get Pr(better code). The corrected 0.01 cutoff for 73 comparisons is 1.3×10^{-4}, i.e. there is only a 1% probability that we would see *any* values below 13 by chance, while in fact there are 5 such values.

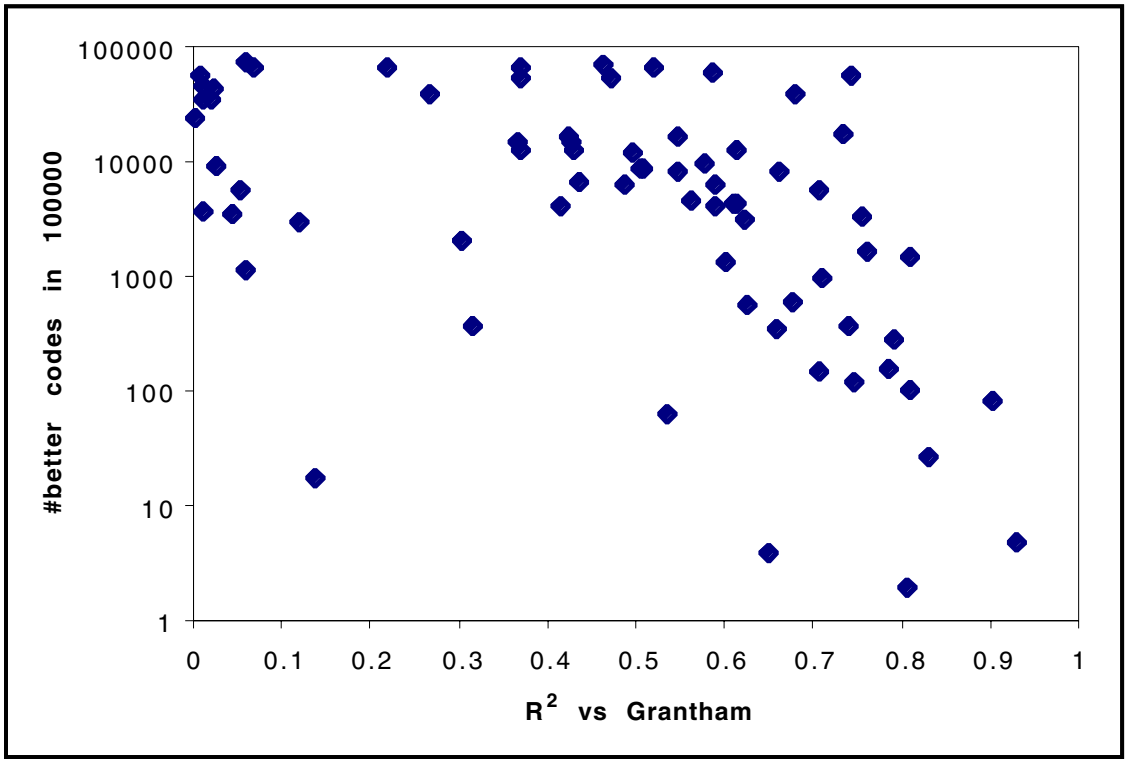


Figure 5: Lack of strong association between correlation with the single best measure and highly optimal codes. Grantham's polarity is the best individual measure (no better codes in 100 000), but the other measures that make the code look highly optimal can be only weakly correlated with it: average for prebiotic measures, no better codes, $r^2 = 0.93$; long-range non-bonded energy per atom, 2 better codes, $r^2 = 0.80$; free energy of transfer of pentapeptides to water, 4 better codes, $r^2 = 0.65$; polar requirement, 5 better codes, $r^2 = 0.92$; Trifonov's consensus temporal order of amino acid entry into the code, 18 better codes, $r^2 = 0.13$.

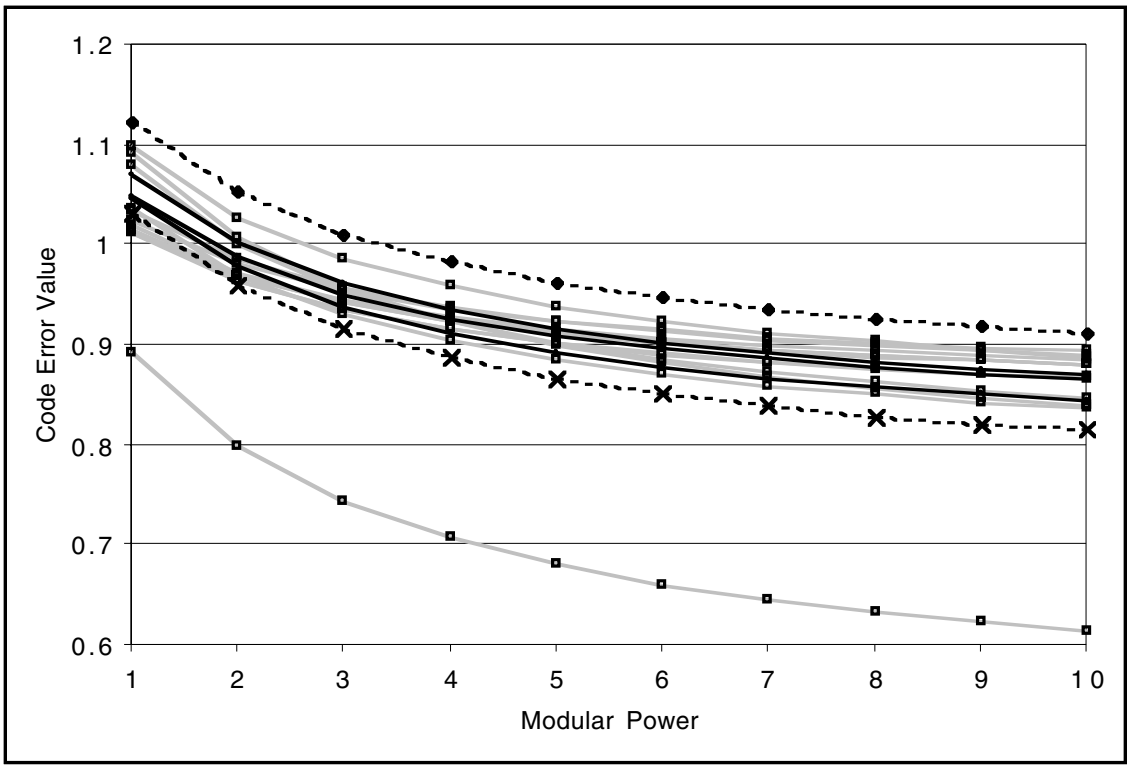


Figure 6: Average error under point substitution (y axis) of variant codes with increasing modular power (x axis). Codes found only in mitochondria are represented as hollow squares connected by gray lines; codes found only in nuclei are solid circles connected by black lines; codes found in both are solid diamonds connected by dashed lines, except for the canonical code which is marked by crosses. All errors measured using Fauchere's index (Fauchere and Pliska 1983).

3 Diversification of Modern Genetic Codes

This section deals with the recent evolution of the genetic code, which has changed in a variety of nuclear and mitochondrial lineages. The recent diversification of genetic codes may be important for understanding the evolution of the canonical code, since similar processes could presumably have occurred between the invention of tRNA and aminoacyl-tRNA synthetases, but is also interesting in its own right.

Chapter 3.1 reviews the mechanisms by which the genetic code can and has changed, and summarizes the vast wealth of detailed information on the biochemical mechanisms of codon reassignment that has accumulated over the last few years. It also introduces the three main hypotheses about the evolutionary mechanisms of recent codon reassignment: Genome Minimization, Codon Capture, and Codon Ambiguity.

Most genetic code changes happen in mitochondria, so it seemed that closer examination of this system would be rewarding. Chapter 3.2 provides explicit, quantitative tests of the three hypotheses, with reference to the hundred or so mitochondrial genomes available in GenBank at the time it was written. Comparative analysis made it possible to rule out the Genome Minimization hypothesis completely (or, at least, to the extent that anything can be ruled out in biology), and showed that codons need not disappear to be reassigned.

Surprisingly many changes in the genetic code have taken place in ciliates. In collaboration with Catherine Lozupone, I analyzed the sequences of the eukaryotic release factor eRF1 to test the hypothesis I advanced in Chapter 1.4, that a single amino acid substitution in the conserved NIKS motif conferred altered stop codon specificity. The reality turned out to be more complicated, but I was able to develop a statistical method for correlating molecular change to phylogeny that surprisingly recaptured results from yeast genetics and x-ray crystallography of the same molecule, in essence using 'Nature's Genetic Screen' to identify mutants with interesting phenotypes in the field instead of in the lab (admittedly, **these** ***mutants were actually different species).***

Finally, the genetic code is not just the pattern of codon assignments: it is also the frequency with which each codon assignment is used. That codon usage is unequal is not just of theoretical interest: heterologous expression of transgenes often requires that the codon usage be matched to the host to give reasonable protein yields, and the black art of primer design owes many of its complications to the difficulty of predicting even 'highly conserved' sequences across species with different, arbitrary codon preferences. In Chapter 1.5 I make what was (to me, at least) the shocking discovery that codon usage, and even amino acid usage, can be explained almost entirely by forces acting at the nucleotide level. In other words, both codon and amino acid usage can be predicted just from the GC content (or, better, the ratio of all four bases) in an organism's genome! Some amino acids, such as arginine, vary 5-fold or more in frequency depending on genome GC content, which has profound implications for molecular evolution.

3.1 Rewiring the Keyboard: Evolvability of the Genetic Code

Since the discovery that human mitochondria use a different genetic code from the nucleus, a bewildering variety of variant genetic codes has been found in organisms and organelles. In this chapter, I review the specific biochemical *mechanisms of these changes. See the following chapter for a more thorough analysis of the* evolutionary *mechanisms involved.*

This paper has previously appeared in the Jan 2001 issue of Nature Reviews Genetics:

Knight, R. D., S. J. Freeland and L. F. Landweber (2001). “Rewiring the keyboard: evolvability of the genetic code.” Nat Rev Genet 2: 49-58.

The paper was primarily my work, although Dr. Freeland’s monumental effort in producing the composite phylogeny of variant codes (and also the somewhat less arduous task of producing Table 2) should be specifically acknowledged. Fig. 1 was contributed by a staff artist at Nature Reviews Genetics.

REWIRING THE KEYBOARD: EVOLVABILITY OF THE GENETIC CODE

Robin D. Knight, Stephen J. Freeland and Laura F. Landweber

The genetic code evolved in two distinct phases. First, the 'canonical' code emerged before the last universal ancestor; subsequently, this code diverged in numerous nuclear and organelle lineages. Here, we examine the distribution and causes of these secondary deviations from the canonical genetic code. The majority of non-standard codes arise from alterations in the tRNA, with most occurring by post-transcriptional modifications, such as base modification or RNA editing, rather than by substitutions within tRNA anticodons.

DIPLOMONADS
Among the earliest-diverging eukaryotes, these unicellular organisms have two nuclei, but lack mitochondria. The gastrointestinal parasite *Giardia* is a diplomonad.

Department of Ecology and Evolutionary Biology, Princeton University, Princeton, New Jersey 08544, USA. e-mails: rdknight@princeton.edu; sfreelan@princeton.edu; lfl@princeton.edu
Correspondence to L.F.L.

The genetic code, which translates nucleotide sequences into amino-acid sequences (FIG. 1), was long thought to be an immutable 'frozen accident', incapable of further evolution even if it were far from optimal[1]. Any change in the genetic code alters the meaning of a codon, which, analogous to reassigning a key on a keyboard, would introduce errors into every translated message. Although this might have been acceptable at the inception of the code, when cells relied on few proteins, the forces that act on modern translation systems are likely to be quite different from those that influenced the origin and early evolution of the code[2,3].

The observation that the vertebrate mitochondrial and nuclear codes differ[4] prompted a search for other variants, several of which have now been found in both nuclear and mitochondrial systems (reviewed in REF. 5). Curiously, many of the same codons are reassigned in independent lineages, frequently between the same two meanings[6], indicating that there may be an underlying predisposition towards certain reassignments. At least one of these changes seems to confer a direct selective advantage[7], showing that the code is evolvable in the formal sense that the mapping between genotype and phenotype allows adaptive changes[8].

Secondary changes in the code pose three problems. First, what are the sources of variability in codon assignments? Second, what constraints, if any, limit changes in the code? And last, what causes a variant code to become fixed in a lineage once it has arisen?

Recent advances in genome sequencing, and in identifying specific base modifications that alter codon–anticodon pairing, allow us to evaluate the hypotheses that have been proposed to explain the mechanisms of codon reassignment.

Where do changes occur?

The genetic code varies in a wide range of organisms (FIG. 2), some of which share no obvious similarities. Sometimes the same change recurs in different lineages: for instance, the UAA and UAG codons have been reassigned from Stop to Gln in some DIPLOMONADS, in several lineages of ciliates and in the green alga *Acetabularia acetabulum* (reviewed in REF. 5). Similarly, animal and yeast mitochondria have independently reassigned AUA from Ile to Met. The bacterial *Mycoplasma* species, which are obligate intracellular parasites, share several features with animal mitochondria, such as small, A+T-rich genomes, and both translate UGA as Trp (reviewed in REF. 9).

Where research has focused on the taxonomy of change, the results are often surprising: the same changes seem to have occurred several times independently in closely related lineages, implying multiple gain and/or loss of a change on a relatively short timescale of tens to hundreds of millions of years. This is true for yeasts[10], cil-

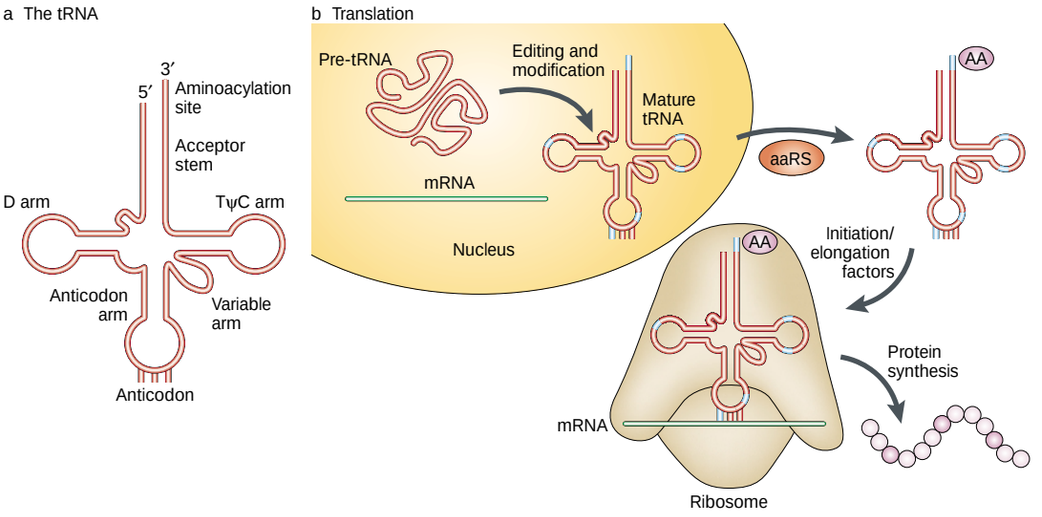


Figure 1 | **Overview of translation. a** | Secondary structure of tRNA, showing major features. The anticodon, which pairs with the codon on the mRNA, is at the opposite end of the molecule from the acceptor stem, to which the amino acid is attached. The D arm and the TψC arm take their names from characteristic base modifications. **b** | tRNA and mRNA are transcribed from genes in the nucleus. Both tRNA and mRNA can be edited before translation; in particular, specific enzymes modify many bases in tRNA, which can change the decoding ability of the tRNA. In the cytoplasm, aminoacyl-tRNA synthetases (aaRSs) specifically charge the tRNAs with amino acids, and, at the ribosome, the three-base anticodon of the tRNA pairs with the complementary three-base codon in the mRNA. The amino acid is added to the growing polypeptide chain, which is extruded through a channel in the ribosome. Mitochondria have their own, separate translation machinery (not shown), but they only encode some of it (primarily tRNAs and rRNAs): they do not encode their own synthetases or editing enzymes, which must be imported from the nucleus.

RELEASE FACTORS
Proteins involved in translation termination that specifically recognize stop codons and catalyse the disassembly of the translation complex.

AMINOACYL-tRNA SYNTHETASE
The enzyme that attaches an amino acid to its cognate tRNA(s).

WOBBLE RULES
An extension to Watson–Crick base pairing, these rules indicate that, in the context of the first anticodon position of the tRNA (complementary to the third codon position), more flexibility allows non-standard base pairs (such as G with U rather than with C).

RIBOZYME
RNA that can perform a catalytic task, such as the self-splicing Group I intron in the ciliate *Tetrahymena*.

THIOL
A thiol (or sulphydryl) group is a chemical group that contains sulphur and hydrogen.

SUPPRESSOR MUTATION
A mutation that counteracts the effects of another mutation. A suppressor mutation maps at a different site to that of the mutation that it counteracts, either within the gene or at a more distant locus. Mutations in tRNAs often act as suppressors, because they can change the meaning of the mutated codon back to the original (albeit usually at a low level, because efficient suppressors are often lethal).

iates[11] and the mitochondria of diatoms[12], algae[13,14] and metazoa[15]. Many more reassignments probably await discovery in taxa that have been less well studied.

Some codons seem to be reassigned frequently, but to various alternatives. For instance, UAG has been reassigned from Stop to Leu, Ala and Gln, and AGA and AGG have been reassigned from Arg to Ser, Gly and Stop. Termination codons may be particularly labile either because they are rare (occurring only once per gene, and therefore causing minimal damage if they are reassigned) or because changes to RELEASE FACTORS are easy to effect[16]. Traditionally, changes have been considered in two separate groups: the (more frequent) changes in mitochondrial codes, and those found in the primary genome. However, all codons that have been lost or reassigned in nuclear lineages have also been lost or reassigned in mitochondrial lineages (FIG. 3); if independent processes were at work in the two systems then the probability of this would be only about 10^{-5} (by Fisher's exact test). This surprising conformity indicates that universal mechanisms, such as the thermodynamics of base pairing, may be at work. For instance, G•C pairs are stronger than A•T pairs, and so the range of possible misreadings or reassignments may differ depending on the composition of particular codons (and of the tRNA anticodons that decode them)[17].

How do changes occur?

Matching codons in mRNA to specific amino acids involves two distinct steps. First, an AMINOACYL-tRNA SYNTHETASE (aaRS) specifically recognizes and covalently links the tRNA and the amino acid (FIG. 1). Then, at the ribosome, the anticodon of the charged tRNA base-pairs with the correct codon on the mRNA, using somewhat extended 'WOBBLE RULES' (TABLE 1). This indirect recognition removes any requirements for direct stereochemical association between trinucleotides and amino acids that may have guided the original codon assignments in an RNA world[1,18]. Other components of the translation apparatus include release factors that specifically recognize termination codons using a tripeptide 'anticodon'[19], and specific enzymes that modify and/or edit the tRNA to alter its recognition at the synthetase[20] and/or the ribosome[21].

Ancient base modifications. Some base modifications, such as U→pseudouridine and A→inosine, are common to all three domains of life — the archaea, the bacteria and the eukaryotes — and are probably at least as old as the last universal common ancestor[22]. The successful selection of RIBOZYMES *in vitro* that can join a base to a sugar to form a nucleoside[23] and the prebiotic synthesis of some modified purines (for example, by reaction of the unmodified purines with methylamine under conditions simulating an evaporating lagoon[24]) may bring some modified bases into the plausible scope of an RNA world. THIOLATED uridines are present throughout the three domains of life (reviewed in REF. 25), and, because they alter decoding at the ribosome by restricting base pairing (TABLE 1), their ubiquitous distribution may indicate that this base modification influenced the code from the beginning.

Mutation of tRNAs. Changes to many components of the translation apparatus can and do alter the genetic code in experimental systems (see supplementary infor-

mation online, Table S1). Many non-standard codon assignments have been traced to changes in tRNAs: unlike other components, mutant tRNAs are easy to characterize because they are small and relatively stable, and because single nucleotide substitutions have direct, specific effects on decoding (reviewed in REF. 26). SUPPRESSOR MUTATIONS are usually altered tRNA genes (reviewed in REF. 27). Other components seem less amenable to change: both the tRNA-binding and amino-acid binding domains of the aaRSs involve many amino-acid residues in precise alignment (reviewed in REF. 28), and mutations would probably alter the translation of multiple codons. Changes in the ribosome, apart from those affecting release-factor binding, are more

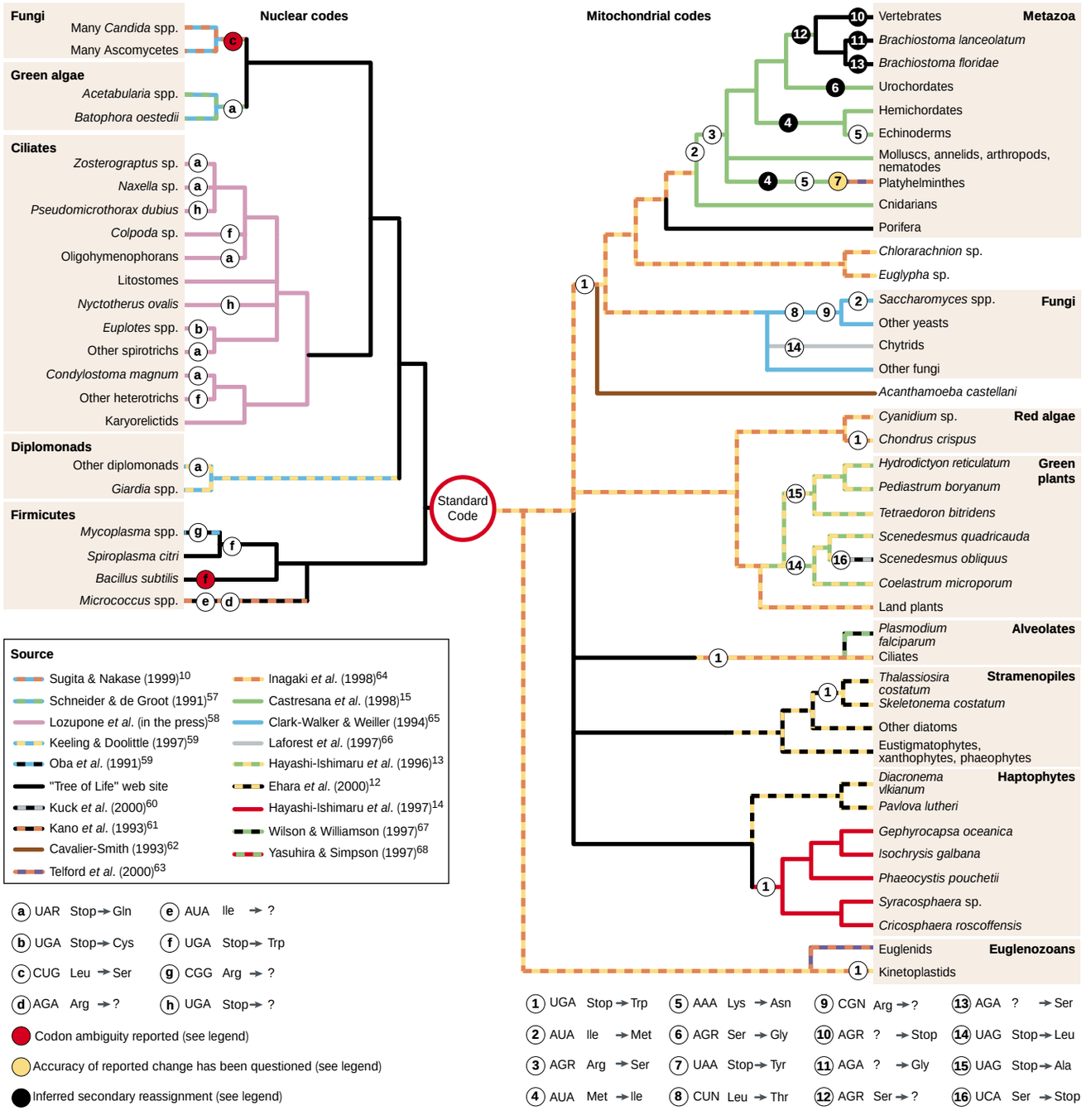


Figure 2 | **Composite phylogeny of variant codes.** Note that the same few changes have taken place repeatedly and independently in different taxa. Relationships are assembled from different studies (see source key) so that branching order, not length, is meaningful. Black discs denote further changes in codons that already deviate from the canonical code; red discs denote instances of codon ambiguity, with specific codons translated either as the canonical amino acid or as the indicated non-standard meaning, depending on the circumstance (for yeast, see REF. 7; for bacteria see REF. 69). The yellow disc denotes a reported reassignment that has recently been challenged on the basis of new sequence data[63]: the reassignment may be limited to one or a few species of *Platyhelminth*. The Tree of Life is a project containing information about the diversity of organisms on Earth and their history (see LINKS). (sp, single unspecified species; spp, multiple unspecified species.)

RNA EDITING
Changes in the RNA sequence after transcription is completed. Examples include modification of C to U or of A to I by deamination, or insertion and/or deletion of particular bases.

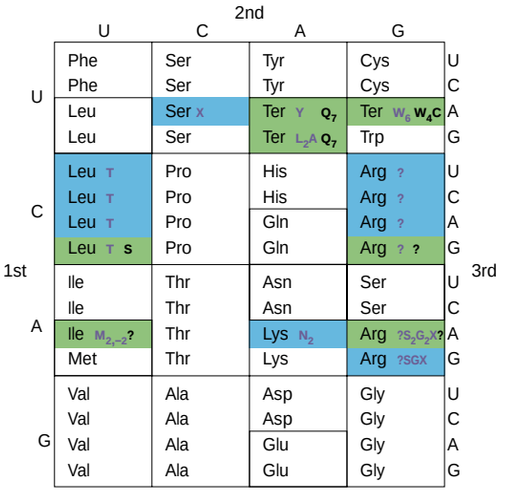


Figure 3 | **The genetic code and its variants.** Blue letters: changes in mitochondrial lineages. Bold letters: changes in nuclear lineages. Blue boxes: codons that have changed only in mitochondria. Green boxes: codons that have changed both in mitochondrial and in nuclear lineages. No codons have changed in nuclear lineages only. (Standard one-letter codes are used for reassigned amino acids; ?, unassigned. Subscripts give number of changes; the minus sign indicates reverse change.)

likely to have a deleterious effect on all codon–anticodon interactions than to affect specific binding. Consequently, some components of the translation apparatus have a greater capacity for adaptive change than others (TABLE 2).

Recent base modifications. Fewer types of change have been observed in natural systems: this may be due to observer bias, or evolution may favour particular mechanisms. However, a surprising number of natural variant codes have been traced to alterations in base modification (see supplementary information online, Table S1). For example, squid and starfish mitochondria translate AGR, where R is any purine, as Ser instead of Arg. The gene for the single tRNA that decodes the AGN block, where N is any nucleotide, has a GCT anticodon, which should only pair with AGY, where Y is any pyrimidine. However, conversion of the first position of the anticodon from G to 7-methylguanosine allows it to decode any nucleotide[29,30].

RNA editing. Editing of C to U, and changes in the targeting of molecules to particular organelles, can also cause changes in the genetic code in mitochondria. The kinetoplastid protist *Leishmania tarentolae*, which encodes all its tRNAs in the nucleus and imports them into the mitochondria, has a single $tRNA^{Trp}$. This tRNA has a CCA anticodon, which decodes only UGG in the cytoplasm. However, mitochondrial RNA EDITING converts the anticodon to UCA, which decodes both UGG and UGA, permitting codon reassignment in the mitochondria but not in the nucleus[31]. In this case, compartmentalization of the editing activity is crucial to avoid altering the nuclear code as well; this implies that changes in organellar targeting of either modification enzymes or edited tRNAs could result in codon reassignment. (No plausible candidate proteins for base modification have been identified in the mitochondrial genome.) Conversely, C to U editing prevents reassignment in marsupial mitochondria, which edit the anticodon of $tRNA^{Asp}$ from GCC to the standard GUC. Changes in targeting of tRNAs and aaRSs, which can be shared among cellular compartments (reviewed in REF. 32), may generally be important because recognition mechanisms can evolve to differ in specific ways[33].

What explains variation in the code?

Francis Crick's seminal observations on genetic-code evolution included speculative, but remarkably prescient, proposed mechanisms for codon reassignment. In 1963, he proposed that biased mutation could render specific codons very rare, permitting their reassignment[34]. Three years later, he suggested a specific example whereby anticodon base modification could induce reassignment of AUA from Met to Ile, through a stage in which the codon is translated ambiguously[35]. However, subsequent findings regarding the apparent universality of the standard genetic code led him to an increasingly strong conviction that codon-reassignment events were limited to primordial evolution when "the genetic message of the cell coded for only a small number of proteins which were somewhat crudely constructed"[1].

More recently, there have been three main attempts to explain variation in the code. The 'codon capture' hypothesis[5,16,36] proposes that fluctuations in mutation bias that influence G+C content can eliminate codons from the entire genome, after which they can be reassigned by neutral processes (that is, without selection) by mutation of other tRNAs. The 'ambiguous intermediate' hypothesis[6,37] notes that tRNA mutations at locations other than at the anticodon can cause translational ambiguity, and ultimately fixation of the new meaning if it is adaptive. The 'genome streamlining' hypothesis[9,38] suggests that code change, at least in mitochondria and obligate intracellular parasites, is driven by selection to minimize the translation apparatus.

These theories, and other suggestions about variation in the code, are not mutually exclusive, especially when we examine their components (TABLE 3). Different codon reassignments may result from different causes, both proximate and ultimate. For instance, codons that have been made rare during extreme G+C pressure might be easier to reassign via an ambiguous intermediate, as the impact of mistranslation at those codons would be ameliorated by their scarcity[6]. Where possible, we test the implications of each theory for codon reassignments that have occurred and those that are possible.

Limits on mutation: chemistry versus history

A novel code must be both chemically plausible and mutationally accessible from its immediate ancestor. The former restriction is liberal, relying primarily on rules of codon–anticodon pairing (although there are many specific mechanisms for change: see below). The

FAMILY BOX
A set of four codons that are identical at the first two positions, differ only at the third position and code for the same amino acid. For instance, GUU, GUC, GUA and GUG comprise a family box for valine in all known codes.

Table 1 | Codon/anticodon base-pairing rules

Crick's wobble rule 1st anticodon	3rd codon	Modified wobble rule 1st anticodon	3rd codon	Usage	Taxa
U	A G	U	U C A G	Family boxes	Mt, ch, *Mycoplasma* spp.
		xo^5U	U A G	Family boxes (Ser, Val, Thr, Ala)	Eubacteria
		xm^5Um, Um, xm^5U	A G	Two-codon sets	Mt, bacteria, eukaryotes
		xm^5s^2U	A (G)	Two-codon sets	Eubacteria, eukaryotes
		Gψ (1st, 2nd)	U C A	Asn AUU, AUC, AUA	Echinoderm mt[70]
C	G	C	G	All	All
		Cm	(A) G	Leu UUR	*E. coli* $tRNA^{5Leu}$ (REF. 71)
		f^5C	A G	Met AUR	Nematode, bovine, squid mt[72]; *Drosophila* mt[46]
		U	A G	Trp UGR	Leishmania mt[31]
		L	A	Ile AUA	Eubacteria, plant mt
A	U	A	U C G (A)	Thr ACU, Arg CGN	*Mycoplasma* spp., yeast mt; nematode mt[73]; artificial *E. coli* $tRNA^{Thr}$ (REF. 74)
		I	U C (A)	Arg CGN	Eubacteria[48]
		I	U C A	Family boxes except Gly GGN	Eukaryotes
G	U C	G	U C	Two-codon sets	All
		G	U C	Family boxes	Eubacteria
		Q	U C	Two-codon sets	Eubacteria, eukaryotes
		m^7G	U C A G	Ser AGN	Echinoderm mt[29], squid mt[30]
		?	U C A	Cys UGU, UGC, UGA	Euplotes[75]

Modified from REFS 76–79 except where noted. Structures can be found in REF. 80. Entries shaded in beige seem to have contributed to changes in the genetic code in some lineages. (Ch, chloroplast; f, formyl; I, inosine; L, lysidine; m, methyl; mt, mitochondria; N, any nucleotide; Q, queosine; R, any purine; s, thiol-substituted; xo^5U, hydroxymethyluridine derivative; xm^5U, methyluridine derivative; Ψ=pseudouridine.)

latter restriction is relatively severe, because a codon reassignment must require only a few mutations in the translation apparatus (as the probability that any change will be other than deleterious is small). Consequently, the existing state of the system influences which variants can be reached.

Codon–anticodon pairing. Recognition rules for the third base of the codon are somewhat expanded from Watson–Crick pairing because of the unusual conformation of the anticodon loop and because of pervasive base modification (TABLE 1). Although U may be specifically recognized by A in certain structural contexts[39], no known base recognizes C uniquely at the third codon position. Consequently, although NNA and NNG can have distinct meanings, NNU and NNC always encode the same amino acid. This wobble-imposed constraint prevents all four codons in a FAMILY BOX from specifying different amino acids. However, the range of possible base modifications indicates that wobble may only be a proximal explanation for modern patterns of degeneracy: if it were advantageous to split NNU and NNC, perhaps some base modification could be found that would do this[40].

Degeneracy in the canonical code seems to follow simple chemical rules. The pattern depends on the G+C content of particular codons: in all known codes, doublets (the first two bases) composed only of G and C form a family box (fourfold degenerate), whereas those composed only of A and U are split between two alternatives. Mixed doublets are split if the second base is a purine, and unsplit otherwise. Because G+C pairs are stronger than A+U pairs, this might indicate a thermodynamic rationale for degeneracy[41]. However, reassignment of CUG from Leu to Ser splits a family box, whereas reassignment of AGY from Ser to Arg creates one, showing that codon reassignments are not constrained by the second observation.

Table 2 | Relative effects of mutations in the translation apparatus

Component mutated	Potential scope of effect	Probability of being neutral or advantageous to organism
mRNA	A single protein	Small
tRNA	All incidences of associated codon(s)	Very small
Aminoacyl-tRNA synthetase	All incidences of associated amino acid	Vanishingly small
Ribosome	All translation	Close to zero

Existing tRNA identities. Because the anticodon is often an identity element for recognition by the aaRS, mutations in anticodons need not alter codon assignments: identities for both aminoacylation and decoding can

Table 3 | Theories explaining variation in the code

	Codon capture[16,36]	Ambiguous intermediate[6,37]	Genome streamlining[9,38,55]	Other suggestions
Sources of variability	tRNA point substitution in anticodon	tRNA mutations at locations other than anticodons that cause ambiguous reading	tRNA deletion	'Anticodon shift' from indels in anticodon loop[27]
	Loss of modification enzyme	Change in release factors	Release factor deletion	Alteration in modification enzyme-recognition sites[27]
	Change in release factor activity		tRNA mutation at anticodon	
	Change in aminoacylation specificity of synthetases[45]			
	Change in ribosome pairing[45]			
Limitations of variability	New tRNAs usually single point substitution from old ones	Altered specificities primarily from G•U mispairing at the first codon position or from C•A mispairing at the third codon position	The last tRNA for an amino acid cannot disappear	G+C content of doublet may make discrimination of third base easier or more difficult, restricting which quartets are split and which are unsplit[17]
	Only some anticodons have modifications to lose	Reassignment might be easier for rare codons, but this is not required	Amino acids with multiple tRNAs should have codons that are frequently reassigned	
	Ambiguity is assumed never to occur: codon must disappear entirely for reassignment			
	Reassigned codons should be rare codons			
Forces promoting fixation	G+C or A+T bias in mutation first causes codons to disappear, then reappear	If alternative reading is favourable in some circumstances, selection minimizes and can eventually eliminate original reading while maximizing new reading	Muller's ratchet and/or faster replication of smaller genomes	Selection for error-minimization better than that found in the canonical code[45]
	Codon loss and original tRNA disappearance occur by drift			
	New tRNA appearance driven by selection to decode reappearing codons under altered bias			

change simultaneously[42]. However, the diversity of suppressor tRNAs indicates that few restrictions on charging exist. For instance, tRNAs that bind UAG termination codons have been derived from all specificities except for $tRNA^{Asn}$ and $tRNA^{Val}$, although some suppressor tRNAs are mischarged with Gln or Lys (reviewed in REF. 43). Similarly, there are missense suppressors derived from single point mutations in the $tRNA^{Gly}$ anticodon (presumably still charged by GlyRS) for 16 codons, which cover a quarter of the code (reviewed in REF. 44). This indicates that synthetase specificities do not greatly restrict codon reassignment.

Wobble pairing allows a single tRNA to decode multiple codons, so changes in certain codons can be correlated: a mutation in a single tRNA might cause multiple reassignments in a two-codon set or family box. Traditionally, it has been assumed that mutations in anticodons lead to most codon reassignments[45]. However, as discussed above, some changes cannot be explained by single point mutations; even where point mutation could achieve a given change, the available data indicate that anticodon base modification predominates. For instance, many animal mitochondria have reassigned AUA from Ile to Met. AUA is normally recognized exclusively by a $tRNA^{Ile}$ with a CAU anticodon in which the C is modified to lysidine to pair with A but not G at the third codon position. This $tRNA^{Ile}$ has vanished, but instead of mutating the anticodon of $tRNA^{Met}$ from CAU to UAU, which would allow it to recognize both AUA and AUG, nematode, squid, bovine and *Drosophila* mitochondria modify the wobble position C of $tRNA^{Met}$ to 5-formylcytidine, which confers the same specificity[46].

Additionally, mutations elsewhere in the tRNA can alter decoding[26,37,47]. In particular, Schultz and Yarus[6] suggest that most codon reassignment takes place through structural tRNA mutants that promote C•A or G•A mispairing at the third codon position or G•U mispairing at the first codon position. However, as noted above, the actual mechanism for reassignment may instead be base modification: because the efficiency of modification can be affected in many ways, this might provide a finely adjustable mechanism for introducing and modifying patterns of coding ambiguity.

Limits on fixation: history versus selection

The subset of variant codes that are actually fixed in modern populations may differ significantly from the set of possible variants. This may be because particular variants are adaptive, or because consistent mutational pressure favours some types of non-adaptive change.

History: fluctuating genome composition. The deleterious effects of codon reassignment can be avoided if codons are absent from all protein-coding genes in the

genome when they are reassigned: although it is nearly impossible[48] to write an English novel without the letter 'e', it would be much easier to write one without the letter 'x'. Species differ vastly in their genome composition, from 74% G+C (*Micrococcus capricolum*) down to only 25% G+C (*Mycoplasma luteus*), which is reflected in extremely biased preferences for synonymous codons. Codons can disappear entirely under directional mutation pressure, allowing their cognate tRNAs to mutate without ill effects. Because A and T are complementary, and C and G are complementary, pressure towards A+T (or G+C) will simultaneously favour disappearance of a codon and its anticodon, because mutation will be in the same direction for both tRNAs and protein-coding mRNA sequences. If the direction of the mutation pressure changes, the codons would reappear, and, if a tRNA with a different charging specificity now recognizes these codons, they will be read as a different amino acid. This codon capture hypothesis[36] provides a strictly neutral[49] model for codon reassignment, which differs from Crick's original proposal primarily in the requirement that the codon must disappear from the genome completely: "Central to codon reassignment is the principle that a codon cannot have two assignments simultaneously, because this would be lethal to an organism. Before reassignment, a codon must disappear from coding sequences in the genes of organisms."[36]

Paradigmatic examples of the codon capture hypothesis are the reassignment of AUA from Ile to Met, AAA from Lys to Asn, and UGA from Stop to Trp. In each case, extreme mutational pressure towards increased genomic G+C content is supposed to have eliminated the A-rich codon entirely, in favour of G-ending codons with equivalent function. The first base of the anticodon of the tRNA is then free to mutate in the same direction, from U to C, restricting pairing from A or G to just G (or, in the case of Ile, the lysidine modification of $tRNA^{Ile}$ could be lost once it is no longer necessary to read AUA). Later, A+T pressure would result in back-mutation, producing the codon again. At this stage, the codon could be captured by another tRNA[36]. For instance, in a split family box, an unassigned NNA codon could be read by either the pyrimidine-reading tRNA (changing its first anticodon base to I to pair with U, C or A), or by the purine-reading tRNA (changing its first anticodon base to U).

If G+C content were the main force driving codon reassignment, we might expect to find a link between the two in cases where the same codon has been reassigned numerous times in related taxa. However, this is not the case in ciliates[11] or diatom mitochondria[12]. The diplomonads provide another counterexample, reassigning UAR from Stop to Gln (in addition to the normal GAR) despite having coding regions relatively rich in G and C (REF. 50).

One potential objection is that G+C content fluctuates rapidly, and so we should not expect to see any association between modern G+C contents and variant genetic codes. However, most changes occur in mitochondria, which are all A+T-rich. If codon reassignment were linked to disappearance under A+T pressure, significantly more changes should take place in codons ending in G and C than in those ending with A and T (because the third codon position is not nearly as functionally constrained as the first and second positions, and responds much faster to mutational bias). This should be especially true when a codon changes by itself, rather than as part of a block of two or four codons in the same lineage (which might occur by a different mechanism). However, we actually observe the opposite: of codon reassignments across all mitochondrial lineages, counting reassignments in different lineages separately, less than one-third (only 8 out of 28) actually involve C- or G-ending codons. Of these, only three (UAG from Stop to Ala, and to Leu twice) involve a G-ending codon changing alone. By contrast, 15 independent reassignments involve an A-ending codon changing by itself (the remaining ten reassignments are block reassignments, such as CUN from Leu to Thr in yeast). So the codon capture model does not seem to explain adequately the pattern of codon reassignments.

Selection: error resistance, adaptive ambiguity or genome minimization? There are three ways in which a change in the code could be adaptive: directly, in that it minimizes the possibility or impact of errors; indirectly, in that individual reassignments are adaptive in some context (such as suppressor mutations); or perversely, in that selection for something else entirely, such as reduction of the translational apparatus, overrides the maladaptive decoding consequences of changes in the code. Various authors have suggested each of these mechanisms.

The universal genetic code seems highly optimal, in that the arrangement of codon assignments minimizes the impact of errors[51]. Although mitochondrial variant codes are close to this optimum, they all seem slightly worse than the canonical code[52]. This does not necessarily mean that the mitochondrial codes are not more suited to their particular translation system (the vastly reduced subset of proteins translated in mitochondria may respond differently to translation error than does the set in a complete, free-living genome), but it is consistent with the idea that the changes are all mildly deleterious variants in terms of minimizing translation error.

In contrast to the strictly neutral codon-disappearance hypothesis[49], selection can drive the process of codon reassignment if there is an intermediate stage in which translation is ambiguous[6]. According to this hypothesis, a mutation in a tRNA alters its decoding efficiency or specificity, causing a single codon to have more than one meaning. If the new meaning is advantageous in some circumstances, selection can favour it over the original meaning, increasing the proportion of one amino acid over another at the ambiguous sites. Eventually, the tRNA that produces the old meaning disappears by mutation or deletion, leaving the new meaning unambiguous. So selection, rather than drift, accelerates every step in the reassignment process[53]. Interestingly, CUG is actually translated ambiguously as both Ser and Leu in some *Candida* species. When the *Candida* tRNA is expressed heterologously in

Table 4 | Testable predictions of the various models

	Codon capture[16,36]	Ambiguous intermediate[6,37]	Genome streamlining[9,38,55]
Predicted code changes not yet observed	Changes accessible by point substitutions, insertions and deletions in the anticodon, but not predicted from the other models. Changes accessible by transitions should be more common if mutation is limiting	G•U 1: UCY Ser→Pro, UCA Ser→Pro, UCG Ser→Pro, UGA Stop→Arg	UUA Leu→Phe, UUG Leu→Phe, UAG Stop→Tyr, AGY Ser→Arg
		C•A 3: Both changes already observed	Reassignments of codons to termination
		G•A 3: *UUA Leu→Phe*, CAA Gln→His, GAA Glu→Asp	
	Changes of G-ending codons other than UAG in mitochondria		
Other predictions	Codons that disappear in some lineages should be more likely to be reassigned in other lineages	tRNAs from species with reassigned codons should show specific changes that cause ambiguous translation when introduced into model organisms	Genetic codes that deviate further from the standard code should encode fewer tRNAs
	Codon frequency should be predictable from genome composition	Mutations that enhance G•A mispairing at the third position in model organisms should be demonstrated	Smaller genomes should encode fewer tRNAs
			Experimental deletion of tRNAs from mitochondria (rescued by import from nucleus) should provide measurable selective advantage

The change in italics (*UAA Leu → Phe*) is consistent with more than one model.

Saccharomyces cerevisiae, it produces misfolded peptides that induce heat-shock proteins, allowing the transformants to survive various environmental insults such as heat, oxidation, heavy metals, cycloheximide and 1.5 M sodium chloride[7]. This provides a direct rationale for considering ambiguous translation as advantageous, rather than deleterious, in some circumstances.

Although the obvious pathway for ambiguous translation is to have two tRNAs that decode the same codon but accept different amino acids, it is also possible for a single tRNA to be charged ambiguously (as in the case of *Candida* mentioned above). However, changes distant from the anticodon at the top of the anticodon helix and in the D arm can cause altered decoding, particularly C•A and G•A mispairing at the third codon position and G•U wobble at the first position[37]. Unlike codon capture, this hypothesis makes specific predictions about which tRNAs can usurp particular codons, limiting the identity of the possible changes. Most observed changes can be explained by point substitutions in the anticodon, which account for 312 out of 800 possible changes in the identity of a codon block. However, of these 312 changes, only 15 are consistent with the three mechanisms identified above. So we can estimate whether more of the observed changes are consistent with codon ambiguity than chance would predict using a simple BINOMIAL TEST.

These mispairing mechanisms cannot account for all of the observed changes (for example, CUN from Leu to Thr). However, of the 15 observed codon reassignments that could potentially have been effected by single-base misreadings or mutations, nine are consistent with Schultz and Yarus's proposed mechanism, with a probability of 5×10^{-9} if such changes are random. This estimate counts each change only once: if we count the number of times each change has occurred (for example, UGA to Trp as ten changes instead of one), this probability drops to about 4×10^{-37}. However, the fact that codons have disappeared from certain mitochondrial genomes for hundreds of millions of years[15] suggests that ambiguous intermediates need not be involved in all cases. Interestingly, although the paper by Castresana and colleagues[15] has been widely cited as support for the codon capture hypothesis, the codon that disappears, AAA, would not have been expected to disappear in an A+T-rich mitochondrial genome (the *Balanoglossus carnosus* mitochondrial genome is 51% A+T).

Finally, it is possible that genetic code evolution is driven by another force entirely, such as genome minimization[9]. This predicts the simplification of the repertoire of tRNAs and modifying enzymes. Family boxes can be recognized by a single tRNA with a U at the first anticodon position; two-codon sets can be most efficiently recognized by a single tRNA with either a G (to recognize U and C), or a modified U (to recognize A and G) at the first anticodon position. Because decoding A with I is inefficient[54], and because the A to I editing step requires a deaminase, three-codon sets should be infrequent because they require an extra tRNA to pair with NNA. Conversely, amino acids that have a single NNG codon should expand to take over NNA by replacing C with modified U at the first anticodon position. Additionally, amino acids, such as Ser, that are encoded both by a four-codon set and a two-codon set should lose the two-codon set to the amino acid with the two adjacent codons, which would eliminate the requirement for another tRNA. Finally, tRNAs can also be lost if the release factors change to recognize the appropriate codons as termination codons[55].

This predicts the following ten missense changes: AUA from Ile to Met; UGA from Stop to Trp; UGA from Stop to Cys; AGA and AGG from Arg to Ser; UAA and UAG from Stop to Tyr; AGY from Ser to Arg; UUA and UUG from Leu to Phe. Also, there are another 37 nonsense changes, of which 13 are accessible by single-base changes. There are therefore 23 possible changes accessible by point mutations. Of 15 codon reassignments by point substitution, eight are consistent with

BINOMIAL TEST
If an observation has only two possible outcomes and there are multiple observations, the binomial distribution gives the probability that *x* or more outcomes of a given type would occur by chance.

genome minimization ($P = 3 \times 10^{-6}$) if each is counted once; counting repeated changes, 18 out of 40 are consistent ($P = 10^{-12}$). So genome minimization may also be important in determining which codons are reassigned. Metazoan mitochondrial genomes have an average size of about 16 kb; elimination of a single tRNA would therefore reduce the genome by nearly 1%, plausibly enough to confer an evolutionarily significant replication advantage.

Conclusions

Inferences about the recent history of the genetic code are constrained by the few variants that are known. Many theories predict the same changes, but for different reasons; for instance, reassignment of AUA between Ile and Met can be explained by loss or gain of base modification, fluctuating G+C pressure that removes the codon from the genome, C•A and G•U mispairing caused by mutations in the D-arm of the tRNA, loss of the tRNA that specifically assigns AUA to Ile or mutation at the anticodon. Therefore, more precise estimates of the relative contributions of each process await discovery of more variants in the future. Some specific, testable predictions of each of the models discussed above are summarized in TABLE 4. However, it is clear that many distinct (but not necessarily mutually exclusive) processes have been involved in producing variant codes.

The observation that well-studied taxa seem to repeat the same codon reassignments indicates that variant codes may be more pervasive than suspected at present, and that either some lineages are predisposed to certain changes or that certain changes provide a selective advantage in particular ecological circumstances. Further study of tRNA molecules and of release factors in sister taxa with different genetic codes should answer questions about mutations that predispose a group to codon reassignment, especially with respect to the role of ambiguous translation. However, ecological factors that affect code change have received relatively little attention.

Given our knowledge of variant codes, several unseen changes seem to be likely candidates for reassignment. In split family boxes, the NNA codon has acquired a new specificity in five out of eight cases. So it would not be surprising to find lineages that have reassigned UUA to Phe, GAA to Asp or CAA to His. Furthermore, where an amino acid is encoded by two disjoint blocks of codons, each of those blocks is susceptible to reassignment (the change is not necessarily deleterious, as would be the case if all codons for an amino acid were reassigned, because some codons with the original specificity remain). Thus, the UUR Leu block, and the UCN and AGY Ser blocks, might be reassigned similarly to the CUN Leu block and the AGR Arg block.

These two processes — reassignment between two amino acids that share a family box and wholesale replacement of one two-codon or four-codon set by another specificity — together allow any two amino acids to interchange their positions in the code table. However, certain swaps require fewer steps than others, and should therefore occur more frequently. This capacity for rearrangement could permit optimization of a primitive code to a highly adapted state, especially in a less intricate early translation system. However, the observation that several amino acids show an intrinsic affinity for their cognate codons in the canonical code (reviewed in REF. 56) may place limits on the actual impact of such rearrangements. Extrapolation of these processes back from the canonical genetic code may indicate how selection and chemistry shaped the code before the last universal common ancestor.

Overall, we may conclude that the code is far from frozen, and is still evolving in many lineages. The scope and extent of variation increases as new sequence data accumulate, which underscores the importance of related work in understanding how and why the standard code evolved in the way it did (reviewed in REF. 2). Furthermore, it provides the basis for asking important questions about the link between code structure and the process of molecular evolution. Increasingly, comparative genomics is moving beyond the analysis of individual gene sequences and towards the analysis of assemblages of genes and the common evolutionary mechanisms that govern their alteration and rearrangement. As more and more non-standard codes are discovered, and the mechanisms that underlie codon reassignments are further clarified, it becomes possible to explore these subtle relationships between coding rules and genome evolution.

Links

FURTHER INFORMATION The Tree of Life | The RNA world web site | Laura Landweber's lab page | RNA editing

ENCYCLOPEDIA OF LIFE SCIENCES Transfer RNA | RNA editing | Genetic code and its variants | tRNA modification

1. Crick, F. H. C. The origin of the genetic code. *J. Mol. Biol.* **38**, 367–379 (1968).
Seminal introduction to the origin and evolution of the genetic code, best known for its exposition of the 'frozen accident' theory (that the code became fixed at a suboptimal state, because to change it would be deleterious).
2. Knight, R. D., Freeland, S. J. & Landweber, L. F. Selection, history and chemistry: the three faces of the genetic code. *Trends Biochem. Sci.* **24**, 241–247 (1999).
3. Szathmáry, E. The origin of the genetic code: amino acids as cofactors in an RNA world. *Trends Genet.* **15**, 223–229 (1999).
4. Barrell, B. G., Bankier, A. T. & Drouin, J. A different genetic code in human mitochondria. *Nature* **282**, 189–194 (1979).
5. Osawa, S. *Evolution of the Genetic Code* (Oxford Univ. Press, Oxford, 1995).
Exposition of the 'codon capture' hypothesis, which proposes a neutral mechanism for codon reassignment through a stage in which the codon disappears from the genome entirely.
6. Schultz, D. W. & Yarus, M. On malleability in the genetic code. *J. Mol. Evol.* **42**, 597–601 (1996).
Exposition of the 'ambiguous intermediate' hypothesis, which suggests that the genetic code changes through a state in which some codons have more than one meaning.
7. Santos, M. A., Cheesman, C., Costa, V., Moradas-Ferreira, P. & Tuite, M. F. Selective advantages created by codon ambiguity allowed for the evolution of an alternative genetic code in Candida spp. *Mol. Microbiol.* **31**, 937–947 (1999).
Provides experimental support for the idea that ambiguous decoding can be advantageous in some circumstances.

8. Wagner, G. P. & Altenberg, L. Complex adaptations and the evolution of evolvability. *Evolution* **50**, 967–976 (1996).
9. Andersson, S. G. & Kurland, C. G. Genomic evolution drives the evolution of the translation system. *Biochem. Cell Biol.* **73**, 775–787 (1995).
Most complete exposition of the 'genome reduction' hypothesis, which suggests that pressure to minimize mitochondrial genomes leads to the reassignment of specific codons.
10. Sugita, T. & Nakase, T. Non-universal usage of the leucine CUG codon and the molecular phylogeny of the genus Candida. *Syst. Appl. Microbiol.* **22**, 79–86 (1999).
11. Tourancheau, A. B., Tsao, N., Klobutcher, L. A., Pearlman, R. E. & Adoutte, A. Genetic code deviations in the ciliates: evidence for multiple and independent events. *EMBO J.* **14**, 3262–3267 (1995).
12. Ehara, M., Inagaki, Y., Watanabe, K. I. & Ohama, T. Phylogenetic analysis of diatom *cox*I genes and implications of a fluctuating GC content on mitochondrial genetic code evolution. *Curr. Genet.* **37**, 29–33 (2000).
13. Hayashi-Ishimaru, Y., Ohama, T., Kawatsu, Y., Nakamura, K. & Osawa, S. UAG is a sense codon in several chlorophycean mitochondria. *Curr. Genet.* **30**, 29–33 (1996).
14. Hayashi-Ishimaru, Y., Ehara, M., Inagaki, Y. & Ohama, T. A deviant mitochondrial genetic code in prymnesiophytes (yellow-algae): UGA codon for tryptophan. *Curr. Genet.* **32**, 296–299 (1997).
15. Castresana, J., Feldmaier-Fuchs, G. & Pääbo, S. Codon reassignment and amino acid composition in hemichordate mitochondria. *Proc. Natl Acad. Sci. USA* **95**, 3703–3707 (1998).
16. Osawa, S., Jukes, T. H., Watanabe, K. & Muto, A. Recent evidence for evolution of the genetic code. *Microbiol. Rev.* **56**, 229–264 (1992).
17. Lagerkvist, U. Unorthodox codon reading and the evolution of the genetic code. *Cell* **23**, 305–306 (1981).
18. Knight, R. D. & Landweber, L. F. Guilt by association: the arginine case revisited. *RNA* **6**, 499–510 (2000).
19. Ito, K., Uno, M. & Nakamura, Y. A tripeptide 'anticodon' deciphers stop codons in messenger RNA. *Nature* **403**, 680–684 (2000).
Experimental demonstration that bacterial release factors use only a few amino acids to recognize the specific mRNA stop codons.
20. Perret, V. *et al.* Relaxation of a transfer RNA specificity by removal of modified nucleotides. *Nature* **344**, 787–789 (1990).
21. Muramatsu, T. *et al.* Codon and amino-acid specificities of a transfer RNA are both converted by a single post-transcriptional modification. *Nature* **336**, 179–181 (1988).
22. Cermakian, N. & Cedegren, R. C. in *Modification and Editing of RNA* (eds Grosjean, H. & Benne, R.) 535–541 (American Society for Microbiology, Washington, 1998).
Reviews the distribution of modified bases throughout the three domains of life, and argues that many of the modifications pre-date the last common ancestor of extant life.
23. Unrau, P. J. & Bartel, D. P. RNA-catalysed nucleotide synthesis. *Nature* **395**, 260–263 (1998).
24. Levy, M. & Miller, S. L. The prebiotic synthesis of modified purines and their potential role in the RNA world. *J. Mol. Evol.* **48**, 631–637 (1999).
25. Edmonds, C. G. *et al.* Posttranscriptional modification of tRNA in thermophilic archaea (Archaebacteria). *J. Bacteriol.* **173**, 3138–3148 (1991).
26. Giege, R., Sissler, M. & Florentz, C. Universal rules and idiosyncratic features in tRNA identity. *Nucleic Acids Res.* **26**, 5017–5035 (1998).
Excellent review of tRNA identity.
27. Murgola, E. J. tRNA, suppression, and the code. *Annu. Rev. Genet.* **19**, 57–80 (1985).
28. Arnez, J. G. & Moras, D. Structural and functional considerations of the aminoacylation reaction. *Trends Biochem. Sci.* **22**, 211–216 (1997).
29. Matsuyama, S., Ueda, T., Crain, P. F., McCloskey, J. A. & Watanabe, K. A novel wobble rule found in starfish mitochondria. Presence of 7-methylguanosine at the anticodon wobble position expands decoding capability of tRNA. *J. Biol. Chem.* **273**, 3363–3368 (1998).
This is one of a series of papers from Watanabe's lab, and shows the role of specific base modifications in changing the genetic code in mitochondria.
30. Tomita, K., Ueda, T. & Watanabe, K. 7-Methylguanosine at the anticodon wobble position of squid mitochondrial tRNA(Ser)GCU: molecular basis for assignment of AGA/AGG codons as serine in invertebrate mitochondria. *Biochim. Biophys. Acta* **1399**, 78–82 (1998).
31. Alfonzo, J. D., Blanc, V., Estevez, A. M., Rubio, M. A. & Simpson, L. C to U editing of the anticodon of imported mitochondrial tRNA(Trp) allows decoding of the UGA stop codon in *Leishmania tarentolae*. *EMBO J.* **18**, 7056–7062 (1999).
32. Small, I., Wintz, H., Akashi, K. & Mireau, H. Two birds with one stone: genes that encode products targeted to two or more compartments. *Plant Mol. Biol.* **38**, 265–277 (1998).
33. Mazauric, M. H., Roy, H. & Kern, D. tRNA glycylation system from *Thermus thermophilus*. $tRNA^{Gly}$ identity and functional interrelation with the glycylation systems from other phylae. *Biochemistry* **38**, 13094–13105 (1999).
34. Crick, F. H. C. The recent excitement in the coding problem. *Prog. Nucleic Acids* **1**, 163–217 (1963).
35. Crick, F. H. Codon–anticodon pairing: the wobble hypothesis. *J. Mol. Biol.* **19**, 548–555 (1966).
36. Osawa, S. & Jukes, T. H. Codon reassignment (codon capture) in evolution. *J. Mol. Evol.* **28**, 271–278 (1989).
37. Schultz, D. W. & Yarus, M. Transfer RNA mutation and the malleability of the genetic code. *J. Mol. Biol.* **235**, 1377–1380 (1994).
38. Andersson, S. G. & Kurland, C. G. Reductive evolution of resident genomes. *Trends Microbiol.* **6**, 263–268 (1998).
39. Takai, K., Takaku, H. & Yokoyama, S. *In vitro* codon-reading specificities of unmodified tRNA molecules with different anticodons on the sequence background of *Escherichia coli* $tRNA^{Ser}$. *Biochem. Biophys. Res. Commun.* **257**, 662–667 (1999).
40. Szathmáry, E. Codon swapping as a possible evolutionary mechanism. *J. Mol. Evol.* **32**, 178–182 (1991).
41. Lagerkvist, U. 'Two out of three': An alternative method for codon reading. *Proc. Natl Acad. Sci. USA* **75**, 1759–1762 (1978).
42. Saks, M. E., Sampson, J. R. & Abelson, J. Evolution of a transfer RNA gene through a point mutation in the anticodon. *Science* **279**, 1665–1670 (1998).
Demonstration that a single-base change at the anticodon of a tRNA can change both its decoding and aminoacylation specificities. This paper has important implications for the use of tRNA phylogeny to track the evolution of the genetic code.
43. Pallanck, K., Pak, M. & Schulman, L. H. in *tRNA: Structure, Biosynthesis, and Function* (eds Söll, D. & RajBhandary, U.) 371–394 (American Society for Microbiology, Washington, 1995).
44. Murgola, E. J. in *tRNA: Structure, Biosynthesis and Function* (eds Söll, D. & RajBhandary, U.) 491–509 (American Society for Microbiology, Washington, 1995).
45. Jukes, T. H. Genetic code 1990. Outlook. *Experientia* **46**, 1149–1157 (1990).
46. Tomita, K. *et al.* Codon reading patterns in *Drosophila melanogaster* mitochondria based on their tRNA sequences: a unique wobble rule in animal mitochondria. *Nucleic Acids Res.* **27**, 4291–4297 (1999).
47. Schimmel, P., Giege, R., Moras, D. & Yokoyama, S. An operational genetic code for amino acids and possible relationship to genetic code. *Proc. Natl Acad. Sci. USA* **90**, 8763–8768 (1993).
48. Wright, E. V. *Gadsby: a story of over 50,000 words without using the letter 'E'* (Wetzel, Los Angeles, 1939).
49. Jukes, T. H. Neutral changes and modifications of the genetic code. *Theor. Popul. Biol.* **49**, 143–145 (1996).
50. Keeling, P. J. & Doolittle, W. F. Widespread and ancient distribution of a noncanonical genetic Code in diplomonads. *Mol. Biol. Evol.* **14**, 895–901 (1997).
51. Freeland, S. J. & Hurst, L. D. The genetic code is one in a million. *J. Mol. Evol.* **47**, 238–248 (1998).
A statistical argument to show that the actual genetic code minimizes the effects of error far better than would be expected by chance.
52. Freeland, S. J., Knight, R. D., Landweber, L. F. & Hurst, L. D. Early fixation of an optimal genetic code. *Mol. Biol. Evol.* **17**, 511–518 (2000).
53. Yarus, M. & Schultz, D. W. Response: Further comments on codon reassignment. *J. Mol. Evol.* **45**, 1–8 (1997).
54. Curran, J. F. Decoding with the A:I wobble pair is inefficient. *Nucleic Acids Res.* **23**, 683–688 (1995).
55. Andersson, G. E. & Kurland, C. G. An extreme codon preference strategy: codon reassignment. *Mol. Biol. Evol.* **8**, 530–544 (1991).
56. Yarus, M. RNA-ligand chemistry: a testable source for the genetic code. *RNA* **6**, 475–484 (2000).
57. Schneider, S. U. & de Groot, E. J. Sequences of two rbcS cDNA clones of *Batophora oerstedii*: structural and evolutionary considerations. *Curr. Genet.* **20**, 173–175 (1991).
58. Lozupone, C. A., Knight, R. D. & Landweber, L. F. The molecular basis of nuclear genetic code change in ciliates. *Curr. Biol.* (in the press).
59. Oba, T., Andachi, Y., Muto, A. & Osawa, S. CGG: an unassigned or nonsense codon in *Mycoplasma capricolum*. *Proc. Natl Acad. Sci. USA* **88**, 921–925 (1991).
60. Kuck, U., Jekosch, K. & Holzamer, P. DNA sequence analysis of the complete mitochondrial genome of the green alga *Scenedesmus obliquus*: evidence for UAG being a leucine and UCA being a non-sense codon. *Gene* **253**, 13–18 (2000).
61. Kano, A., Ohama, T., Abe, R. & Osawa, S. Unassigned or nonsense codons in *Micrococcus luteus*. *J. Mol. Biol.* **230**, 51–56 (1993).
62. Cavalier-Smith, T. Kingdom protozoa and its 18 phyla. *Microbiol. Rev.* **57**, 953–994 (1993).
63. Telford, M. J., Herniou, E. A., Russell, R. B. & Littlewood, D. T. Changes in mitochondrial genetic codes as phylogenetic characters: two examples from the flatworms. *Proc. Natl Acad. Sci. USA* **97**, 11359–11364 (2000).
64. Inagaki, Y., Ehara, M., Watanabe, K. I., Hayashi-Ishimaru, Y. & Ohama, T. Directionally evolving genetic code: the UGA codon from stop to tryptophan in mitochondria. *J. Mol. Evol.* **47**, 378–384 (1998).
65. Clark-Walker, G. D. & Weiller, G. F. The structure of the small mitochondrial DNA of *Kluyveromyces thermotolerans* is likely to reflect the ancestral gene order in fungi. *J. Mol. Evol.* **38**, 593–601 (1994).
66. Laforest, M. J., Roewer, I. & Lang, B. F. Mitochondrial tRNAs in the lower fungus *Spizellomyces punctatus*: tRNA editing and UAG 'stop' codons recognized as leucine. *Nucleic Acids Res.* **25**, 626–632 (1997).
67. Wilson, R. J. & Williamson, D. H. Extrachromosomal DNA in the Apicomplexa. *Microbiol. Mol. Biol. Rev.* **61**, 1–16 (1997).
68. Yasuhira, S. & Simpson, L. Phylogenetic affinity of mitochondria of *Euglena gracilis* and kinetoplastids using cytochrome oxidase I and hsp60. *J. Mol. Evol.* **44**, 341–347 (1997).
69. Lovett, P. S. *et al.* UGA can be decoded as tryptophan at low efficiency in *Bacillus subtilis*. *J. Bacteriol.* **173**, 1810–1812 (1991).
70. Tomita, K., Ueda, T. & Watanabe, K. The presence of pseudouridine in the anticodon alters the genetic code: a possible mechanism for assignment of the AAA lysine codon as asparagine in echinoderm mitochondria. *Nucleic Acids Res.* **27**, 1683–1689 (1999).
71. Horie, N. *et al.* Modified nucleosides in the first positions of the anticodons of tRNA(Leu)4 and tRNA(Leu)5 from *Escherichia coli*. *Biochemistry* **38**, 207–217 (1999).
72. Tomita, K., Ueda, T. & Watanabe, K. 5-formylcytidine (f5C) found at the wobble position of the anticodon of squid mitochondrial tRNA(Met)CAU. *Nucleic Acids Symp. Ser.* **37**, 197–198 (1997).
73. Watanabe, Y. *et al.* Primary sequence of mitochondrial tRNA(Arg) of a nematode *Ascaris suum*: occurrence of unmodified adenosine at the first position of the anticodon. *Biochim. Biophys. Acta* **1350**, 119–122 (1997).
74. Boren, T. *et al.* Undiscriminating codon reading with adenosine in the wobble position. *J. Mol. Biol.* **230**, 739–749 (1993).
75. Grimm, M., Brunen-Nieweler, C., Junker, V., Heckmann, K. & Beier, H. The hypotrichous ciliate *Euplotes octocarinatus* has only one type of $tRNA^{Cys}$ with GCA anticodon encoded on a single macronuclear DNA molecule. *Nucleic Acids Res.* **26**, 4557–4565 (1998).
76. Watanabe, K. & Osawa, S. in *tRNA: Structure, Biosynthesis, and Function* (eds Söll, D. & RajBhandary, U.) 225–250 (American Society for Microbiology, Washington, 1995).
77. Yokoyama, S. & Nishimura, S. in *tRNA: Structure, Biosynthesis, and Function* (eds. Söll, D. & RajBhandary, U.) 207–223 (American Society for Microbiology, Washington 1995).
78. Björk, G. R. in *Modification and Editing of RNA* (eds Grosjean, H. & Benne, R.) 577–581 (American Society for Microbiology, Washington, 1998).
79. Curran, J. F. in *Modification and Editing of RNA* (eds Grosjean, H. & Benne, R.) 493–516 (American Society for Microbiology, Washington, 1998).
Excellent review of the base-pairing roles of normal and modified bases at the wobble position in the tRNA anticodon.
80. Motorin, Y. & Grosjean, H. in *Modification and Editing of RNA* (eds Grosjean, H. & Benne, R.) 543–549 (American Society for Microbiology, Washington, 1998).

Acknowledgements

S.J.F. is supported by a Human Frontier Science Programm fellowship.

3.2 How Mitochondria Redefine the Code

Most variant genetic codes are in mitochondria, which have the nice property that their genomes are small and there are many complete sequences in GenBank. I was surprised that no-one had used this obvious resource to test the various hypotheses about the evolution of variant genetic codes: each made specific predictions that could easily be tested with reference to the whole genome. Specifically, the Codon Capture model implied (a) that codons that disappear [in some species] should be the same codons that respond strongly to mutation bias and are reassigned [in other species], and (b) that the codons that respond strongly to mutation bias should be the same ones that disappear; the Genome Minimization model implied that smaller genomes should have fewer tRNAs and more changes in the code.

Surprisingly, although mitochondria can be under intense selection for minimal genome size, and can even shorten the length of each *tRNA by 20 bases on average, tRNA loss is not a strategy they use for genome reduction! I also found that, in mitochondria, the frequencies of the four bases in coding regions are uncorrelated (except for tradeoffs due to deamination), in contrast to every nuclear genome.*

This chapter is forthcoming in the Journal of Molecular Evolution:

Knight, R. D., L. F. Landweber, and M. Yarus (2001). "How mitochondria redefine the code." J Mol Evol, forthcoming 2001.

Prof. Yarus contributed several important insights in reconciling the data with both the Codon Ambiguity hypothesis and certain aspects of Codon Capture.

3.2.1 Abstract

Annotated, complete DNA sequences are available for 213 mitochondrial genomes from 132 species. These provide an extensive sample of evolutionary adjustment of codon usage and meaning spanning the history of this organelle. Because most known coding changes are mitochondrial, such data bear on the general mechanism of codon reassignment. Coding changes have variously been attributed to loss of codons due to changes in directional mutation affecting the genome GC content (Osawa and Jukes 1988), to pressure to reduce the number of mitochondrial tRNAs in order to minimize genome size (Andersson and Kurland 1991) and to the existence of transitional coding mechanisms in which translation is ambiguous (Schultz and Yarus 1994). We find that no single proposed mechanism completely explains observed codon reassignments.

In particular: (1) Genomic variation in the prevalence of a codon's 3rd position nucleotide predicts relative mitochondrial codon usage well, though GC content does not. This is because A and T, and G and C, are uncorrelated in mitochondrial genomes. (2) Codons predicted to reach zero usage (disappear) do so more often than expected by chance, and codons that do disappear are disproportionately likely to be reassigned. However, codons predicted to disappear are not significantly more likely to be reassigned. Indefinitely low codon frequency is apparently not sufficient for disappearance, and disappearance is not sufficient for reassignment. (3) Changes in the genetic code are not more likely to accompany smaller numbers of tRNA genes, and are not more frequent in smaller genomes.

Thus, mitochondrial codons are not reassigned during demonstrable selection for decreased genome size. Instead, the data suggest that both codon disappearance and codon reassignment depend on at least one other event. This mitochondrial event (leading to reassignment) occurs more frequently when a codon has disappeared, and produces only a small subset of possible reassignments. We suggest that coding ambiguity, the extension of a tRNA's decoding capacity beyond its original set of codons, is the second event. Ambiguity can act alone, but often acts in concert with codon disappearance, which broadly promotes codon reassignment.

3.2.2 Introduction

Although most nuclear genomes share a single genetic code inherited from their common ancestor, codon reassignments are pervasive in mitochondria. Current evidence suggests as many as 27 codon reassignments in mitochondria, although several specific changes (e.g. UGA Ter → Trp) seem to have recurred frequently (Knight, Freeland and Landweber, in press *Nature Genetics Reviews*). It is hard to understand how these changes take place. Mitochondria (and especially animal mitochondria) encode a small set of highly-conserved genes that are critical for respiration, and so any change in the genetic code would change most proteins and be expected to be deleterious.

However, mitochondrial genomes may be predisposed to change by their differences from nuclear genomes. First, they are much smaller: *C. elegans* has fewer than 14 000 base pairs in its mtDNA. Second, they encode few proteins, importing most of what they need from the nucleus, and may thus not be under strong selection for translational accuracy (especially since the proteins they do encode are essential for respiration, not their own replication). Third, they are extremely biased in nucleotide composition, and are all AT-rich.

Some of these differences should explain why we see more code changes in mitochondria than in nuclear genomes. However, there are grounds for caution: changes that have taken place in nuclear genomes are a subset of the changes that have taken place in mitochondrial genomes. No codon has changed in a nuclear lineage that has not changed in a mitochondrial lineage. In fact, the probability that the sets of codons known to have changed in mitochondria and nuclei are independent is only 3.3×10^{-4} (by Fisher's Exact Test – see Table 1). Thus, common processes may be at work in both cases: the peculiarities of mitochondria may alter the rate at which changes occur, but not the types of changes that occur.

We summarize the hypotheses about changes in the genetic code, each based on one or more of the three major peculiarities of mitochondrial genomes, as follows:

1. The 'codon capture' or 'codon disappearance' hypothesis (Jukes, Osawa et al. 1987; Osawa, Jukes et al. 1987; Osawa and Jukes 1988; Osawa, Ohama et al. 1988; Osawa and Jukes 1989; Osawa, Jukes et al. 1992; Osawa 1995) proposes that specific codons, made rare by AT or GC pressure, disappear from the genome entirely. Any mutation to the tRNAs that translate these codons will be allowed, since such mutations will be neutral. If the mutation pressure reverses, causing these codons to reappear, they may now code for a different amino acid. Thus the process of codon reassignment can be 'entirely neutral' (Jukes 1996). This argument accounts for the frequency of reassignments in mitochondria on the grounds that mitochondria are (a) AT-rich (so GC-rich codons should disappear frequently), and (b) small (so a codon really could disappear entirely by directional mutation pressure). It is also supported by the fact that the small, AT-rich Mycoplasmas have the same UGA → Trp reassignment as do mitochondria (Yamao, Muto et al. 1985).
2. The 'genome minimization' hypothesis (Andersson and Kurland 1990; Andersson and Kurland 1991; Andersson and Kurland 1995; Andersson and Kurland 1998) proposes that mitochondria are under extreme selection to reduce their genome size, which should confer an advantage in replication. tRNAs are usually about 74 bases long, so a single full-length tRNA corresponds to nearly 0.5% of the smallest genomes. If replication is proportional to genome length, an advantage of 0.5% per generation may be very significant in evolutionary terms, and should lead to fixation of genomes lacking particular tRNAs (if this is not otherwise too deleterious). Consequently, changes in the code that allow the elimination of tRNAs or release factors, such as UGA to Trp, should be favored. This argument relies only the small size of mitochondrial genomes. Like the codon capture hypothesis, it is supported by the fact that the small *Mycoplasma* genomes use UGA for Trp, and lack the release factor

RF2 that would normally cause termination at this codon (Inagaki, Bessho et al. 1993; Inagaki, Bessho et al. 1996).

3. The 'ambiguous intermediate' hypothesis (Schultz and Yarus 1994; Schultz and Yarus 1996; Yarus and Schultz 1997) proposes that, rather than disappearing entirely, codons undergo a period of ambiguous translation in which a single codon is read in two ways. This hypothesis draws support from the fact that tRNAs are observed to misread in specific ways (G-A and C-A pairing at the third position, and G-U at the first position) that parallel changes most commonly observed in variant codes (whether nuclear or mitochondrial) (Schultz and Yarus 1994; Schultz and Yarus 1994). It is also supported by the fact that ambiguous translation between Ser and Leu at the codon CUG does occur in some yeast (Santos and Tuite 1995; Santos, Perreau et al. 1996; Santos, Cheesman et al. 1999) that have reassigned CUG from Leu to Ser. Additionally, *Bacillus subtilis* translates UGA ambiguously, sometimes inserting Trp (Lovett, Ambulos et al. 1991; Matsugi, Murao et al. 1998). So far, the ambiguous intermediate hypothesis has not tried to explain why changes are more frequent in mitochondrial lineages, but the extensive alteration of mitochondrial tRNAs (some of which lack the entire T or D arm (Wolstenholme, Macfarlane et al. 1987; Watanabe, Tsurui et al. 1994)) could provide opportunity for structural changes that cause ambiguity.

A difficulty in comparing these hypotheses is that the same evidence tends to support more than one of them. For instance, the properties of the mycoplasmas are taken to support both codon capture and genome reduction as mechanisms for code change, since their genomes happen to be both small and AT-rich. Here we evaluate evidence for and against the three hypotheses as obtained from a large sample of mitochondrial genomes: 213 completely sequenced and annotated genomes from 132 different species. In particular we ask:

1. Do the codons that should disappear from AT-rich mitochondrial genomes really disappear?
2. Is there a link between codon disappearance and codon reassignment?
3. Do smaller genomes have fewer tRNAs, or more changes in the code?
4. Are the patterns of codon reassignment consistent with change through ambiguous intermediates?

3.2.3 Materials and Methods

We downloaded all mitochondrial genomes available through NCBI (http:www.ncbi.nlm.nih.gov) on 8/30/2000. This set consisted of 218 records from a variety of eukaryotes, but not including kinetoplastids (which have highly unusual mitochondria). Of these 218 records, two were excluded because they were for plasmids (*Neurospora crassa),* and three were excluded because only one gene was annotated (one each for *Homo sapiens, Pan troglodytes, and Gorilla gorilla)*. Additionally, the record for *Physarum polycephalum* had no structural RNAs annotated (although all protein-coding sequences were annotated): we exclude this genome from analyses of tRNA and rRNA length and composition.

All annotated features of the genomes (rRNA, tRNA, CDS, and intron) were extracted and analyzed using a custom program written in C. Individual exons were not considered: where introns occurred, the bases involved in the entire CDS record were linked together into a single string and then analyzed. We defined "spacer" regions as those nucleotides not belonging to any feature (including features such as the origin of replication, variations, etc., which we extracted but did not analyze). Since the 'spacer' includes any features that are present but not annotated, we do not consider it a reliable indication of the properties of nucleotides not under selection.

Base compositions and codon usages of the different types of feature were summed for each genome, and further analyzed using Microsoft Excel. Amino acid frequencies were estimated using the feature's in-frame codon usage (determined from the primary sequence) and the translation table given in the genome record. Comparison with the amino acid sequences read directly from the CDS record indicates that RNA editing does not greatly affect the overall amino acid usage in the genomes under consideration here.

Where multiple genomes were available for a single species, we averaged the codon usages across all available genomes. Thus each species is given equal weighting in subsequent analyses: we did not account further for the effects of phylogeny, although this is unlikely to affect the conclusions because distantly related taxa tend to fall on the same regression lines.

Statistical tests for independence were calculated using the G test with Williams' correction (Sokal and Rohlf 1995), implemented in Microsoft Excel. Binomial probabilities were calculated using Excel's built-in BINOMDIST function for individual terms and then summing them. We note that the cumulative BINOMDIST function can give wildly inaccurate results for low probabilities. Consequently, we checked the results using an independent method (by evaluating the area under the curve directly using the incomplete beta function) as described and implemented in Press et al. (1992).

The null hypothesis for patterns of codon reassignment was determined as follows. NNU and NNC must always code for the same amino acid, since no first-position base in the anticodon discriminates efficiently between these nucleotides. Thus there are 48 blocks of codons that could, in principle, change independently. Of these, 8 blocks (e.g. UGG Trp) normally code for a single amino acid, and so cannot be reassigned. Each of the 40 remaining blocks could, in principle, be reassigned to 20 alternative meanings, giving 800 possible reassignments. Most reassignments are single-base changes in meaning: if we consider only these changes, the structure of the canonical code allows 312 nonsynonymous changes. Of these, only fifteen changes are allowed by the three types of mispairing invoked by the Schultz/Yarus model. Thus we can compare the actual changes in the code to a binomial distribution where the probability that each observed change is consistent with the model is 15/800 if all changes are considered, or 15/312 if we consider only point substitutions, and ask whether more agree with the model than would be expected by chance.

3.2.4 GC Content and Codon Reassignment

Forces common to the whole genome influence the nucleotide content of coding sequences and other functional molecules. Consequently, the GC content of the 1st, 2nd, and 3rd position of codons, of tRNAs, of rRNAs, and of spacer elements are highly correlated (Muto and Osawa 1987). This is consistent with the idea that the bulk of change at the molecular level is driven by mutation, as limited by the functional consequences of mutation (Sueoka 1962; Kimura 1968; King and Jukes 1969). Figure 1 shows the relationship between these variables for mitochondria: as in nuclear genomes, the GC content of the different elements is highly correlated. In fact, the GC content accounts for 98% of the variance in GC content in coding sequences, and 84% of the variance in GC content in rRNA genes.

As a result, it is meaningful to consider the mutation spectrum as a single force that affects all parts of the genome, subject to the selective constraints at each position. GC content varies widely in a variety of lineages: in this data set, for instance, the chlorophytes vary from 22% to 36% GC, and the crustaceans vary from 15% to 38% GC. Different lineages have explored more or less the same range of GC contents (though always less than 50% GC). Therefore we can evaluate the effect of GC content on codon disappearance and codon reassignment. Suppose that codon disappearance is a necessary prerequisite to codon reassignment, and the same forces affect codon usage in all genomes. Then the set of reassigned codons should be a subset of the set of codons that disappeared. Moreover, since AT pressure is acting in all mitochondrial genomes, (a) the codons that disappear should be GC-rich codons, (b) the codons that are reassigned (in some lineages) should also be found to disappear (in other

lineages), and, by inference, (c) the codons that are reassigned should be the ones that would be predicted to disappear based on GC content. Since the third-position base is under less selection (since changes are usually synonymous), it changes faster than the other two bases in relation to genome GC content. Thus, the codons that have been reassigned in mitochondria would likely contain G or C at the third position, and should not contain A or U.

A brief examination of Table 1 should be sufficient to convince even the most ardent proponent of codon disappearance that this is not the case. Of the 10 codons that have been reassigned, 6 have A or U at the third position: this is what we would expect if codon reassignment were random with respect to GC content! In particular, AAA should never have disappeared or been reassigned in any mitochondrion, yet it is absent in *Balanoglossus carnosus* (Castresana, Feldmaier-Fuchs et al. 1998). There are two possible ways of explaining this. First, perhaps the lineages that have reassigned AT-rich codons underwent a period of GC pressure sufficient to make those codons disappear. If this is true, it is surprising that no GC-rich mitochondria at all survive to the present, in any of the 132 species examined to date! Second, it is possible that GC content is not a useful measure of genome composition bias. We explore this latter idea further in the next section.

3.2.5 Mitochondrial Mutation: Beyond GC Content

In bacteria, the complete set of protein-coding sequences follows Chargaff's Rule, in that C=G and A=T, despite sampling only one strand. Since genes tend not to overlap, this is not a consequence of chemistry (as is Chargaff's Rule for double-stranded DNA) but of statistics. If there are no differences in selection or mutation between strands, or if genes are more or less evenly distributed between strands, a forward mutation on one strand is indistinguishable from the complementary mutation on the other strand. Thus Chargaff's Rule should hold even within a strand (Sueoka 1995). Consequently, from the frequency of one nucleotide, such as T, it is possible to make very good predictions about the remaining three (Fig. 2a).

In contrast, mitochondria tend to encode all their genes on a single strand. Additionally, the mutation spectra can be very different between the two strands (De Giorgi, De Luca et al. 1991). We find only and precisely two correlations between nucleotide frequencies in mitochondrial coding: T is negatively correlated with C ($r = -0.95$) and A is negatively correlated with G ($r = -0.85$). For no other pair of nucleotides can knowledge of the frequency of one explain more than 5% of the variance in the other (Fig. 2b)! This is consistent with the idea that nearly all mutations in mitochondria are transitions induced during the long single-stranded stage in replication.

We explored nucleotide bias further by plotting least-squares linear regression lines for each of the 64 codons against each of the 4 nucleotide frequencies at the third position. We defined a codon as 'in range' for disappearance if the x intercept for its regression line was within the range of values that mitochondria have actually taken. For instance, the amount of C at the third position ranges from 1.3% to 44%. Consequently, any codon for which the x intercept vs. C at the third position was between 1.3% and 44% was considered to be in range for disappearance. This is a liberal criterion for disappearance since 40 codons (nearly 2/3 of the total) are in range for disappearance on one graph or another. However, as we discuss later, no other plausible criterion does a better job of predicting which codons will disappear.

For 44 of the 64 codons, the frequency of at least one of the bases explains more than half the variance (i.e. $|r| > 1/\sqrt{2}$) in the frequency that codon. On average, we can predict 60% of the variance in the usage of a particular codon by regressing it on the frequency of one of the four 3rd position nucleotides in that genome. But to attain this accuracy we need to test it against all 4 nucleotides and decide *post hoc* which gives the best association. If we instead use the particular nucleotide at the codon's 3rd position as the independent variable, we do almost as well, explaining 54% of the variance in codon usage (i.e. more than half). Thus, from first principles, more than half of codon usage in mitochondria is explained by the nucleotide

composition of the genome in which the codons exist. Therefore genomic nucleotide composition seems to be the primary determining factor in mitochondrial codon usage.

How much better is this approach than use of third-position GC content? Because the third position changes much more rapidly (presumably because it is under much less selection) than the other positions, changes in codon frequency can, to a rough approximation, be reduced to the frequency at the 3rd position overall of their 3rd position base. This works for 2/3 of codons. We find that for 46 out of 64 codons, the 3rd position base is a better predictor than any other nucleotide. This gives much better correlations with individual codon frequencies than 3rd position GC: on average, third position GC content explains only 35% of the variance in the frequency of individual codons. On average, 32% of the variance not explained by GC content is explained by one of the 4 nucleotides individually. Thus predictions based on genome GC content are inaccurate compared to predictions based on individual nucleotide frequencies.

Now using our best predictor, Figure 3 shows the various possibilities for predicted and actual disappearance. Fig. 3a gives an example of a codon (CUC) that is driven out of existence by mutation bias. As the frequency of C at the third position lessens, its frequency approaches, and finally becomes, zero at low C. However, Fig. 3b shows an example of a codon (UGU) that is also heavily influenced by mutation bias and should disappear but never actually does. Similarly, Fig. 3c gives an example of a codon that is not predicted to disappear and doesn't. Fig. 3d is an example of a codon that is not predicted to disappear but in fact does, and at genome compositions where disappearance is surprising. Thus all combinations of predicted disappearance and actual disappearance coexist; genomic composition and codon disappearance have no obligatory interrelation. Figure 3b in particular suggests that mutation pressure is not sufficient to make a codon disappear.

Fig. 3e is AAA, an example cited as a paradigm of codon disappearance (Castresana, Feldmaier-Fuchs et al. 1998). However, the genome in which it disappears still has over 25% A overall at the third position. Thus it seems unlikely that nucleotide bias caused AAA to disappear from hemichordate genomes.

Remarkably, similar relationships hold even where the identity of codons has changed. Fig. 4 gives the regressions for two codons, AUA (Fig. 4a) and AGA (Fig. 4b), that have switched meaning repeatedly in different lineages. The distributions for the AUA response to A at the third codon position are indistinguishable whether AUA encodes Ile or Met; the distributions for AGA = Ser and Arg are similar. In contrast to drift, note that selection can efficiently make a codon disappear: when AGA = termination, it is much rarer than would be predicted from the third position A content, over a wide range of genome compositions.

3.2.6 Testing the Codon Capture model

The regression analysis of codon frequencies on base composition therefore indicates that we can predict a codon's frequency quite well from the genome composition. However, predicting whether a codon actually *disappears* because of biased composition is another matter. Although codon disappearance has been noted in several mitochondrial genomes, it is a relatively rare event. Consequently, the factors involved in a codon disappearance are not identical to those that govern its frequency of occurrence across all genomes.

Genome composition consistently affects the frequency of particular codons. If this factor is of paramount importance in codon reassignment (as suggested by Osawa and Jukes), the codons that disappear should be the same as the codons that are reassigned. This is testable, because several different groups of organisms span the full range of nucleotide contents. We therefore have several independent samples of mitochondrial genome evolution under the same genomic composition. Consequently, if universal forces are at work, the same codons should repeatedly disappear and/or be reassigned in different taxa. The fact that UGA has been reassigned to Trp many times independently, in 6 different lineages of mitochondria and

in the Mycoplasmas, confirms that specific codons are prone to reassignment along a particular pathway.

We now apply standard statistical calculations to determine whether observed codon reassignments are associated with predicted events under various theories. In this way, we can bring an objective, quantitative and generally-agreed criterion to bear on questions of this form: does the natural history of the mitochondrion support the predictions of codon capture, genome compression and/or the ambiguous intermediate? Here we use simple 2 x 2 tests for independence to test the following:

1. Are the codons made rare by nucleotide composition fluctuations the same codons that disappear?
2. Are the codons that disappear the same codons that are reassigned?
3. Are the codons made rare by nucleotide composition the same codons that are reassigned?

We classify each codon (e.g. has disappeared or not; has been reassigned or not), then test for association between classifications. If the x intercept for codon frequency versus any of the four genomic base frequencies is in the range of base composition that actual mitochondria have explored, we count the codon as predicted to disappear because of directional mutation pressure. We use the G test for independence since it is more robust to small sample sizes (Sokal and Rohlf 1995), but the more familiar chi-squared test gives qualitatively the same results (data not shown).

The relationship between codon disappearance and codon reassignment is quite clear-cut (Table 2a): of the 11 codons that have been reassigned in some lineage, 8 have disappeared in some other lineage. The probability that disappearance and reassignment are independent events is only 0.004. Similarly, there is a strong relationship between predictions of disappearance and actual disappearance; probability of independence P = 0.006. However, there is no significant association between the codons predicted to disappear and those that are actually reassigned, P = 0.22!

How can we explain this surprising result? Fig. 5 is a Venn diagram that shows which codons have disappeared, which are predicted to disappear, and which have been reassigned, along with all possible combinations of these three states. The lack of association between predicted disappearances and actual reassignments is due to the large number of codons predicted to disappear but not actually reassigned. In light of this, it is possible that our criterion (x intercept in range for any of the 4 bases) is too liberal in predicting disappearance. However, other initially plausible criteria do much worse at predicting which codons will disappear, which will be reassigned, or both. Neither the 3rd position GC content nor N3 (where N is the nucleotide at the third position of the codon) do significantly better than chance at predicting which codons disappear or which are reassigned (Table 2). The members of a transition pair are highly correlated, and so it might be the case that one would (by chance) do better than the other at predicting the codon usage. However, this does much worse both at predicting disappearances and reassignments than does our initial criterion (Table 2). Thus, no alternative criterion narrows only the range of codons that are predicted to disappear but neither disappear nor are reassigned.

These data are compatible with the following scenario: biased mutation can and does cause codons to disappear from the genome. These codons are more likely to be reassigned, since changing a rare codon is less likely to be deleterious. Some codons disappear for reasons unconnected to mutation bias, and these are also more likely to be reassigned. However, actual disappearance is not a necessary condition for reassignment, implying that at least in some cases a codon is read with two meanings as an intermediate stage. Since we know that this is possible in both the *Candida* nuclear genome and in *Bacillus*, it should not come as a surprise that it can also happen in mitochondrial genomes.

3.2.7 Does genome size matter?

Here we test the hypothesis that mitochondrial code evolution is driven by genome minimization (Andersson and Kurland 1991; Andersson and Kurland 1995) – the idea that mitochondrial genomes are selected to be as small as possible. The size decrease involved in losing a tRNA arguably confers sufficient selective advantage to counterbalance the effects of losing it. Thus code evolution could be driven by loss of tRNAs.

If this were true, we would expect that genomes that deviate further from the canonical code would be smaller and use fewer tRNAs, and that smaller genomes would use fewer tRNAs.

Surprisingly, we find that none of these initially plausible predictions is borne out. Fig. 6a and 6b show the relationship between genome size and number of deviations from the canonical code for (a) all organisms, and (b) just the metazoa. In neither case is there any relationship (the regression line in Fig. 6a is primarily influenced by *Beta vulgaris*, a land plant that has a huge mitochondrial genome). Fig. 6c shows the relationship between the number of tRNAs encoded by the mitochondrion and the number of codons in which each code differs from the canonical one. The number of tRNAs varies widely (the jellyfish *Metridium senile* has only 2, presumably importing the rest from the nucleus), but is not linked to the number of deviations from the canonical code. Metazoans other than Metridium occupy the cluster of 5 dark points above the regression line; there is no association for the metazoans considered separately from other groups. Note that the "association" here is not significant, but more importantly, the regression line actually slopes in the opposite direction from that predicted. Fig. 6d and 6e show that the number of tRNAs is not correlated with genome length either, whether across all organisms or just within the metazoa.

This last result, especially, is rather surprising. Why do smaller mitochondria not, in general, encode fewer tRNA molecules? In order to test whether we would have seen an effect had it been there, we tested several other variables that we thought might correlate with genome size. Smaller genomes do have significantly fewer bases involved in coding sequences (Fig. 6f), ribosomal RNA (Fig. 6g), and spacer regions (Fig. 6h) than do larger genomes.

Most of the observed changes in the code are in metazoa (although this could be observer bias: 113 of the 132 species for which complete genes are available are metazoa). Thus we investigated whether the metazoa have particularly extreme strategies for genome minimization. The metazoa encode a common and much reduced set of critical proteins, so it is unsurprising that average protein coding sequence length does not respond greatly to changes in genome size (Fig. 6i). More surprising is that the average length of each tRNA varies significantly with genome size (Fig. 6j), reaching a low of about 55 bases on average in the nematodes: this corresponds to a loss of nearly 20 nucleotides in every tRNA! Similarly, the average length of rRNA molecules declines precipitously in extremely small genomes (Fig. 6k), although the rRNA have not shortened appreciably within the chordates.

Given that mitochondria really do take extreme measures to minimize their genomes, the fact that losing tRNA molecules is not one of these measures is remarkable. Nonetheless, if such an association had been present we should have found it: that we uniformly found no trace of the predicted tendencies suggests that genome minimization has not been a factor in shaping mitochondrial genetic codes.

3.2.8 The Ambiguous Intermediate Revisited

There are two plausible pathways that could lead to ambiguous translation in mitochondria. The first is alteration of the tRNA at locations distant from the anticodon, inducing G-U mispairing at the first codon position and C-A or G-A mispairing at the third codon position. The second is loss or gain of base modification, which (because modification can be incomplete) allows complete gradation in the relative frequencies of two interpretations of a

codon. Neither of these states can be inferred from the DNA sequence of tRNA genes or from other DNA data, since both are 'hidden' in sites distant from the anticodon.

Nevertheless, we can ask whether the set of observed changes is more compatible with the ambiguous intermediate model than chance would predict. In particular, the ambiguous intermediate model predicts that codons should be reassigned to amino acids that are accessible via known translational ambiguities, whereas codon capture merely predicts that the change in tRNA should be a single point mutation. Thus the ambiguous intermediate model predicts the amino acids as well as the codons that should be reassigned. We can thus test whether surprisingly many of the observed changes are consistent with the model, given the fraction of possible changes that are consistent.

In one version of this analysis, we use a composite phylogeny of lineages with variant nuclear and/or mitochondrial codes (Knight, Freeland, and Landweber, in press *Nature Genetics Reviews*), the number of changes in which is presented in Table 3. In a second type of calculation, we conservatively count each reassignment only once, ignoring its occurrence in disparate lineages.

Changes involving NNU or NNC are never independent (since both are always recognized by G, and no anticodon base at the first position can recognize NNC uniquely). Therefore treating NNY as one codon, there are 48 'effective codons'. If there were no restrictions on how codon identity could change, each of these 48 blocks could be reassigned independently to each of the 20 alternatives (either the 20 amino acids, for a stop codon, or the other 19 amino acids plus stop, for a sense codon). This yields 960 possible changes, of which 800 are allowed. There are 8 blocks that are the only block for an amino acid (e.g. UUY = Phe), which would presumably have deleterious results if changed.

If we alternatively assume that only single point mutations are allowable, the canonical code structure allows 312 possible changes. The number 312 takes account of changes that result in the same amino acid, and of amino acids that are not accessible by any point mutation for a given codon. We consider changes of NNA and NNG to be potentially independent even when they occur together: in ciliates, there is evidence that the tRNAs decoding UAG and UAA as sense codons have changed because of independent changes in two tRNAs (Hanyu, Kuchino et al. 1986).

First we count each change in identity only once, no matter how many times it has occurred. This maximizes the robustness of the result, since errors in the phylogeny cannot affect it. However, it throws away much of the information we have. The second calculation takes account of multiple reassignments by weighting each change by the number of times it seems to have occurred in independent lineages. We do each test both for mitochondria and for the combined nuclear/mitochondrial data set. since the ambiguous intermediate model depends primarily on the physical chemistry of tRNA/mRNA/ribosome interactions and thus anticipates similar changes in mitochondria and nuclei.

Of the 800 possible changes (or 312 single-base changes), only 15 are compatible with G-U mispairing at the first position, or C-A or G-A mispairing at the third position. Thus, we would expect only 15/800 changes (on average) to be compatible with the model if all nucleotide changes are possible, or 15/312 changes to be compatible if only single point substitutions are allowable. Some known changes (such as the CUN Ser → Thr reassignment in yeast) are not single point mutations, so we exclude these changes when only point mutations are allowable. Table 4 gives the probability of observing at least as many changes compatible with the model by chance for the various circumstances discussed above.

In fact, the ambiguous intermediate model makes the best predictions of all models (not shown), explaining the observed changes far better than chance would predict under all circumstances (Table 4). Strikingly, it does well even using the most conservative assumptions (when changes are not weighted by their frequency of occurrence) and does better when all available data are brought to bear (nuclear changes are considered in addition

to mitochondrial changes). This supports the idea that ambiguous intermediates play a general role in codon reassignment.

3.2.9 Conclusions

We show that mutation pressure affects and can be used to predict the frequency of particular codons in mitochondrial genomes. Codon frequencies also yield moderately good predictions about which codons will disappear from genomes entirely. Additionally, we find a link between codon disappearance and codon reassignment. However, the codons that disappear and are reassigned are not necessarily the codons that disappear because of mutation pressure!

Consequently, the causes of codon disappearance and reassignment may be more complex than the original neutral model proposed by Osawa and Jukes.

Some codons do not disappear despite an apparently strong tendency to do so, and codons also disappear despite contrary predictions (Figure 3 above). Conceivably, this behavior reflects the need for sporadic and unpredictable intermediate mutations, such as mutations that make the meaning of these codons ambiguous. The incomplete but significant tendency of codons to be reassigned on return from disappearance (Figure 5) might also be explained this way. There may be no phenotypic penalty at all for a translational ambiguity that extends to a codon already disappeared. Thus the return of the codon with a new meaning may be facilitated, dependent on an ambiguity that has become possible during the period of disappearance. The codon may also return with an unambiguous meaning, as originally suggested by Jukes and Osawa (Osawa and Jukes 1988). However, the mutational target in tRNAs alone for ambiguity due to both sequence changes (Schultz and Yarus 1994; Schultz and Yarus 1994) and modification changes (Muramatsu, Nishikawa et al. 1988; Perret, Garcia et al. 1990; Matsuyama, Ueda et al. 1998; Tomita, Ueda et al. 1999) is large, and therefore this may be a prevalent mechanism.

We find no evidence that changes in the genetic code are linked to genome minimization. Mitochondria go to startling lengths to minimize their genomes, but changes in the number of tRNAs are not linked to changes in genome size, nor to the number of deviations from the canonical genetic code. Genome minimization can thus be ruled out as a factor in generating the variety of genetic codes found in extant mitochondria. The coding changes that previously (Knight, Freeland, and Landweber, in press *Nature Genetics Reviews*) seemed consistent with genome compression (Andersson and Kurland 1995) are mostly expansions that extend wobble pairing within a 4-codon box. These may be reinterpreted as transient coding ambiguity at the third codon position (known to be intrinsically the most ambiguous), followed by loss of the original tRNA specific to the smaller, changed subset of codons.

A parallel wobble expansion (UGG Trp → UGA, UGG Trp), the most frequent of all codon reassignments, is apparently taking place in *Bacillus subtilis* (Lovett, *et al*., 1991). This change has been traced to a $tRNA^{Trp}$ with expanded capability for C-A wobble in the third codon position (Matsugi, *et al*., 1998). *Bacillus* $tRNA^{Trp}$ has D-arm (Hirsh, 1971) and anticodon stem sequences (Schultz and Yarus, 1994c) previously shown to enhance C-A wobble in E. coli. The resulting ambiguity of UGA (as stop and Trp) is sufficiently effective that *Mycoplasma* genes (where UGA = Trp) can be read *in vivo* in *Bacillus* to give full-length products (Kannan and Baseman, 2000). Thus *Bacillus subtilis* UGA, where a release factor could be lost in conformance with genome compression, is actually proceeding through an ambiguous intermediate codon, as suggested above.

The role we find for mutation pressure is somewhat as envisioned by Tom Jukes (King and Jukes 1969; Osawa, Jukes et al. 1992). Disappeared codons are more easily reassigned (Figure 5). However, the repetitive, specific nature of most reassignments requires another explanation, because drift does not constrain the possible reassignments. Codon reassignment does not require codon disappearance (Figure 5), suggesting a second route.

We suggest that codons that transiently possess two meanings (ambiguous codons) supply this route. Coding ambiguity is demonstrated for CUG reassignment in *Candida* (Santos, Cheesman et al. 1999) and in the assignment of UGA as both stop and Trp in *Bacillus* (Lovett, Ambulos et al. 1991). Further, three independently demonstrated tRNA coding ambiguities explain the observed selectivity of codon reassignment very well, using few assumptions (Table 4). Therefore we suggest that drift of codon frequencies and coding ambiguity often combine in a reassignment pathway (Schultz and Yarus, 1996) whose overall properties reflect both the stimulatory effect of the former and the specificity of the latter.

Table 1: code changes in mitochondria and nuclei

	Change in Nucleus	No Change in Nucleus
Change in Mitochondria	UAA, UAG, UGA, CUG (4 codons)	CUU, CUC, CUA, AAA, AGA, AGG (6 codons)
No Change in Mitochondria	—	54 codons

This table partitions the codons by whether or not they have changes in the nucleus (columns) and whether or not they have changed in mitochondria (rows): data is from the translation tables at NCBI. The probability of independence between the set of codons that have changed in nuclei and mitochondria is 3.3×10^{-4} (by Fisher's Exact Test). If we include known unassignments as well as reassignments, the probability of independence drops to 1.0×10^{-5}. Thus, the same changes take place in both systems.

Table 2: relation between predicted codon disappearance, actual disappearance and codon reassignment. The P value gives the probability that discrepancies as large as observed would be found if the two variables in each table were really independent. Each table is a 2 x 2 contingency table giving the number of codons that falls into either of the categories listed across the columns or down the rows. '+' indicates that the appropriate event applies to the codons in that row or column; '–' indicates that the event does not apply. The G value is calculated according to standard procedure, incorporating the Williams correction (Sokal and Rohlf 1995).

The first table gives the relationship between actual reassignment and actual disappearance, which are significantly associated (P = 0.004). The remaining paired tables use different criteria for predicting which codons will disappear. Varied criteria change the relationships between predicted and actual disappearance and between predicted disappearance and actual reassignment (the relationship between actual disappearance and actual reassignment is not affected).

Criteria for predicted disappearance are as follows. OR: a codon is predicted to disappear if the x intercept of the regression of its frequency on any of the 4 nucleotide frequencies is inside the range of frequencies of that 3rd position nucleotide in any known mitochondrial genome. GC3: predicted to disappear if the x intercept on the 3rd position GC regression is within the range of known 3rd position GC contents. Note that using 3rd position GC, actual and predicted disappearance are not significantly associated!

	Disappear?	
Reassigned?	+	-
+	8	3
-	15	38
	G = 7.17, P = 0.004	

Using OR:

	Disappear?			Reassigned?	
Predicted?	+	-	Predicted?	+	-
+	19	21	+	8	32
-	4	20	-	3	21
	G = 6.43, P=0.006			G = 0.59, P = 0.22	

Using GC3:

	Disappear?			Reassigned?	
Predicted?	+	-	Predicted?	+	-
+	7	8	+	3	12
-	16	33	-	8	41
	G = 0.92, P = 0.17			G = 0.09, P = 0.38	

Codon	From	To	Mt	Nuc	Total	G-U 1?	C-A 3?	G-A 3?	SY?
UGA	Ter	Trp	6	4	10		+		TRUE
AUA	Ile	Met	2		2		+		TRUE
AGA	Arg	Ser	1		1			+	TRUE
AGG	Arg	Ser	1		1				FALSE
AUA	Met	Ile	2		2			+	TRUE
AAA	Lys	Asn	2		2			+	TRUE
AGA	Ser	Gly	1		1				FALSE
AGG	Ser	Gly	1		1				FALSE
UAA	Ter	Tyr	1		1			+	TRUE
CUY	Leu	Thr	1		1				FALSE
CUA	Leu	Thr	1		1				FALSE
CUG	Leu	Thr	1		1				FALSE
AGA	unassigned (Ser)	Ter	1		1				FALSE
AGG	unassigned (Ser)	Ter	1		1				FALSE
AGA	unassigned (Ser)	Gly	1		1				FALSE
UAG	Ter	Leu	2		2				FALSE
UAG	Ter	Ala	1		1				FALSE
UCA	Ser	Ter	1		1				FALSE
UAA	Ter	Gln		7	7	+			TRUE
UAG	Ter	Gln		7	7	+			TRUE
UGA	Ter	Cys		1	1			+	TRUE
CUG	Leu	Ser		1	1				FALSE

Table 3: all known changes in the genetic code (from Knight, Freeland and Landweber, in press *Nature Genetics Reviews)*. Column headings: 'From': original meaning assigned to the codon. 'To': new meaning assigned to the codon. 'Mt': number of independent changes in mitochondrial lineages. 'Nuc': number of independent changes in nuclear lineages. 'G-U 1?': Plus if the reassignment is consistent with G-U mispairing at the first codon position. 'C-A 3?': Plus if change is consistent with C-A mispairing at the third codon position. 'G-A 3?': Plus if change is consistent with G-A mispairing at the third codon position. 'SY?': TRUE if change is consistent with any of the three mispairing mechanisms originally suggested by Schultz and Yarus (Schultz and Yarus 1994); FALSE otherwise.

Table 4: Tests of concordance between code changes and predictions of the ambiguous intermediate hypothesis. There are 15 detectable Schultz-Yarus changes (Methods). P is the cumulative binomial probability of getting as many or more Schultz-Yarus changes as observed when random choice would yield 15/800 successes (for all changes), or 15/312 successes (allowing only point substitutions). Column headings: 'All changes': considers all possible changes in the code. 'Point mutation only': considers only changes that could have arisen by single nucleotide substitutions (excluding, for instance, the CUN Ser → Thr reassignment in yeast mitochondria). '+': number of changes consistent with the Schultz-Yarus ambiguity mechanism. '-': number of changes inconsistent with the Schultz-Yarus reassignment mechanism. 'P': probability of observing at least as many changes consistent with this mechanism if all point changes were equally likely. Row headings: 'weighted': each phylogenetically distinct change (from Table 3) is counted as independent. 'Unweighted': each reassignment is counted only once, even when it has apparently recurred in different lineages. 'Mt only': only changes in mitochondria are counted. 'Mt and nuc': changes in both mitochondrial and nuclear lineages are counted.

		All changes			Point mutation only		
		+	-	P	+	-	P
weighted	Mt only	14	13	1.1E-17	14	8	7.8E-14
	Mt and nuc	33	14	2.7E-46	33	8	2.1E-36
unweighted	Mt only	6	11	4.5E-07	6	6	8.9E-06
	Mt and nuc	9	12	6.9E-11	9	6	5.3E-09

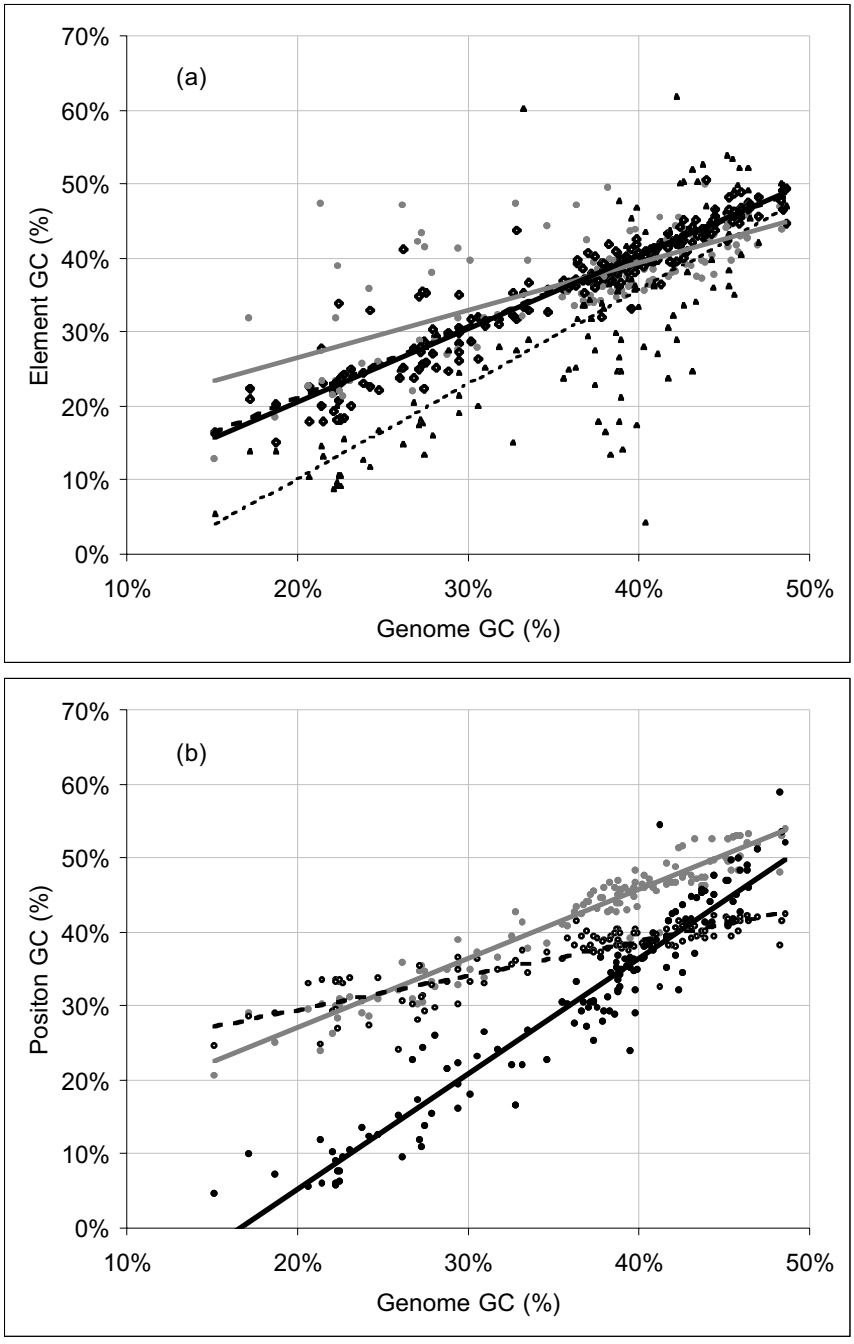


Figure 1: Relations between GC content of genome components. (a) Spacer: small white triangles with a thin, dashed line; protein-coding regions: black diamonds, solid black line; rRNA: white diamonds, thick dashed line; tRNA: gray circles, gray line. (b) 1st, 2nd, and 3rd position in mitochondrial genes. The 3rd position (black) changes much more rapidly in response to GC content than either the 1st (gray) or 2nd (white, dashed line) position.

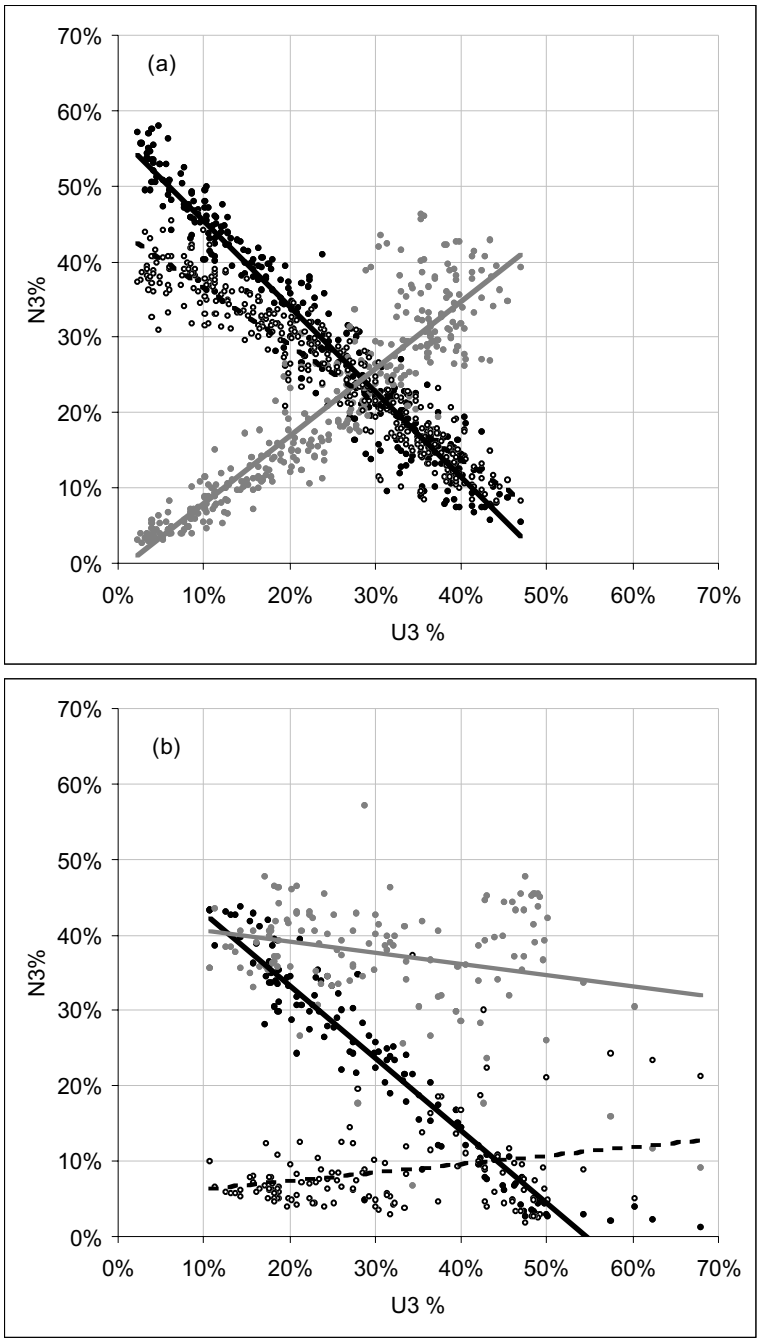
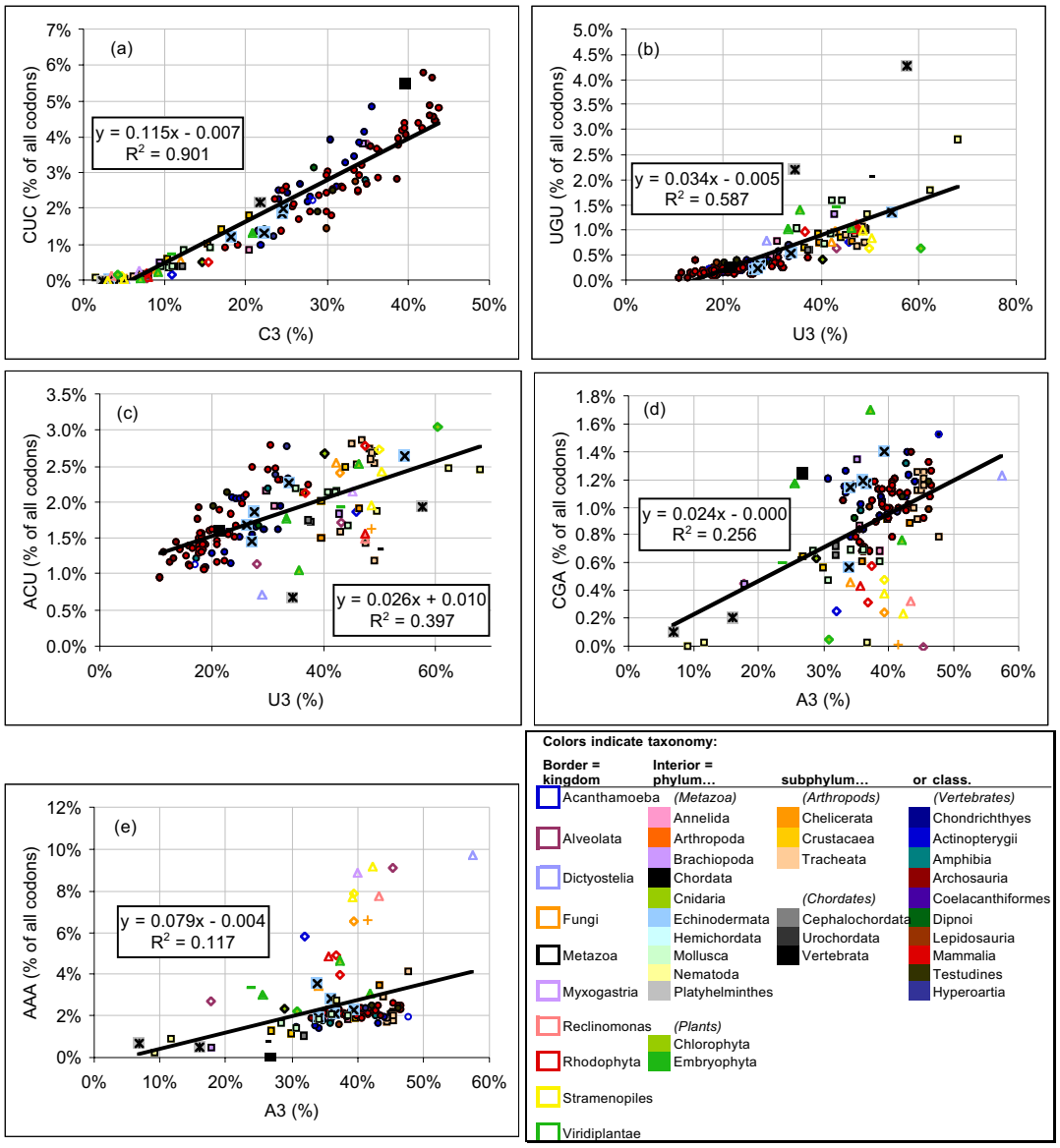
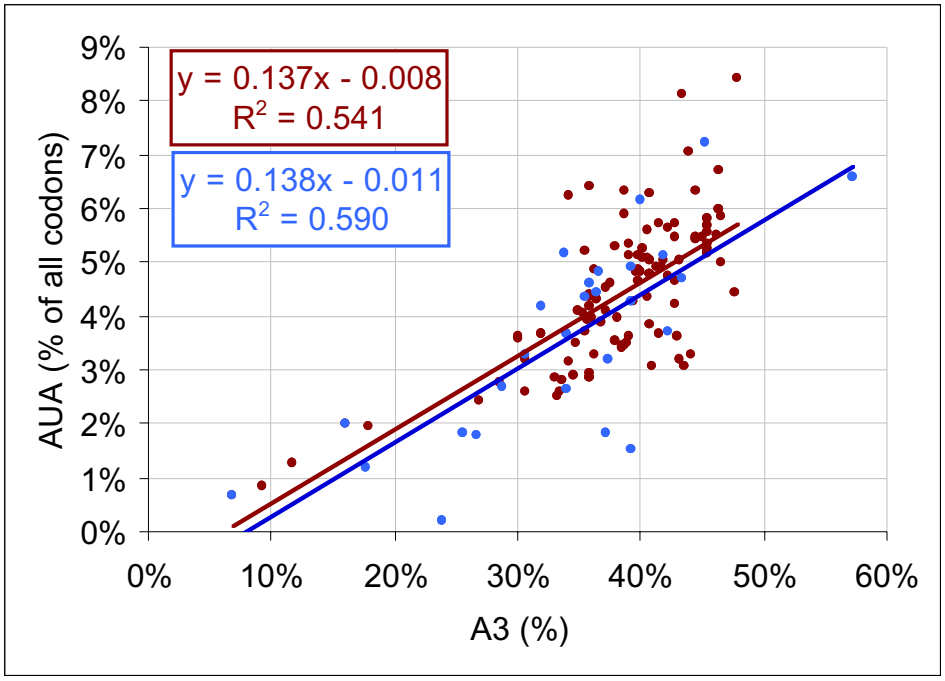


Figure 2: U at 3rd codon position vs. other 3rd codon nucleotide frequencies in (a) Bacteria/Archaea and (b) mitochondria. In nuclear lineages, all four nucleotides are highly correlated with one another ($r^2 > 0.85$), but in mitochondria only C (black) is correlated with U; A (gray) and G (white) are uncorrelated (though they are correlated with each other).



Symbols indicate translation table:		
Trans. Table	# species	symbol
1	10	Δ
2	72	o
3	1	+
4	9	◇
5	31	□
9	6	×
13	1	–
14	2	*
22	1	—

Figure 3: Effects of nucleotide bias on individual codons. The regression line in each case is for the sum of all species. Examples are shown for codons that are or are not predicted to disappear, and that do or do not actually disappear.



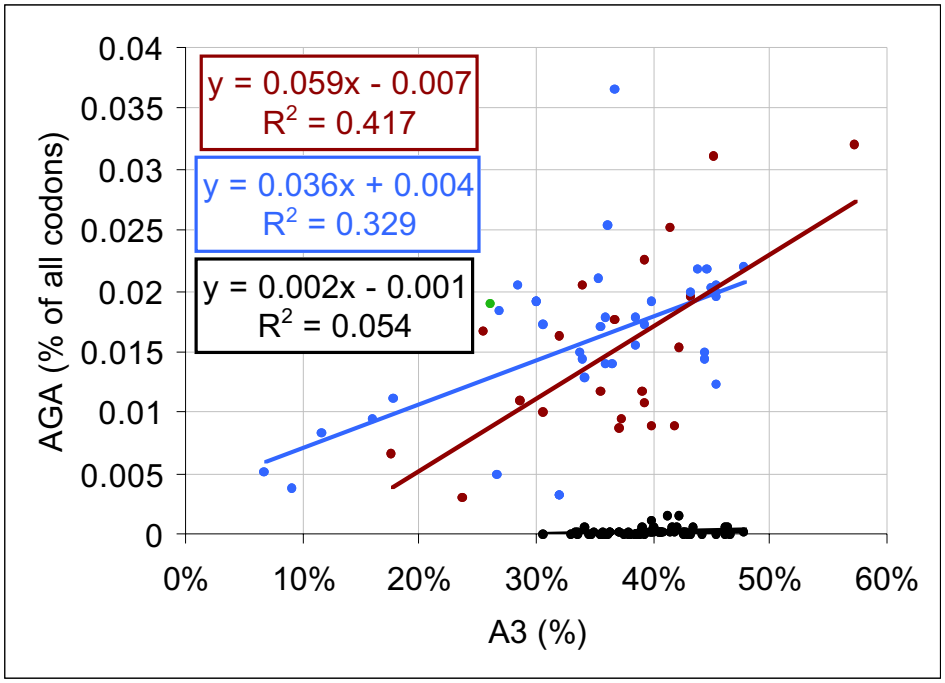


Figure 4: Codon usage with multiple reassignments. (a) AUA = Met (Red), Ile (Blue). Two meanings have the same regression line. (b) AGA = Arg (Blue), Ser (Red), Gly (Green), and Stop (Black). Distributions intermingle, except where AGA = Stop. Ser and Arg distributions do not differ (2-dimensional Kolmogorov-Smirnov test: N(R) = 21; N(S) = 39; D = 0.28, P = 0.27).

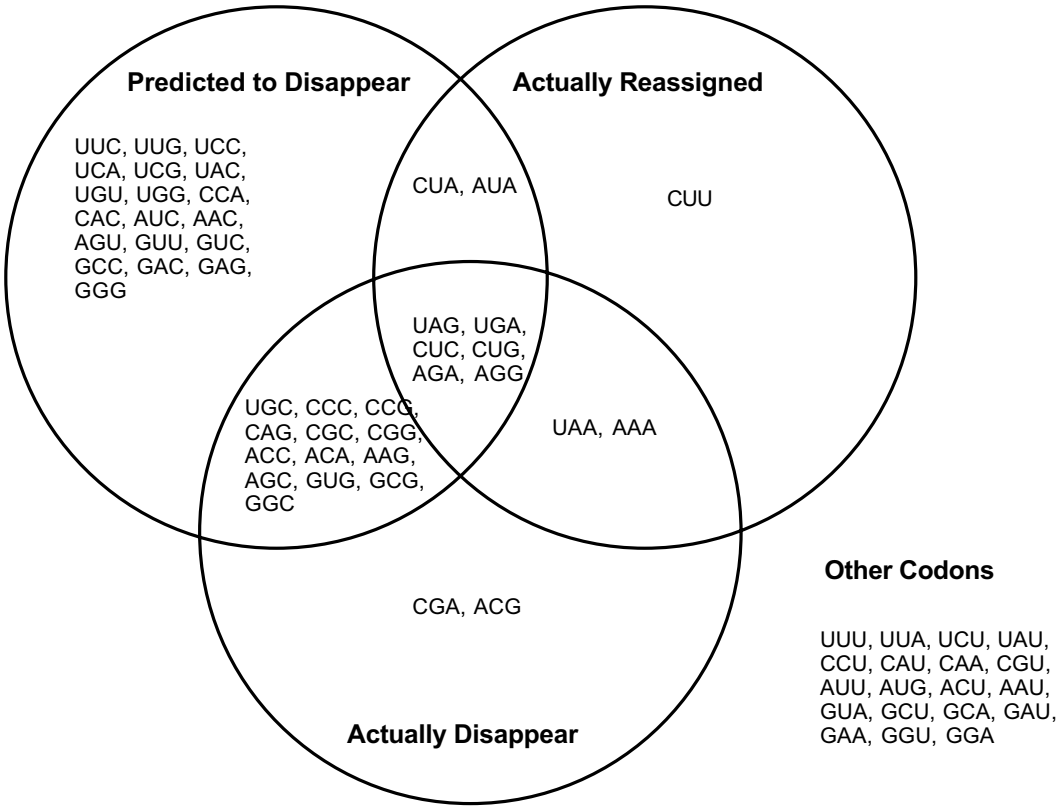
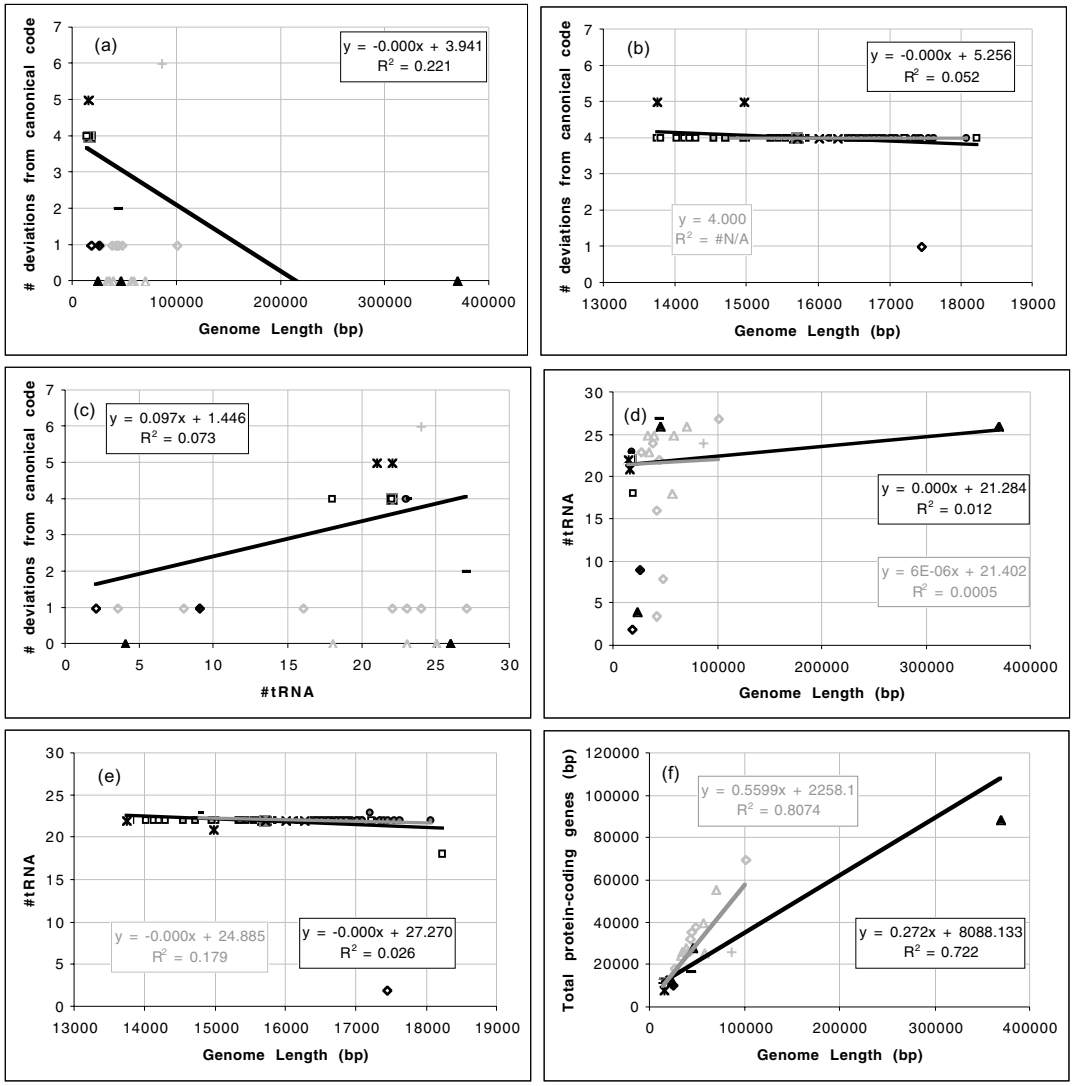
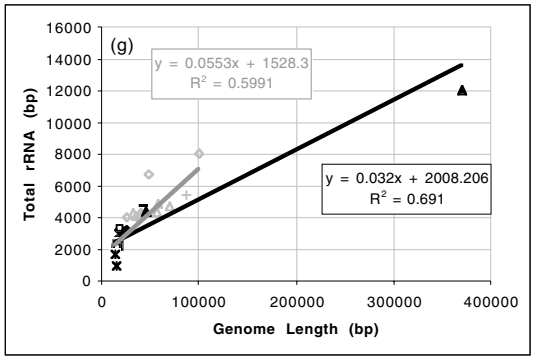
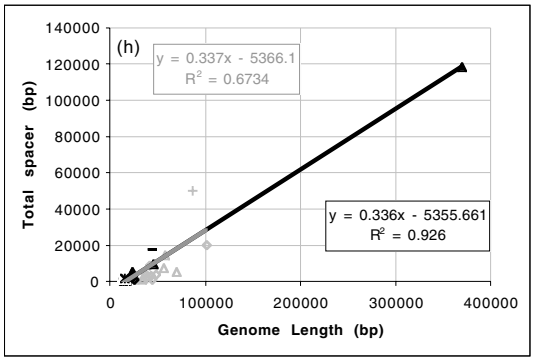


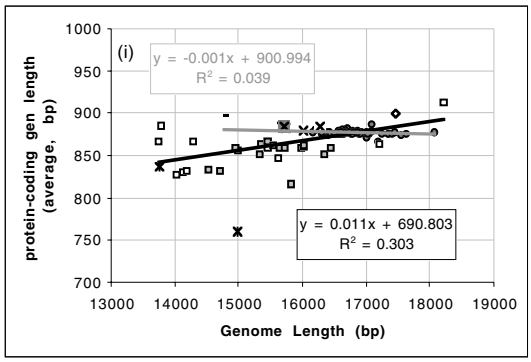
Figure 5: Venn diagram relating actual and predicted codon disappearance, and codon reassignment. Although actual disappearance is linked to reassignment, it seems that (a) some codons disappear from genomes for reasons unconnected with mutation pressure, and (b) some codons can be reassigned without actually disappearing. The latter observation implies a stage of ambiguous translation.

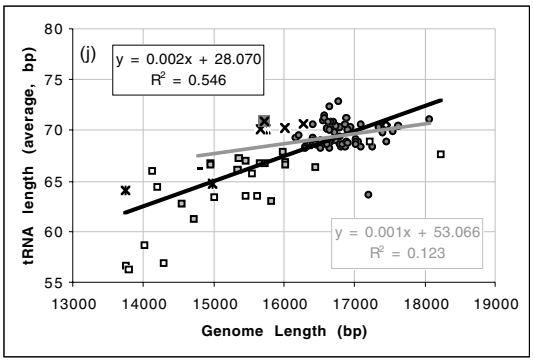
Fig. 6 Relationships between genome size and other gene and genome characteristics. Different symbols denote different translation tables, as in Fig. 4. Symbols for metazoa have black outlines: chordates have dark gray interiors, arthropods light gray, and other metazoa white. Plants are denoted by solid black symbols: other taxa (protists, fungi) by gray outlined symbols. Where there is one regression line on a graph, that line applies to all organisms. For graphs with two regression lines, either (1) the black line applies to all organisms and the gray line applies to non-plants only (d, f, g, h), or (2) the black line applies to all metazoa, and the gray line applies to chordates specifically (b, e, i, j, k). The genome length and number of deviations from the genetic code, either for all organisms (6a) or for the metazoa specifically (6b). The number of tRNA molecules a mitochondrial genome encodes and the extent to which that genome's code differs from the canonical code (6c); the number of tRNAs and the genome's length for all taxa (6d) or for the metazoa specifically (6e). Genome length and number of bases involved in protein-coding genes (CDS) across all taxa (6f), in rRNA (6g), and in spacer (6h). Genome length and average length of genes in metazoa (6i). Genome length and average tRNA length (6j) or average rRNA length (6k).

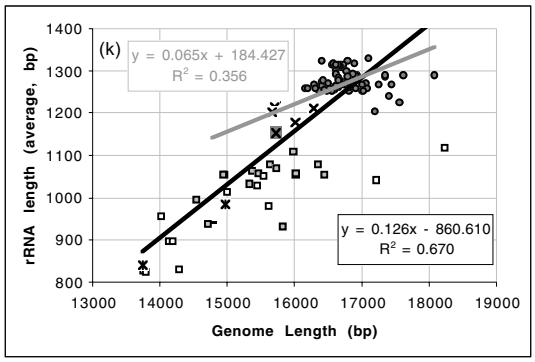



(g)
y = 0.0553x + 1528.3
R² = 0.5991
y = 0.032x + 2008.206
R² = 0.691
Total rRNA (bp)
Genome Length (bp)


(h)
y = 0.337x - 5366.1
R² = 0.6734
y = 0.336x - 5355.661
R² = 0.926
Total spacer (bp)
Genome Length (bp)


(i)
y = -0.001x + 900.994
R² = 0.039
y = 0.011x + 690.803
R² = 0.303
protein-coding gen length (average, bp)
Genome Length (bp)


(j)
y = 0.002x + 28.070
R² = 0.546
y = 0.001x + 53.066
R² = 0.123
tRNA length (average, bp)
Genome Length (bp)


(k)
y = 0.065x + 184.427
R² = 0.356
y = 0.126x - 860.610
R² = 0.670
rRNA length (average, bp)
Genome Length (bp)

3.3 The Molecular Basis of Nuclear Genetic Code Change in Ciliates

In Chapter 1.4, I conjectured that a particular change in Tetrahymena eRF1, in the highly conserved NIKS domain, led to a change in stop codon specificity (eRF1 normally recognizes the three stop codons UAA, UAG, and UGA). Prof. Landweber suggested that we sequence eRF1 from a range of other ciliates with variant genetic codes, which were known to have evolved convergently, to test whether independent code changes in different lineages had the same molecular basis. To this end, Catherine Lozupone sequenced eRF1 from several lineages of distantly related ciliates, spanning more genetic diversity than is found in the plants, animals, and fungi combined.

Unfortunately, the pattern was not as simple as a single amino acid change, and none of the other ciliates had that particular change. In order to test whether any *changes correlated with change in the genetic code, I developed a new statistical technique for correlating molecular changes with changes on a phylogeny. A phylogeny is fundamentally a graph, where each species is a vertex and the transition between each ancestral sequence and its descendant is an edge. My method basically does many replicated tests for independence between edges in which the code changes and edges in which the sequence changes, identifying which parts of the sequence change in the same evolutionary interval when a change in the code occurred.*

This technique was able to recapture the principal functional domains of the protein (the NIKS anticodon recognition site and the GGQ ribosome-binding site), as revealed by the crystal structure, and highlighted specific areas of the protein that conferred suppressor activity when mutated in yeast by Ian Stansfield's lab. I also set the program to extract states of amino acids that occurred if and only if the code was nonstandard. Surprisingly, this identified changes at exactly the same positions, and, in some cases, to the same altered amino acid, as had been found in the mutant yeast. This reinforces the idea that extant species can be used as a natural genetic screen for interesting phenotypes.

This chapter has appeared as a paper in Current Biology:

Lozupone, C. A., R. D. Knight and L. F. Landweber (2001). "The molecular basis of genetic code change in ciliates." Current Biology 11: 65-74.

The sequencing, phylogeny building, and database analysis were entirely performed by Ms. Lozupone. I developed and performed the statistical analysis, contributed Fig. 4, and wrote most of the material relating the work to changes in the genetic code in other taxa.

The molecular basis of nuclear genetic code change in ciliates

Catherine A. Lozupone, Robin D. Knight and Laura F. Landweber*

Background: The nuclear genetic code has changed in several lineages of ciliates. These changes, UAR to glutamine and UGA to cysteine, imply that eukaryotic release factor 1 (eRF1), the protein that recognizes stop codons and terminates translation, changes specificity. Here we test whether changes in eRF1 drive genetic code evolution.

Results: Database sequence analysis reveals numerous genetic code alterations in ciliates, including UGA → tryptophan in *Blepharisma americanum* and the distantly related *Colpoda*. We sequenced eRF1 from four ciliates: *B. americanum*, a heterotrich that independently derived the same eRF1 specificity as *Euplotes*, and three spirotrichs, *Stylonychia lemnae*, *S. mytilus*, and *Oxytricha trifallax*, that independently derived the same genetic code as *Tetrahymena* (UAR → glutamine). Distantly related ciliates with similar codes show characteristic changes in eRF1. We used a sliding window analysis to test associations between changes in specific eRF1 residues and changes in the genetic code. The regions of eRF1 that display convergent substitutions are identical to those identified in a recently reported nonsense suppression mutant screen in yeast.

Conclusions: Genetic code change by stop codon reassignment is surprisingly frequent in ciliates, with UGA → tryptophan occurring twice independently. This is the first description of this code, previously found only in bacteria and mitochondria, in a eukaryotic nuclear genome. eRF1 has evolved strikingly convergently in lineages with variant genetic codes. The strong concordance with biochemical data indicates that our methodology may be generally useful for detecting molecular determinants of biochemical changes in evolution.

Address: Department of Ecology and Evolutionary Biology, Princeton University, Princeton, New Jersey 08544

Correspondence: Laura F. Landweber
E-mail: lfl@princeton.edu

Received: **20 October 2000**
Revised: **26 November 2000**
Accepted: **27 November 2000**

Published: **23 January 2001**

Current Biology 2001, 11:65–74

0960-9822/01/$ – see front matter

Background

The genetic code was once thought to be universal among all organisms: once fixed, any change – tantamount to rewiring a keyboard – would cause deleterious changes in every protein [1]. We now know that the code can change; alternative genetic codes are found in most mitochondrial genomes [2], the nuclear genomes of the eubacterium *Mycoplasma* [3], the yeast *Candida* [4], diplomonads [5], the green alga *Acetabularia*, and a variety of ciliates [e.g., 6–10]. Ciliates are remarkable in this respect. Several different code variants have arisen independently, even within a single class [11]. For example, *Tetrahymena* and *Paramecium*, class Oligohymenophorea, and *Oxytricha* and *Stylonychia*, class Spirotrichea, translate UAA and UAG as glutamine (using only UGA as stop) [6, 12]. *Euplotes*, also a spirotrich, instead translates UGA as cysteine, using UAA and UAG for termination [13]. *Blepharisma*, class Heterotrichea, uses UAA to encode stop [8]: before this study, the translation of UAG and UGA in this species was unknown. Although GenBank includes a separate translation table for *Blepharisma*, called the "Blepharisma code," in which UAA and UGA encode stop and UAG encodes glutamine [8], we find no support for the translation of UAG as glutamine or UGA as stop in this species. In a 1995 study, Baroin Tourancheau et al. [11] sequenced the α-tubulin and phosphoglycerate kinase genes for members of 6 of the 10 currently recognized ciliate classes and suggested that members of the class Litostomatea and the heterotrich *Stentor coeruleus* may use the standard genetic code, whereas the karyorelictid *Loxodes striatus*, the heterotrich *Condylostoma magnum*, and the nassophorean *Zosterographus sp.* appear to use UAA and UAG to encode glutamine. These data indicate that alteration of the ciliate genetic code was not a single, ancient event, as initially supposed [14], but a relatively common event in ciliates [11].

Changes in the genetic code involve changes in tRNAs, tRNA-modifying enzymes, or release factors. These crucial components of the translation apparatus have changed in ciliates, conferring new meanings to specific codons. Altered tRNAs in organisms with variant genetic codes have provided some insight into the mechanism of genetic code changes. In addition to normal $tRNA^{Gln}$, *Tetrahymena thermophila* has two unusual glutamine-specific $tRNA^{Gln}$ with anticodons complementary to UAA and UAG: these unusual tRNAs arose by duplication and divergence of the canonical $tRNA^{Gln}$ [7]. *Euplotes octocarinatus* has only

one $tRNA^{Cys}$ that translates UGA efficiently, despite G:A mispairing at the first anticodon position [15]. This change requires loss of release factor specificity for UGA and/or high concentrations of $tRNA^{Cys}$ relative to other tRNAs.

Because all known ciliate code changes alter stop codon meanings, the eukaryotic release factor 1 (eRF1) must have evolved alternate specificities. In eukaryotes, eRF1 recognizes the three standard stop codons in mRNA at the ribosomal A site and terminates translation by peptidyl tRNA hydrolysis. Archaea have an eRF1 homolog, aRF1, which is highly conserved across domains, and aRF1 even functions with eukaryotic ribosomes [16].

Although eRF1 sequences from organisms with altered termination are of particular interest, only one eRF1 sequence was in GenBank for an organism with a nonstandard genetic code (the ciliate *Tetrahymena thermophila* [17]). However, this eRF1 sequence suggested a mechanism for the specificity change: the NIKS motif, conserved across all eukaryotes, is NIKD in *Tetrahymena*. Together with the recent crystal structure of human eRF1 [18] and mutational evidence that changes adjacent to NIKS abolish stop codon recognition in vitro [19], we suggested that this specific mutation might be the molecular cause of *Tetrahymena's* altered genetic code [20].

In order to test this hypothesis and to further examine the biochemical basis for altered stop codon recognition in ciliates, we sequenced the complete gene encoding eRF1 in three spirotrichs, *Stylonychia lemnae*, *S. mytilus*, and *Oxytricha trifallax*, that independently derived the same genetic code as *Tetrahymena*. For comparison, we also sequenced the gene in *Blepharisma americanum*, an early diverging ciliate that uses UAA as stop [8], and found evidence that this species uses UGA to encode tryptophan. Using a novel statistical approach that uses sliding window analysis to associate changes in specific regions of the protein with changes in the genetic code, as well as mapping some of the convergent amino acid substitutions in lineages that independently evolved the same eRF1 specificity onto the protein crystal structure, we identify several candidate amino acid residues or regions of the protein that may underlie the altered codon specificity of eRF1 and genetic code change in ciliates.

Results

Isolation of eRF1

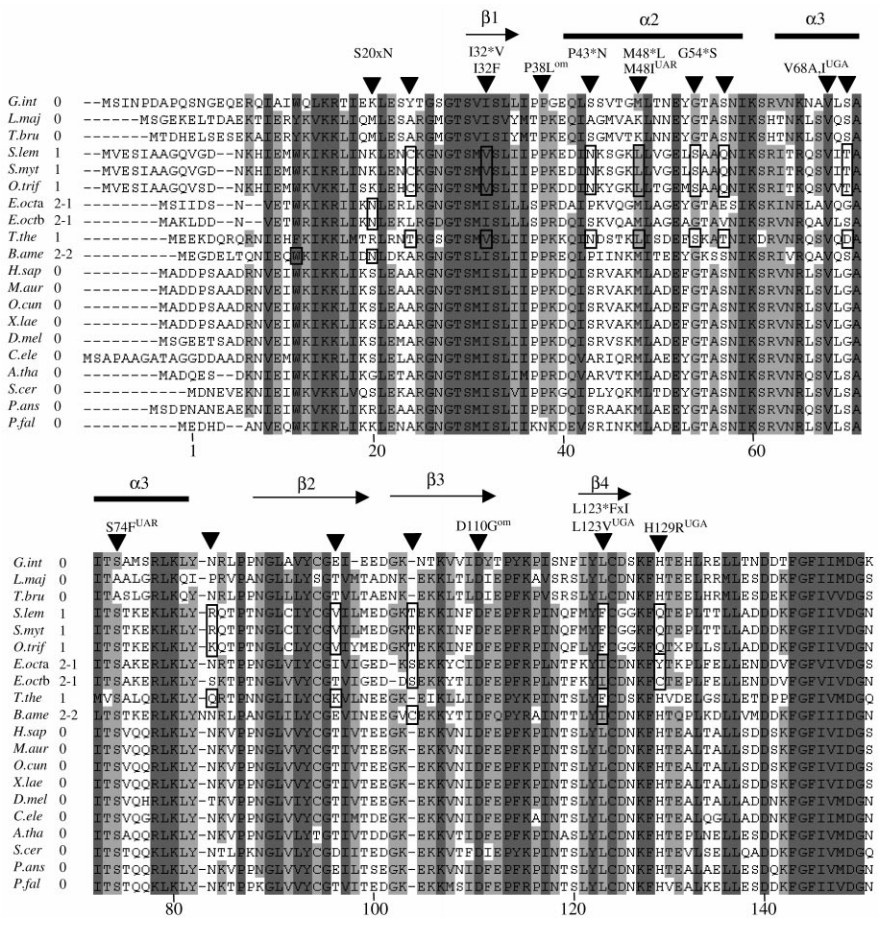
We determined the complete macronuclear sequence of *Stylonychia lemnae*, *S. mytilus*, and *Oxytricha trifallax* eRF1 gene-sized chromosomes, which are all predicted to encode proteins 445 amino acids long, as well as *Blepharisma americanum* eRF1, predicted to encode a 436 amino acid protein. *S. lemnae*, *S. mytilus*, and *O. trifallax* each appear to have a phase I intron, 32, 32, and 38 nucleotides long, respectively, at position 78 in the amino acid alignment (Figure 1); *B. americanum* has no intron at this position. The eRF1 gene has 2 in-frame UGA codons in *B. americanum* and numerous in-frame UAR codons in *S. lemnae*, *S. mytilus* and *O. trifallax*.

Genetic database analysis

Protein sequence data were available in the database for 7 of the 10 currently recognized [21] classes of ciliates and for a member of the order Amorphoridae classified as *sedis mutabilis* in the subphylum Intramacronucleata [21]. We analyzed, for the first time, data for members of the class Colpodea and for *Nyctotherus*, of the order Amorphoridea. In addition, we expanded analysis of the six previously studied classes [11] to include more species allowing better definition within these groups. Figure 2 is a composite tree assembled from the literature, using both 28S large subunit [11, 22] and 18S small subunit [23, 24, 25] congruent rDNA phylogenies, for the purpose of character mapping of genetic codes. It lists all of the genera for which information on genetic code usage is available, except for many of the Spirotrichs analyzed, which form a monophyletic group with *Stylonychia*, *Oxytricha*, and *Urostyla*, and all appear to use the same code. These data support that the classes Oligohymenophorea, Spirotrichea, and Litostomatea are each monophyletic. The heterotrichs and the karyorelictids form an early branching monophyletic group, but the relationship of the rest of the classes is largely unresolved. There is some support, however, that the classes Nassophorea, Colpodea, and Oligohymenophorea form a monophyletic group (Figure 2). Members of the classes Karyorelictea and Nassophorea are grouped together by morphological data only [26, 27] and are placed on the tree based on the rRNA sequences of *Zosterograptus* and *Loxodes*, respectively. Table 1 summarizes the database analysis and supplemental Table 1 provides more detail (see Supplementary material).

We found evidence in two distantly related lineages for UGA reassignment to tryptophan (Figure 1 and supplemental Figure 1). The gene for mitotic cyclin-like protein in *Colpoda inflata* contains an in-frame UGA in the position of a conserved tryptophan (supplemental Figure 1). Analysis of five partial protein sequences revealed no in-frame UAR codons in this class (Table 1). In the heterotrichs, members of the genera *Stentor* and *Eufolliculina* both had multiple in-frame UGA codons, and *Eufolliculina* and *Blepharisma* use UAA to encode stop (Table 1; [8]). Alignment of the *Stentor* and *Eufolliculina* proteins containing in-frame UGA codons with orthologs retrieved using BLAST did not suggest which amino acid UGA was coding for because the alignment was in variable regions of the proteins. However, the *B. americanum* eRF1 sequence generated in this study had two in-frame UGA codons, both in the location of conserved tryptophan residues (Figure 1 and supplemental Figure 1). Based on the close

Figure 1



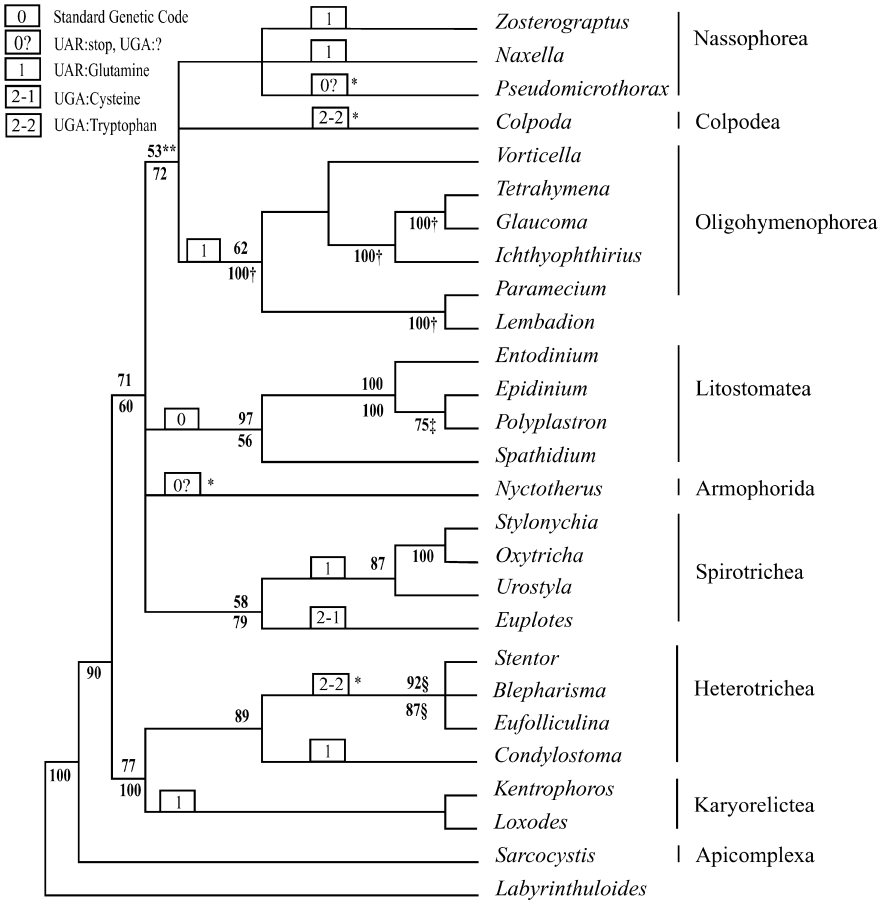
Alignment of the N terminus of all available eRF1 proteins. Areas of the alignment corresponding to structural features of domain 1 (α helices and β sheets) are indicated. Genetic code usage of each species is listed next to the abbreviated Latin name using the code notation described in Figure 2. Residues marked with an arrow are shown in Figure 3 and those that Bertram et al. [30] identified are indicated with their notation plus om (omnipotent suppression), UAR, or UGA to indicate the new recognition suppression of the mutant. Sites of interest are boxed and sites of convergent evolution between *Stylonychia/Oxytricha* and *Tetrahymena* (*) or *Euplotes* and *Blepharisma* (x) are annotated with the symbols (* or x) between the residues that mutated and the amino acid position in yeast (e.g., L123*F convergently changed to F at yeast position 123 in *Stylonychia/Oxytricha* and *Tetrahymena*). The boxed W at position 11 in the *Blepharisma* sequence is encoded as UGA (Supplemental Figure 1). Residues shaded in dark gray are highly conserved, variable residues are unshaded, and intermediate residues are light gray. Numbering according to the yeast sequence, *Saccharomyces cerevisiae* (*S.cer*, CAA51935) as in [30]. *G.int.*, *Giardia intestinalis* (AF198107 [41]); *L.maj*, *Leishmania major* (CAB77686); *T.bru*, *Trypanosoma brucei* (AAF86346); *S.lem*, *Stylonychia lemnae* (AF31784, this study); *S.myt*, *Stylonychia mytilus* (AF31783, this study); *O.tri*, *Oxytricha trifallax* (AF31782, this study); *E.octa*/b, *Euplotes octocarinatus* eRF1a and b [31]; *T.the*, *Tetrahymena thermophila* (P46055); *B.ame*, *Blepharisma americanum* (AF31781, this study); *H.sap*, *Homo sapiens* (P46055); *M.aur*, *Mesocricetus auratus* (X81626); *O.cun*, *Oryctolagus cuniculus* (AB029089); *X.lae*, *Xenopus laevis* (P35615); *D.mel*, *Drosophila melanogaster* (AAF51574); *C.ele*, *Caenorhabditis elegans* (T31907); *A.tha*, *Arabidopsis thaliana* (P35614); *P.ans*, *Podospora anserina* (AAC08410); and *P.fal*, *Plasmodium falciparum* (AAC71899).

phylogenetic relationship of *E. uhligi* and *S. coeruleus* to *B. americanum* [11, 28] and the presence of in-frame UGA codons in protein-coding genes for these species, it is likely that *E. uhligi* and *S. coeruleus* use UGA to encode tryptophan as well. This genetic code has been observed in mitochondrial genomes [2] and in the eubacterial *Mycoplasma* [3] but has never before been described for eukaryotic nuclear genomes. The heterotrichs are also an example of members of the same class using two different divergent codes: the early diverging heterotrich *Condylostoma magnum* has in-frame UAA and UAG codons at a position where conserved glutamines usually occur in orthologs [11].

Our results support the finding of Baroin Tourancheau et al. [11] that the reassignment of UAR to glutamine appears to have occurred multiple times independently in ciliates. Within the classes Oligohymenophorea and Spirotrichea, *Tetrahymena*, *Paramecium*, *Stylonychia*, *Oxytricha*, and *Glaucoma* all use this code [11]. We find evidence for use of UAR as glutamine in other oligohymenophorans, *Ichthyophthirius multifilis* and *Vorticella convallaria*, as well as many spirotrichs (Table 1). This shows that use of this code is common to members of these classes, except for the spirotrich *Euplotes*, which uses UGA to encode cysteine and UAR as stop [13]. The three other lineages that use UAR to encode glutamine are *Loxodes striatus* (Karyorelictea), *Zosterographus* (Nassophorea), and *Condylostoma* (Heterotrichea) [11]. We strengthen these findings by showing in-frame UAR codons in the karyorelictid *Kentrophoros sp.* and the nassophorean *Naxella sp.* (Table 1).

The class Nassophorea also displays use of different genetic codes within a ciliate class, although no molecular data are available on the monophyly of this class. While

Figure 2



Mapping of genetic code changes (boxes) mapped onto a composite phylogenetic tree of ciliate genera compiled from congruent published analyses of 28S ([11] bootstrap values indicated above the line in boldface; [22] marked **) and 18S ([23] bootstrap values below the line in bold; [25] marked †; [24] ‡) rRNA sequences. Bootstrap values marked § represent 28S and 18S studies, respectively, that grouped *Blepharisma* with *Stentor* 92% [11] and with *Eufolliculina* 87% [28]. A single asterisk indicates lineages with a code change described for the first time in this study.

in-frame UAR codons are present in *Zosterograptus sp*. and *Naxella sp*., three protein sequences in *Pseudomicrothorax dubius* have no in-frame stop codons and use UAG and UAA to terminate translation. The coding of UGA in this species was ambiguous. The sequence of the hydrogenase protein in *Nyctotherus ovalis*, of the order Amorphorida, has no in-frame stop codons and uses UAA to encode stop. Additional DNA sequences from these two species should be examined for evidence of UGA reassignment to an amino acid.

The class Litostomatea is the only group of ciliates for which strong evidence suggests use of the standard genetic code. None of the 55 protein sequences available for litostomes had in-frame stop codons, and translation was terminated with either UAA or UGA (Table 1 and supplemental Table 1).

Identification of convergent changes

The changes in ciliate genetic codes involve two types of altered stop codon recognition: recoding of either UGA or both UAA and UAG. For each of these two types of change, we have eRF1 sequences from representatives of two different lineages in which the change has occurred: the reassignment of UGA in *Euplotes* and *Blepharisma* and UAR in *Stylonychia/Oxytricha* and *Tetrahymena*. The phylogenetic analysis indicates that both of these changes probably occurred independently. For the UGA change, it is strongly supported that *Euplotes* is more closely related to many genera that use UAR to encode glutamine, such as *Stylonychia*, *Oxytricha*, and *Tetrahymena*, than it is to *Blepharisma* (Figure 2). *Blepharisma*, likewise, is closely related to *Condylostoma magnum*, which also uses UAR to encode glutamine [11]. There is weaker evidence that the change of UAR to glutamine occurred independently in the spirotrichs that use this code and the oligohymenophorans: *Euplotes*, which uses UGA to encode cysteine, appears to have diverged early from a monophyletic lineage that includes the other spirotrichs, and *Colpoda*, in which we find evidence of UGA use to encode tryptophan, appears to form a monophyletic group with the oligohymenophorans (Figure 2). Both of these relationships are

Table 1

Organism	#	3′	aa	UAG	UAA	UGA	stop	Code
Nassophorea	6							0?,1
Naxella	1	0	408	3	2	0	(UGA)	1
Zosterograptus	1	0	382	0	6	0		1
Pseudomicrothorax	3	3	1770	0	0	0	UAA,UAG	0?
Colpodea	4							2-2
Colpoda	4	0	1253	0	0	1		2-2
Oligohymenophorea	6+							1
Vorticella	1	1	181	1	0	0	UGA	1
Glaucoma	2	0	449	3	1	0	(UGA)	1
Ichthyophthirius	2	1	517	1	23	0	UGA	1
Lembadion	1	1	351	0	6	0	UGA	1
Litostomatea	55							0
Entodinium	47	45	9447	0	0	0	UAA,UGA	0
Epidinium	2	1	737	0	0	0	UAA	0
Polyplastron	3	3	930	0	0	0	UAA	0
Spathidium	3	0	1194	0	0	0		0
Armophorida	2							0?
Nyctotherus	2	1	1742	0	0	0	UAA	0?
Spirotrichea	14+							1, 2-1
Pleurotricha	1	0	798	5	31	0	(UGA)	1
Paraurostyla	1	0	966	10	24	0	(UGA)	1
Uroleptus	1	0	973	10	26	0	(UGA)	1
Urostyla	4	2	2606	15	33	0	UGA	1
Histriculus	2	1	755	0	1	0	UGA	1
Halteria	3	1	1697	7	5	0	UGA	1
Holosticha	1	0	966	11	19	0	(UGA)	1
Hypotrichida	1	0	373	0	0	0		
Heterotrichea	32							1, 2-2
Blepharisma	10	2	689	0	0	2	UAA	2-2
Eufolliculina	12	12	3250	0	0	4	UAA	2-2?
Stentor	8	0	1841	0	0	3		2-2?
Condylostoma	1	0	380	1	4	0	(UGA)	1
Karyorelictea	2							1
Kentrophoros	1	0	408	1	0	0		1
Loxodes	1	0	380	6	0	0		1

A summary of ciliate class/genera, total number of protein sequences available; number of sequences that were complete at the 3′ end allowing determination of stop codon usage; total number of amino acids (aa) analyzed; number of in-frame UAA, UAG, and UGA codons, the codon(s) used to terminate translation where available (codons in parentheses are inferred from the data); and the genetic code used: (0) standard genetic code, (0?) UAA and UAG used as stop and translation of UGA is unknown (1) UAR:glutamine, (2) UGA:amino acid, (2-1) UGA:cysteine, (2-2) UGA:tryptophan.

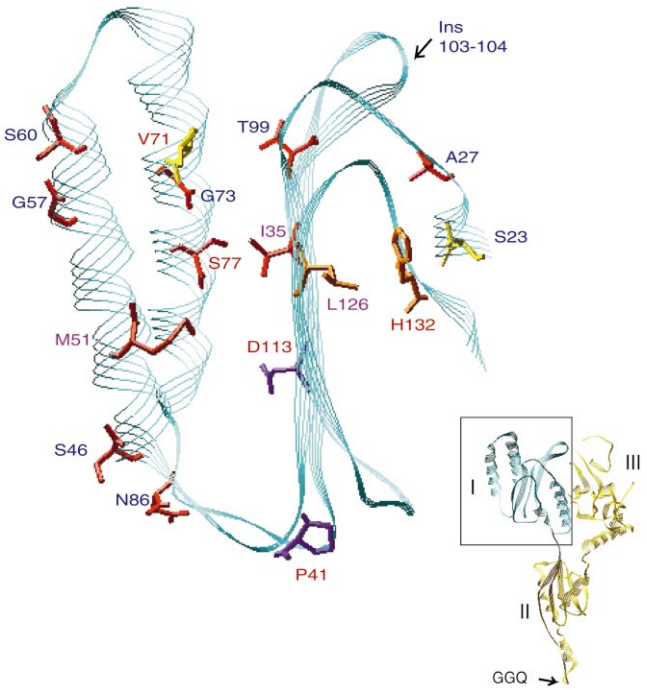
supported in numerous 18S and 28S rRNA trees, generally with greater bootstrap support for studies based on complete 18S small subunit (e.g., [23, 24, 28]) than partial 28S large subunit [11, 22] sequences. Morphological data also support the classification of *Euplotes* as a spirotrich [29]. In addition, phylogenetic analysis of α-tubulin gene sequences weakly supports both of these relationships, and an analysis of phosphoglycerate kinase gene sequences strongly supports the branching of *Euplotes* with *Oxytricha* (97% bootstrap support), though data for *Colpoda* were unavailable [22].

If these phylogenetic inferences are correct, one could still argue that an ancient change of UAR to glutamine in a common ancestor of the spirotrichs and oligohymenophorans might have been followed by UGA reassignment in *Euplotes* and *Colpoda*. It is much more likely, however, that the transition to these alternate genetic codes proceeded from a standard code background. The evolution of the *Euplotes* code from an ancestor that used UAR to encode glutamine, for instance, would potentially require loss of extra glutamine tRNAs, eRF1 mutations that decrease UAR readthrough and increase UGA readthrough, and mutations enabling a $tRNA^{Cys}$ to read UGA. For these reasons, we propose that the UAR-to-glutamine change most likely emerged independently in the spirotrich and oligohymenophoran lineages. We next ask whether there are convergent changes that could confer the altered specificity for both UAR and UGA reassignment. In other words, are there states of particular sites in eRF1 such that all and only ciliates with the same altered translation termination share those particular states? This analysis is itself independent of the direction of change inferred in Figure 2.

We wrote a C program (available from the authors) to scan a set of sequences labeled according to arbitrary criteria (in this case, the type of stop codons used by the species) and to partition amino acid identities at each site of an aligned set of sequences according to which labels

have which sites. We find that the following states are unique to particular translation states (numbered according to yeast eRF1; amino acids to the left of a number occur in organisms with the standard code and the amino acids to the right of the number occur in organisms with altered stop codon specificity): UAR = Stop: (KRGMS)20N, (.)103–104(CS), **L123I**, (NGT)213(DES), N262(ST), (AQ)279(ER), (DEIMQ)366(GV); UGA = Stop: (AY)24(CT), **I32V**, (PSA)43N, **(MK)48L**, G54S, (ES)57(QT), (GS)70(DT), (NPT)83KQR, (ETD)96(KV), (.)103–104T, **L123F**, (SAGKQ)273T, (NDG)322E, (.NDGQ)332(.K), (DECQS)343(NT). Remarkably, 14 of these 22 convergent changes map to domain 1, the putative codon recognition domain [18], and 4 of these changes in 3 different positions (in boldface) are the same changes recently identified by mutational screens for enhanced termination suppression in yeast [30]! Our alignment spanned 417 residues: the probability that we would hit 3 or more of the same residues that the mutational study identified by chance, if the changes were independent, is 0.014. (The mutational study identified 10 of 417 residues as being functionally important. Of these, 9 were in the region covered by our alignment. Since we identified 20 sites as having states unique to particular genetic codes, the 2 × 2 contingency table has counts of 3, 6, 17, and 319 for identification in both studies, in [30], our study, and neither study respectively. The probability that these results are independent is very low [G test for independence applying the Williams correction: df = 1, G = 4.82, $P_{1\text{-tailed}} = 0.014$]. Thus, our identification of several of the same sites is unlikely to be a coincidence.)

Figure 3



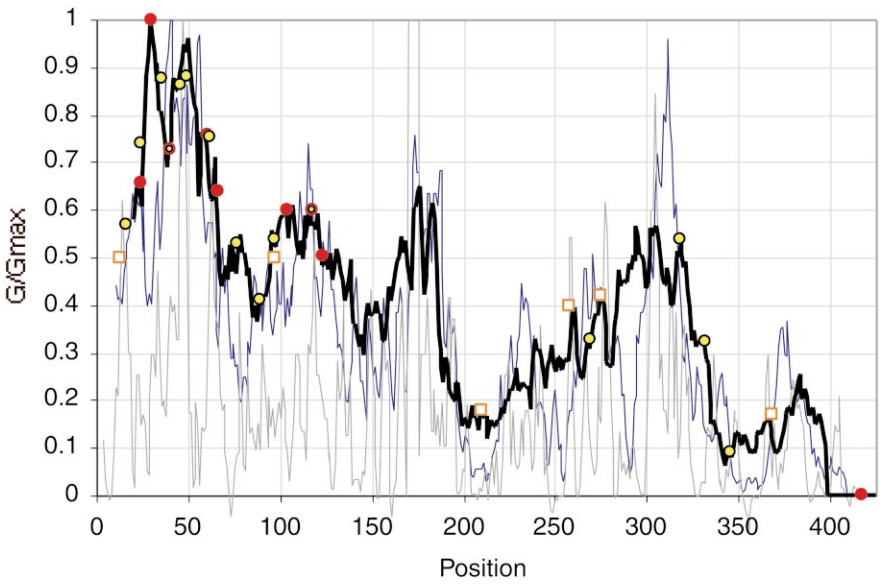
Human eRF1 structure of domain 1. The entire protein ribbon structure of human eRF1 (inset) is shown with domain 1 boxed in blue and domains 2 and 3 in yellow (PDB accession code 1DT9 [18]). Highlighted residues in domain 1 have been identified in this study (blue text), mutational analysis in yeast (red text [30]), or both (purple text). Numbers correspond to human eRF1, +3 from yeast eRF1 and Figure 1 (e.g., human M51 is homologous to yeast M48). Mutations in red residues are associated with UAR suppression; yellow, UGA suppression; and orange, both.

Bertram et al. [30] identified ten changes in domain 1: three changes in mutants exhibiting UGA suppression (V68I, L123V, and H129R) increased UGA readthrough, and three changes in mutants with UAG suppression (M48I, S74F, D110G) increased UAR readthrough. I32F was a potential UGA suppressor and the remaining three isolates were omnipotent suppressors. Our evolutionary approach agrees surprisingly well with these mutational data: the UGA suppressors *Euplotes* and *Blepharisma* have a convergent change from L to I in the homologous position of L123V, and *Euplotes* [31] has a Y or C at the position of H129R, a conserved histidine in all taxa except ciliates with alternate codes (Figure 1). V68I is at a conserved valine across all ciliates, indicating that this site either has not been explored by evolution or is fixed by other constraints. L123V is also a site of a convergent change from L to F in *Stylonychia/Oxytricha* and *Tetrahymena* and is conserved in all other taxa; *Stylonychia* and *Oxytricha* have a nonconservative change from H to Q at site H129R. The homologous residues L126 and H132 in human eRF1 flank a hydrophobic pocket that may recognize the stop codon second position in the model in [30] (Figure 3). As this position distinguishes UAR from UGA, we propose that these changes, in particular the convergent changes at position 123 (Figure 1), play a critical role in the evolution of release factor specificity in ciliates. There is only one additional convergent change in domain 1 of *Euplotes* and *Blepharisma*. In contrast, we identified ten sites in domain 1 that changed in both *Stylonychia/Oxytricha* and *Tetrahymena*. Of these, seven are in α helices 2 and 3 (Figures 1 and 3) and five mutate to the same residue. Two convergent changes, I to V and M to L, are at the same site as I32F and M48I UAR suppression mutations [30], and site 32 is conserved across all taxa but the ones using this code. An important next step would be to engineer these changes into yeast eRF1 and test whether the resulting protein could rescue a *Tetrahymena* eRF1 knockout.

Sliding window analysis: associating eRF1 changes and termination changes

The analysis of individual residues is sometimes difficult to interpret, because apparent associations between change

Figure 4



Sliding window analysis of eRF1 and genetic code change. This figure shows the robustness of peaks to changes in window size, showing windows of size 5 (gray line), 20 (blue line), and 40 (black line). Peaks indicate regions in which changes are maximally associated with genetic code change. The x axis is position in our alignment (Figure 1); the y axis is relative G value (scaled relative to maximum G value for that window size). The first two peaks lie in domain 1, the third in domain 2 near the GGQ motif, and the fourth in domain 3, which binds eRF3 [18]. Red dots mark mutations in yeast eRF1 that affect stop codon readthrough (marked on the 40 residue line); UAR→Gln, yellow dots; UGA ≠ stop, orange boxes. An interactive Excel file that allows the user to view the results with different window sizes is available at ftp://rnaworld.princeton.edu/pub/export/windowresults2.xls

at a single site and change in the genetic code could occur by chance. By pooling counts for a contiguous stretch of sequence, it is possible to test for association between changes in that region and changes in the genetic code. This increases confidence that particular regions are important, at the expense of making inferences about individual sites.

Figure 4 shows results for three window sizes (5, 20, and 40 residues) along with mutations found to affect decoding in yeast eRF1 [30] and those identified in this study. There are four peaks that are stable across a broad range of window sizes. The first two of these peaks correspond to domain 1, and the third one corresponds to the region immediately before the universal GGQ motif. Additionally, the decoding mutants in yeast and the convergent changes identified in this study cluster around the first two peaks.

Discussion

Our database analysis extends both the number of independent code changes in ciliates and the types of code change reported for eukaryotes. Both UAR and UGA have been reassigned independently many times in different lineages of ciliates. Whether ciliates are particularly prone to changes in the genetic code or if they are they just more diverse than other microbial eukaryotes remains open. UGA has been reassigned to both cysteine and tryptophan, showing that either of two tRNAs can expand its decoding ability to cover this codon. The UGA → tryptophan change also occurs in bacteria and mitochondria. Is UGA particularly prone to reassignment?

We observed many fewer convergent eRF1 substitutions in *Euplotes* and *Blepharisma*, which suggests either that independent substitutions may assist stop codon reassignment in these lineages or, more intriguingly, that UGA suppression may be easier to effect than UAR suppression, a conclusion that is also supported by the fact that Bertram et al.'s initial screen for unipotent suppressor phenotypes only identified UGA suppressors [30]. The isolation of UAR suppressors required a plasmid containing a mutant $tRNA^{Ser}$ with weak UAR nonsense suppressor activity.

We initially set out to test the hypothesis that a change of S to D in *Tetrahymena*, within the otherwise universally conserved NIKS motif of domain 1, was responsible for altering the specificity of eRF1 in organisms with this code. We found that *Stylonychia* and *Oxytricha*, which also translate UAR as glutamine, do not have any deviations in the NIKS motif; nor does *Blepharisma*. One of the two release factor genes in *Euplotes octocarinatus*, however, has the sequence SIKS in this region [31]. This indicates that changes in this motif are not necessary for altered stop codon recognition, as we initially proposed, although they may be sufficient. However, several different changes adjacent to this region are heavily implicated in variant codes. Construction of specific mutants in yeast eRF1 will allow finer mapping of the critical residues and perhaps indicate how many ways eRF1 can mutate to generate the same changes in the genetic code.

Changes in ciliate termination may depend on less than 6 residue changes (for loss of UGA recognition) or less than 15 residue changes (for loss of UAR recognition) in eRF1 on the basis of convergent mutations in lineages that evolved the same changes independently. Despite

potentially requiring more mutations in eRF1, UAR has changed to glutamine in at least two other completely independent groups (diplomonads and algae [5, 32]); eRF1 sequences from these species may reveal whether this change in the genetic code has a unique convergent molecular basis.

Most changes in the genetic code involve termination: this may be because stop codons are rare, occurring only once per gene, and so changes in termination are likely to be less deleterious than change in sense codons. This would be particularly true for those species of ciliates whose genes reside on gene-sized chromosomes and/or have short 3′ untranslated regions. In addition, termination is a competition for stop-codon-containing ribosomal A sites between release factors and tRNAs. Consequently, relatively small changes either in the tRNAs or in eRF1 may shift this balance toward partial or complete readthrough in some cases. For instance, *Bacillus subtilis* uses in-frame UGA codons extensively to encode tryptophan; however, this readthrough is inefficient, and UGA is also used as a stop codon [33, 34]. The abundance of stop codon reassignments relative to amino acid codon reassignment, however, could also be an observer bias. In-frame stop codons are much easier to detect in protein coding sequences than amino acid replacements, especially if the latter have similar properties.

Is there any pattern to the identities of the amino acids to which the stop codons are reassigned? The reassignment of UAR to glutamine can be explained either by a transition mutation at the third anticodon position of $tRNA^{Gln}$ (which normally recognizes CAA and/or CAG), alteration in the tRNA elsewhere to enhance G-U mispairing at the first codon position, or both. The second mechanism implies a period of ambiguous translation; interestingly, *Tetrahymena* $tRNA^{Gln}$ contain specific changes distant from the anticodon that increase G-U wobble when introduced into the equivalent tRNA in yeast [35]. Similarly, reassignment of UGA to cysteine and tryptophan can be explained in terms of expansion of wobble in existing tRNAs [35].

We identified convergent changes in three of the exact same sites in eRF1 independently identified by mutational screens in yeast [30]. Additionally, the yeast mutants cluster heavily around the first two major peaks we identified in domain 1 by sliding window analysis. This approach also identified a sharp peak at about position 175, immediately adjacent to the GGQ motif. This motif mimics the tRNA CCA acceptor stem and probably interacts with the peptidyl transferase center of the ribosome to release the nascent peptide [18]. In a previous study, Karamyshev et al. [17] found that *Tetrahymena* eRF1 does not work with yeast ribosomes: this was surprising, since even the distantly related aRF1 from the archaeon *Methanococcus jannaschii* is active with eukaryotic ribosomes [16], and suggests that these structures in *Tetrahymena* have coevolved.

The sliding window analysis proves surprisingly powerful in identifying the sites already known to be particularly important in eRF1 function: domain 1 and the GGQ motif. This type of analysis should be applicable to any character that has changed multiple times independently in different taxa: possible examples include changes in signaling pathways, recruitment of paralogs to new functions, and changes in pathogenicity or host range. Identification of molecular correlates of evolutionary change in this manner may greatly assist corresponding biochemical analyses.

These results show the power of comparative evolutionary techniques in identifying important sites in functional molecules, as Woese et al. showed decades earlier with rRNA [36]. Finding several species that differ naturally in some respect, and determining molecular correlates of this change, may yield the same results as screening for mutants in a single species. This approach may even be more powerful, as change-of-function mutants are often lethal and may not be picked up efficiently in a selection. We predict that many of the changes that we identify above, when introduced into yeast eRF1, will be lethal because of efficient stop codon readthrough.

Conclusions

We have shown that genetic code change in ciliates is even more pervasive than previously thought, with multiple independent lineages evolving changes in stop codon recognition requiring changes in release factor eRF1. We report the first use of UGA as tryptophan in a eukaryotic nuclear genome, in two distinct lineages; therefore, even this change has occurred more than once in ciliates. We detect striking instances of convergent evolution in the amino acid sequence of eRF1, with distantly related lineages with variant codes displaying characteristic substitutions in eRF1, implying that some of these substitutions may drive evolution of the genetic code. The strong agreement between our results and nonsense suppression screens in yeast [30] strengthens our conclusions, as well as the robustness of the statistical analysis. Thus, the combination of unique residue identification and sliding window analysis provides a powerful approach to detecting the molecular determinants of convergent evolutionary change.

Materials and methods

Release factor sequencing and analysis

Macronuclear DNA from *Stylonychia lemnae*, *S. mytilus*, and *Oxytricha trifallax* were a gift from David Prescott (University of Colorado, Boulder). DNA was extracted from *Blepharisma americanum* (Culture Collection of Algae and Protozoa, CCAP 1607/1) using a DNeasy Tissue Kit (QIAGEN). We then aligned all eRF1 sequences available from GenBank and designed degenerate primers flanking the codon recognition domain of eRF1. The forward primer is eRF1100F, ATG(AG)T(TA)(TA)C

(ATC)TT(GA)(GA)T(TC)AT(TC)CC(TA)CC and the reverse primer is eRF1799R, AT(KT)GC(TC)TGKTTKAA(AT)CCKTTNTC(AT)CC(AT)CC (AT)CCKTA. K is a nonstandard base that binds both C and T (Glen Research). We amplified the region encoding the eRF1 protein fragment using 40 cycles of PCR (94°C, 25 s; 50°C, 20 s; 72°C, 1 min; final extension , 72°C, 10 min). We then amplified the 5′ and 3′ ends of the macronuclear eRF1 gene for *Stylonychia lemnae*, *S. mytilus*, and *Oxytricha trifallax* using gene specific primers SORF415F [TATTTTTGC GGTGGTAA(AG)TTCCAGACTGA], SORF680R [(AT)AGTCTCTTAT CGAGCAT(GA)TCAGTCTCA], SORF581F [ACAGAAAGGGAGGT CA(AG)TC(AT)TCAGTCAG], and SORF604R [CTGA(TA)GA(CT)TG ACCTCCCTTTCTGTGCTT] and telomere-specific primers in telomere suppression PCR as described elsewhere [37].

We recovered the 5′ and 3′ ends of the *Blepharisma americanum* eRF1 cDNA using anchor PCR primers and protocols as described in [38] with total RNA extracted using TRIzol reagent (GIBCO–BRL), Superscript II reverse transcriptase (GIBCO–BRL), and primers specific for *B. americanum* eRF1: BGSP1F (TGCCAGGCTGCGTATGGAGTCC), BGSP 2F (CATTGCTGGGTCAGCTGAGTTC), BGSPRTR (AAATCAGACC GTTTGCTGGGAG), BGSP1R (CGCTCTTTTGTGGAGGTAAGAG), and BGSP2R (ACAGCCTGCCTGACGATTCTTG).

All PCR products were cloned using a TOPO TA Cloning Kit (Invitrogen) and multiple clones for each fragment were sequenced at the Syn/Seq Facility of Princeton University. The complete *B. americanum* sequence was confirmed by PCR amplification and sequencing from total DNA. Sequences have been deposited in GenBank (AF31781–AF31784).

We aligned sequences with all known eRF1 homologs using Pileup in Seqlab (Wisconsin GCG version 10.1) and identified nonconservative mutations in otherwise conserved regions of the codon recognition domain in species with nonstandard genetic codes. Mutations common to all and only ciliates with one of the two specific types of change in termination (using either UGA only or UAR only) were of particular interest here.

Additionally, we tested whether changes in specific areas of the eRF1 protein were unusually associated with changes in the genetic code. This relied on the phylogeny, with its inferred ancestral states. We constructed a maximum parsimony phylogeny using the Paupsearch function in Seqlab (Wisconsin GCG version 10.1) and a heuristic tree search using the default settings. We excluded from this analysis the 5′ and 3′ ends of the sequences, as well as 2 residues for which the amino acid identity of *Blepharisma americanum* was ambiguous and a 13 residue region in which a 9 residue insert in one of the sequences left the alignment unresolved. We consistently observed the artificial grouping of *Stylonychia* and *Oxytricha* (class Spirotrichea) with *Tetrahymena* (class Oligohymenophorea) instead of *Euplotes* (class Spirotrichea), but exclusion of five amino acid positions in domain 1, in which there was convergent evolution between *Stylonychia*/*Oxytricha*, and *Tetrahymena*, restored the monophyly of the spirotrichs. These residues were added back to the alignment after performing phylogenetic analysis, and the ancestral states of these residues were inferred based on parsimony [39].

For every edge linking two nodes in the phylogeny, we inferred whether a change in the code had taken place and, for each position in the alignment, whether the amino acid at that position had changed. This gave a table of four counts for each position, according to whether or not changes (in code and sequence) had occurred. It was thus possible to test for independence between code changes and changes at each position in eRF1; if change at a particular position caused changes in the code, the two variables would not be independent. We used the G test for independence [40].

Because there are relatively few sequences, individual residues typically did not give significant associations. In order to identify regions of sequence associated with genetic code change, we performed sliding window analysis, combining counts from contiguous regions of size n ($1 \leq n \leq 100$) to identify regions in which changes were consistently implicated with changes in the genetic code. This analysis was performed using custom programs written in C and Microsoft Excel–hosted VBA.

Genetic database analysis

We analyzed all ciliate protein coding sequences available in genetic databases, using the Seqlab editor of the GCG Wisconsin Package (version 10.1). We excluded sequences from *Tetrahymena*, *Paramecium*, *Stylonychia*, *Oxytricha*, and *Euplotes*, since the genetic code of these genera has been well characterized. We translated sequence data using the standard genetic code and recorded the number of in-frame stop codons and the type of codon actually used for terminating translation in each gene. Wherever possible, we aligned protein sequences displaying in-frame stop codons with orthologs from related species, allowing us to infer the new meaning of the former stop codon. Using congruent published molecular phylogenies based on 18S small subunit [23–25, 28] and 28S large subunit rRNA [11, 22] sequences (Figure 2), we then made a composite phylogenetic tree of the ciliates with known genetic code usage. Molecular data were available to assess the relationships of all groups except the nassophoreans and the karyorelictids, which were grouped according to morphological characters [26, 27].

Supplementary material

Supplementary material including partial eRF1 and cyclin sequence alignments showing evidence of UGA usage at conserved tryptophans and a table recording in-frame stop codon occurrence for the data summarized in Table 1, is available at http://current-biology.com/supmat/supmatin.htm.

Acknowledgements

We gratefully acknowledge David Prescott for the gift of *Stylonychia* and *Oxytricha* DNA and Klaus Heckmann for sharing the *Euplotes octocarinatus* sequences before publication. We also thank Stephen Freeland, Christina Burch, and David Ardell for discussion.

References

1. Crick FH: **The origin of the genetic code.** *J Mol Biol*1968, **38:**367-379.
2. Barrell BG, Bankier AT, Drouin J: **A different genetic code in human mitochondria**. *Nature* 1979, **282:**189-194.
3. Yamao F, Muto A, Kawauchi Y, Iwami M, Iwagami S, Azumi Y, *et al.*: **UGA is read as tryptophan in *Mycoplasma capricolum*.** *Proc Natl Acad Sci USA* 1985, **82:**2306-2309.
4. Santos MAS, Tuite MF: **The CUG codon is decoded in vivo as serine and not leucine in *Candida albicans*.** *Nucleic Acids Res* 1995, **23:**1481-1486.
5. Keeling PJ, Doolittle WF: **A non-canonical genetic code in an early diverging eukaryotic lineage.** *EMBO J* 1996, **15:**2285-2290.
6. Horowitz S, Gorovsky MA: **An unusual genetic code in nuclear genes of *Tetrahymena*.** *Proc Natl Acad Sci USA* 1985, **82:**2452-2455.
7. Hanyu N, Kuchino Y, Susumu N, Beier H: **Dramatic events in ciliate evolution: alteration of UAA and UAG termination codons to glutamine codons due to anticodon mutations in two *Tetrahymena* tRNAs^{Gln}.** *EMBO J* 1986, **5:**1307-1311.
8. Liang A, Heckmann K: ***Blepharisma* uses UAA as a termination codon.** *Naturwissenschaften* 1993, **80:**225-226.
9. Osawa S, Jukes TH, Watanabe K, Muto A: **Recent evidence for evolution of the genetic code.** *Microbiol Rev* 1992, **56:**229-264.
10. Knight RD, Freeland SJ, Landweber LF: **Rewiring the keyboard: evolvability of the genetic code.** *Nat Rev Genet* 2001, **2:**49-58.
11. Baroin Tourancheau A, Tsao N, Klobutcher LA, Pearlman RE, Adoutte A: **Genetic code deviations in the ciliates: evidence for multiple and independent events.** *EMBO J* 1995, **14:**3262-3267.
12. Preer JR Jr, Preer LB, Rudman BM, Barnett AJ: **Deviations from the universal code shown by the gene for surface protein 51A in** *Paramecium*. *Nature* 1985, **314:**188-190.
13. Meyer F, Schmidt HJ, Plumper E, Hasilik A, Mersmann G, Meyer HE, *et al.*: **UGA is translated as cysteine in pheromone 3 of *Euplotes octocarinatus*.** *Proc Natl Acad Sci USA* 1991, **88:**3758-3761.
14. Caron F: **Eucaryotic codes.** *Experientia* 1990, **46:**1106-1117.

15. Grimm M, Brunen-Nieweler C, Junker V, Heckmann K, Beier H: **The hypotrichous ciliate *Euplotes octocarinatus* has only one type of tRNA^{Cys} with GCA anticodon encoded on a single macronuclear DNA molecule.** *Nucleic Acids Res* 1998, **26:**4557-4565.
16. Dontsova M, Frolova L, Vassilieva J, Piendl W, Kisselev L, Garber M: **Translation termination factor aRF1 from the archaeon Methanococcus jannaschii is active with eukaryotic ribosomes.** *FEBS Lett* 2000, **472:**213-216.
17. Karamyshev AL, Ito K, Nakamura Y: **Polypeptide release factor eRF1 from Tetrahymena thermophila: cDNA cloning, purification and complex formation with yeast eRF3.** *FEBS Lett* 1999, **457:**483-488.
18. Song H, Mugnier P, Das AK, Webb HM, Evans DR, Tuite, MF, Hemmings BA, Barford D: **The crystal structure of human eukaryotic release factor eRF1 – Mechanism of stop codon recognition and peptidyl-tRNA hydrolysis.** *Cell* 2000, **100:**311-321.
19. Mironova LN, Zelenaia OA, Ter-Avanesian MD: **Nuclear-mitochondrial interactions in yeasts: mitochondrial mutations compensating the respiration deficiency of sup1 and sup2 mutants.** *Genetika* 1986, **22:**200-208.
20. Knight RD, Landweber LF: **The early evolution of the genetic code.** *Cell* 2000, **101:**569-572.
21. Lynn DH, Small EB: **A revised classification of the Phylum Ciliophora Doflein, 1901.** *Rev Soc Mex Hist Nat* 1997, **47:**65-78.
22. Baroin Tourancheau A, Villalobo E, Tsau N, Torres A, Pearlman RE: **Protein coding gene trees in ciliates: comparison with rRNA-based phylogenies.** *Mol Phylogenet Evol* 1998, **10:**299-309.
23. Stechmann A, Schlegel M, Lynn DH: **Phylogenetic relationships between Prostome and Colpodean ciliates tested by small subunit rRNA sequences.** *Mol Phylogenet Evol* 1998, **9:**48-54.
24. Wright AG, Dehority BA, Lynn DH: **Phylogeny of the rumen ciliates *Entodinium, Epidinium* and *Polyplastron* (Litostomatea: Entodiniomorphida) inferred from small subunit ribosomal RNA sequences.** *J Eukaryot Microbiol* 1997, **44:**61-67.
25. Strüder-Kypke MC, Wright AG, Fokin SI, Lynn D: **Phylogenetic relationships of the subclass Peniculia (Oligohymenophorea, Ciliophora) inferred from small subunit rRNA gene sequences.** *J Eukaryot Microbiol* 2000, **47:**419-429.
26. Small EB, Lynn DH: **A new macrosystem for the Phylum Ciliophora Doflein, 1901.** *Biosystems* 1981, **14:**387-401.
27. Corliss JO: *The Ciliated Protozoa. Characterization, Classification, and Guide to the Literature.* London: Pergamon Press; 1979.
28. Hammerschmidt B, Schlegel M, Lynn DH, Leipe DD, Sogin ML, Raikov IB: **Insights into the evolution of nuclear dualism in the ciliates revealed by phylogenetic analysis of rRNA sequences.** *J Eukaryot Microbiol* 1996, **43:**225-230.
29. Lynn DH, Corliss JO: **Ciliophora**. In *Microscopic Anatomy of Invertebrates. Vol. 1: Protozoa.* Edited by Corliss JO, Harrison FW. New York: Wiley-Liss, Inc.; 1991:333-467.
30. Bertram G, Bell HA, Ritchie DW, Fullerton G, Stansfield I: **Terminating eukaryotic translation: domain 1 of release factor eRF1 functions in stop codon recognition.** *RNA* 2000, **6:**1236-1247.
31. Liang A, Brünen-Nieweler C, Muramatsu T, Kuchino Y, Beier H, Heckmann K: **The ciliate *Euplotes octocarinatus* expresses two polypeptide release factors of the type eRF1**. *Gene* 2001, **262:**161-168.
32. Schneider SU, Leible MB, Yang XP: **Strong homology between the small subunit of ribulose-1,5-bisphosphate carboxylase/oxygenase of two species of Acetabularia and the occurrence of unusual codon usage.** *Mol Gen Genet* 1989, **218:**445-452.
33. Lovett PS, Ambulos NP Jr, Mulbry W, Noguchi N, Rogers, EJ: **UGA can be decoded as tryptophan at low efficiency in *Bacillus subtilis*.** *J Bacteriol* 1991, **173:**1810-1812.
34. Matsugi J, Murao K, Ishikura H: **Effect of *B. subtilis* tRNA(Trp) on readthrough rate at an opal UGA codon.** *J Biochem* 1998, **123:**853-858.
35. Schultz DW, Yarus M: **Transfer RNA mutation and the malleability of the genetic code.** *J Mol Biol* 1994, **235:**1377-1380.
36. Woese CR, Fox GE, Zablen L, Uchida T, Bonen L, Pechman K, *et al.*: **Conservation of primary structure in 16S ribosomal RNA.** *Nature* 1975, **254:**83-86.
37. Curtis EA, Landweber LF: **Evolution of gene scrambling in ciliate micronuclear genes.** *Ann NY Acad Sci* 1999, **870:**349-350.
38. Horton TL, Landweber LF: **Evolution of four types of RNA editing in myxomycetes.** *RNA* 2000, **6:**1339-1346.
39. Harvey PH, Pagel MD: *The Comparative Method in Evolutionary Biology*. Oxford: Oxford University Press; 1991.
40. Sokal RR, Rohlf FJ: *Biometry: The Principles and Practice of Statistics in Biological Research.* New York: W. H. Freeman and Company; 1995.
41. Inagaki Y, Doolittle WF: **Evolution of the eukaryotic translation termination system: origins of release factors.** *Mol Biol Evol* 2000, **17:**882-889.

3.4 A Simple Model Based On Mutation and Selection Explains Compositional Trends Within and Across Genomes

The motivation for this section was to test the Osawa-Jukes model of codon reassignment: I knew of examples of apparently arbitrary codon usage in a range of species, and doubted that GC content would have a strong and consistent effect on codon usage (and certainly did not believe that biased mutation could be sufficient to eliminate a codon from a complete genome). Instead, I discovered that the GC content explains most of the variance in relative usage of every codon and every amino acid in every species (at least on average). Apparently, finding the probability of occurrence of a particular codon is as simple as multiplying the base probabilities together (after allowing for the fact that the different codon positions react to mutation pressure at different rates, probably because of selection), although individual organisms often have unexplained preferences for one or a few codons. The idiosyncratic codon usages of a few model species have thus contributed to a vast literature with thousands of pages of explanations of 'preferences' that disappear when considering species other than Drosophila, *yeast, and* E. coli. *Although it may still be the case that some of these explanations hold for individual species, and may explain deviations from the general trends, they are unable to explain the trends themselves.*

Since the model predicts a specific codon usage for a given GC content, but a given GC content could be attained by a vast number of combinations of codons (e.g. the necessary ratio of just CCC and UUU), this work strongly implies that biased mutation and purifying selection explain most of the variance in codon and amino acid usage across different species. However, it is still possible that GC content is a selected trait, and that codon and amino acid usages are epiphenomenona of it. Interestingly, the model only works for nuclear lineages and non-animal mitochondria, because metazoan mitochondria encode almost all their genes on one strand and so asymmetries in mutation between the leading and lagging strand in replication cause systematic violations in Chargaff's Rule that A=T and C=G (see Chapter 3.3, especially Fig. 2a and 2b for correlations between nucleotide frequency in bacteria and mitochondria). It is possible that the different selection pressures to which metazoan mitochondria are subject, such as coding only a few genes and being highly optimized for genome replication, have influenced the evolution of their genomes uniquely.

This chapter has been published as:

Knight, R. D., S. J. Freeland, and L. F. Landweber. (2001) "A simple model based on mutation and selection explains compositional trends within and across genomes." GenomeBiology 2:4, http://www.genomebiology.com/2001/2/4/research/0010/.

I developed the model and statistical framework, and performed the analyses. Dr. Freeland revised several sections extensively, and suggested the historical link to classical neutral models in population genetics.

Research

A simple model based on mutation and selection explains trends in codon and amino-acid usage and GC composition within and across genomes

Robin D Knight, Stephen J Freeland and Laura F Landweber

Address: Department of Ecology and Evolutionary Biology, Princeton University, Princeton, NJ 08544, USA.

Correspondence: Laura F Landweber. E-mail: lfl@princeton.edu

Published: 22 March 2001

Genome ***Biology*** 2001, **2(4)**:research0010.1–0010.13

The electronic version of this article is the complete one and can be found online at http://genomebiology.com/2001/2/4/research/0010

(Print ISSN 1465-6906; Online ISSN 1465-6914)

Received: 21 November 2000
Revised: 1 February 2001
Accepted: 13 February 2001

Abstract

Background: Correlations between genome composition (in terms of GC content) and usage of particular codons and amino acids have been widely reported, but poorly explained. We show here that a simple model of processes acting at the nucleotide level explains codon usage across a large sample of species (311 bacteria, 28 archaea and 257 eukaryotes). The model quantitatively predicts responses (slope and intercept of the regression line on genome GC content) of individual codons and amino acids to genome composition.

Results: Codons respond to genome composition on the basis of their GC content relative to their synonyms (explaining 71-87% of the variance in response among the different codons, depending on measure). Amino-acid responses are determined by the mean GC content of their codons (explaining 71-79% of the variance). Similar trends hold for genes within a genome. Position-dependent selection for error minimization explains why individual bases respond differently to directional mutation pressure.

Conclusions: Our model suggests that GC content drives codon usage (rather than the converse). It unifies a large body of empirical evidence concerning relationships between GC content and amino-acid or codon usage in disparate systems. The relationship between GC content and codon and amino-acid usage is ahistorical; it is replicated independently in the three domains of living organisms, reinforcing the idea that genes and genomes at mutation/selection equilibrium reproduce a unique relationship between nucleic acid and protein composition. Thus, the model may be useful in predicting amino-acid or nucleotide sequences in poorly characterized taxa.

Background

Different organisms have idiosyncratic, and sometimes extremely biased, preferences for one synonymous codon over another. Although differences in codon usage among genes and species have been widely studied, general principles have been difficult to find. Although it has been known for some time that the frequencies of some codons and amino acids correlate with genome GC content [1], the causality has remained unclear: correlations could exist because selection for a particular codon or amino-acid usage produces a particular genome GC content, or because mutation towards a particular GC content determines codon and

amino-acid usage according to combinatorial principles. Here we show that codon and amino-acid usage is consistent with forces acting on nucleotides, rather than on codons or amino acids, although both mutation and selection play important roles.

Codon usage can be surprisingly biased in different species. For example, the amino acid lysine has two codons, AAA and AAG. Although some organisms, such as *Lactobacillus acidophilus*, use the two codons equally, others show extreme preferences: *Streptomyces venezuelae* uses AAA only 2.2% of the time, whereas *Buchneria aphidicola* uses it for 91% of lysine residues. Amino-acid usage also differs greatly among species: for instance, the amount of arginine varies almost ten-fold, from less than 1.5% of all amino-acid residues in species of *Borrelia* to 12.7% in *Mycobacterium tuberculosis* (data from [2]). Because of these extreme biases, knowing an organism's preferred codon usage is of direct practical relevance in minimizing degeneracy of PCR primers and in maximizing the effectiveness of *in vivo* genetic manipulation. Trends in codon usage across species could also influence molecular phylogenetic reconstruction, and clarify the relative roles of neutral evolution and natural selection in determining nucleotide sequences.

The evolutionary theory of synonymous codon usage began with two separate lines of research, both of which suggested that most substitutions were selectively neutral, but which explained different phenomena. The first line sought to explain interspecific variation in overall sequence composition, and noted correlations between GC content and amino acid content across different species. This suggested that genomes were at equilibrium with respect to mutation, and explained how directional mutation could affect the composition of coding sequences [1,3,4], although it does not explain why species with similar genome compositions have recognizably distinct sequences for individual genes. The second line sought to explain the origin and maintenance of sequence variation within populations, and the fixation of particular alleles between species. This relied on the concept of silent mutations and the relative power of selection and drift in small populations [5,6]. Different usage patterns of synonymous codons are invisible at the protein level: how can selection operate when the amino-acid meaning remains unchanged [7]? However, without directional mutation pressure, the fixation of silent mutants would not lead to the extreme biases in synonymous codon usage actually observed [4].

Subsequently, codon usage in a few species has been extensively characterized, and linked causally to a wide variety of both adaptive and nonadaptive factors including tRNA abundance [8-14], gene expression level [15-23], local compositional biases [24-28], rates and patterns of mutation [29-32], protein composition [33-36], protein structure [37-39], translation optimization [40-42] (but see [43]), gene length [44-47], and mRNA secondary structure [48-52].

In contrast, trends across species have received far less attention. The genome GC content has been shown to correlate with cross-species differences in frequencies of codons [53,54] and amino acids [29,33,34,55-58]. Genome composition may even influence the structure and chemistry of proteins [36,57,59]. Comparing different microbial genomes, codon usage in individual genes also correlates with estimated expression level [60,61] and tRNA copy number [62].

One important point is that these regressions are ahistorical: by predicting a relationship between gene and protein composition, these studies imply that the history of a gene or species is unimportant compared to its current state. This has important implications for species or genes that have uncertain phylogenetic relationships or differ greatly in composition from their close relatives. Although closely related organisms tend to have similar genome compositions, there are considerable exceptions (such as *Mycoplasma pneumoniae* versus *M. genitalium*). If distantly related species with similar GC contents have the same amino-acid or codon usages, we can conclude that phylogenetic constraints are relatively unimportant, and perhaps that genomes are at or near equilibrium with respect to mutation and selection (otherwise, different unrelated species would not attain the same amino-acid composition predicted from the nucleotide composition). Such ahistorical relationships are particularly useful in cases where the goal is prediction of the current state of a sequence (for example, for making PCR primers), rather than reconstruction of its history.

Although regression lines have been fitted to relationships between GC content and codon and amino-acid usage empirically, permitting qualitative inferences, quantitative theoretical predictions relating these responses to each other have thus far had limited success. This can be remedied by taking into account the differential effect of selection on the different positions within codons. Here we present a simple model, based solely on purifying selection and mutation at the nucleotide level, that quantitatively predicts both codon and amino-acid usage trends across archaea, bacteria and eukaryotes on the basis of the genome GC content.

The model also provides insights into the causality between genome composition and protein composition. Every nucleic acid sequence necessarily has an associated GC content, but there need be no similarities in codon usage between different species with the same GC content (for instance, any specified GC content could be obtained by mixing AAA and GGG codons in different ratios). If GC content were an artifact of selection for a particular codon or amino-acid usage, there would be many different ways of arranging the codon frequencies to get the same GC content. If, on the other hand, the codon and amino-acid usage is an artifact of mutation (or selection) towards a particular GC content, the responses of the three codon positions to directional nucleotide substitution predict a single codon or amino-acid usage for each

GC content. Thus, if distantly related species fit the response curves predicted by the model, we can conclude either that forces at the nucleotide level drive codon and amino-acid usage, and there is nothing special about certain codons or amino acids, or that there is a unique spectrum of preferred codon and amino-acid usages that applies to all species, extends over a huge range of compositions, and happens to match the predictions of the model by chance.

Results and discussion

Empirical relationships between GC content and codon/amino-acid frequency

Graphs regressing codon and/or amino-acid frequency onto GC content for subsets of the 64 codons and 20 amino acids (GC response graphs) have been plotted previously, although typically for particular orthologs across up to 30 species [57,58] or for individual genes within a species [29,33,34]. The exception is an analysis of the influence of GC content on average amino-acid composition in 59 bacterial species [56]. Here we plot only the graph for two arginine codons, CGA and CGG (Figure 1a), and for two amino acids, arginine and threonine (Figure 1b), to illustrate that some codons and amino acids, such as CGC and arginine, show a clear relationship with genome GC content; others, such as CGA and threonine, show no relationship whatsoever. All amino acids differ in frequency by two- to ten-fold in different species, however, suggesting that most sites within proteins can tolerate amino-acid substitutions.

The following regression analyses assume a specific flow of causality: that GC content drives codon (and amino-acid) usage. We favor this direction because there are many ways to get a GC content from different codon usages, but only one way to predict a set of codon usages from GC content. Our interpretation follows Muto and Osawa's seminal demonstration of the high correlation between the GC content of noncoding, exon, intron, tRNA and rRNA sequences, which indicates that a common force influences the composition of all parts of the genome [53].

Three different properties of the regression line could be considered as measures of a codon's (or amino acid's) response to GC content: the slope (which describes the rate of response, or sensitivity, to GC content variation), the intercept (marking the lower boundary of GC content, at which the codon/amino acid is predicted to disappear), and the correlation coefficient (describing the degree of variation in codon/amino-acid usage that can be understood in terms of genome GC content). These three measures are in fact highly correlated with one another (Table 1). We use total GC content rather than third position GC content (GC3) for the regressions, as total GC is easily measured directly in the laboratory for otherwise uncharacterized organisms; to estimate GC3, at least some gene sequences are required. In principle, it is preferable to use GC3 where available, as total genome (or coding sequence) GC already contains the data used to measure the GC content at the other positions. Consequently, regressing GC1 or GC2 against total GC might introduce systematic biases, whereas regressing against GC3 better represents deviation from neutrality [4]. However, total GC and GC3 are so highly correlated (in part because GC3 changes much faster than GC1 and GC2) that for practical purposes it makes no difference.

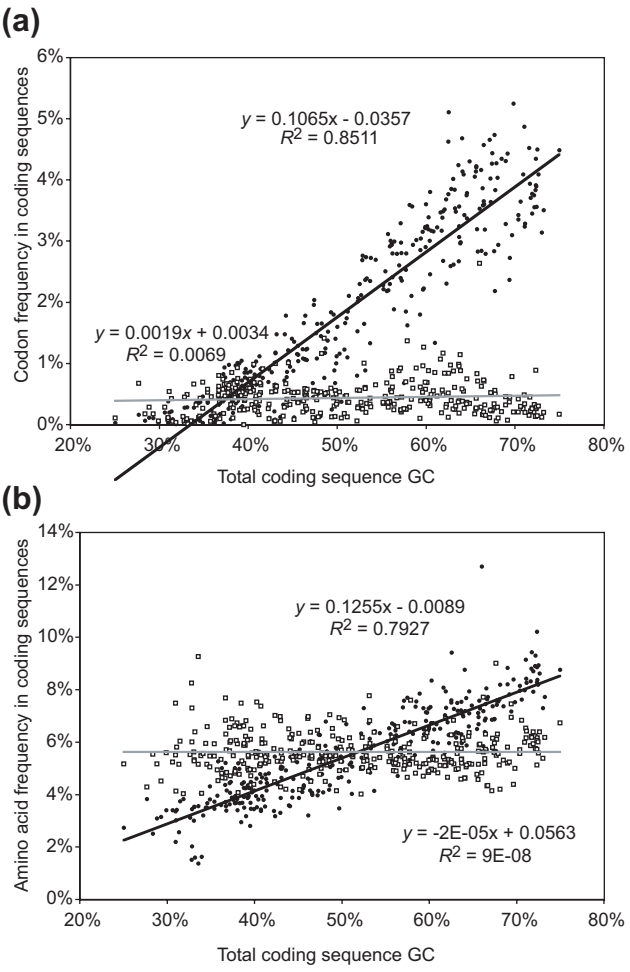


Figure 1
Only some codons and amino acids respond to GC content. **(a)** Plot of codon frequency within coding sequences versus total GC content, for the arginine codons CGA (white squares) and CGC (black circles) in bacteria and archaea. Linear regression lines are shown in black for CGC and gray for CGA. **(b)** A similar plot for the amino acids threonine (white squares) and arginine (black circles) in bacteria and archaea. The plots show that whereas CGC and arginine clearly correlate with GC content, CGA and threonine do not. The three relevant parameters for the response, slope, intercept and correlation coefficient, are all highly correlated with each other (see Table 1).

The factors determining the response of each codon and each amino acid to genome GC content turn out to be surprisingly

Table 1

Pairwise correlations between measures of response to genome GC content

	Amino acids (n = 21)				Codons (n = 64)			
	Archaea/bacteria		Eukaryotes		Archaea/bacteria		Eukaryotes	
	Slope	Intercept	Slope	Intercept	Slope	Intercept	Slope	Intercept
Slope	—	0.95	—	0.96	—	0.90	—	0.91
Correlation coefficient	0.92	0.96	0.96	0.98	0.90	0.97	0.94	0.94

All pairs of measures are highly correlated. Critical values: for amino acids (n = 21 including stop codons), r of 0.9 corresponds to $P = 3 \times 10^{-8}$. For codons (n = 64), r of 0.9 corresponds to $P = 5 \times 10^{-24}$. The x intercept is transformed as $x' = 1/(50\% - x)$ to minimize the effects of extreme values with large errors: this occurs for codons and amino acids with very flat slopes.

simple. An amino acid's response (Figure 2a) is determined by the mean GC content of its codons (that is, the amino acids with particularly AT- or GC-rich codons are most sensitive to genome GC content), which explains 71-79% of the variance in response, depending on the measure used (slope is the poorest fit; correlation coefficient is the best). A codon's response (Figure 2b) is determined by the difference between its GC content and the mean GC content of its synonyms, explaining 71-87% of the variance. (In other words, genome GC content influences the presence or absence of codons with extreme GC content more than synonyms with intermediate GC content. This follows from the fact that the third codon position changes faster than the other two positions, and does not depend directly on the amino acid that each codon encodes.) This relationship applies to both eukaryotes and prokaryotes. Most of the diversity in GC content within eukaryotes, and hence most of the significance of the regressions, comes from unicellular and multinucleate organisms, including fungi and protists, rather than from the multicellular plants and metazoa. This is to be expected, because the relatively early-diverging protist lineages also account for most of the molecular diversity within the eukaryotes [63].

Although the figures reported above include termination codons, excluding these codons does not greatly affect the result (for instance, R^2 increases from 0.72-0.75 for slope in the codon graph for eukaryotes when the three termination codons are excluded). Similarly, excluding tryptophan and methionine, which have only one codon each and thus necessarily fall at the origin, makes no difference, as the best-fit line (for codons) passes through the origin anyway. Excluding stop codons, tryptophan and methionine increases R^2 by only 0.019 on average for the various measures of response. Excluding stop codons makes slightly more difference on the amino-acid graph (increasing R^2 by 0.031 on average, more influential than any single amino-acid point). When stop codons, tryptophan and methionine are all excluded, R^2 increases by only 0.046 on average.

Interestingly, the same sorts of relationships seem to hold within, as well as among genomes. Figure 2c examines the response of codons to GC content across individual genes within sample genomes representing all three domains of life: eukaryotes (*Drosophila*), bacteria (*Synechocystis*), and archaea (*Archaeoglobus*). The codon frequency in individual genes seems to be predictable on the basis of the overall GC content of the coding sequences: the relative GC content of individual codons explains about half the variance in response in their absolute frequencies within each coding sequence (that is, the presence or absence of a particular codon depends on the extent to which alternative synonyms of more 'suitable' GC content exist). This is despite the fact that codon usage in individual genes is known to be influenced by a long list of other factors, including the amino-acid sequence necessary for the gene's function (although the fraction of crucial residues may typically be small). The effect cannot be fully explained by synonymous substitutions, because codons with the same third-position base but different codon doublets respond differently.

Before proceeding, we note that these plots themselves emphasize the causality within the system. Our analysis shows that all amino acids and all codons behave in a predictable manner within each genome, indicating that the GC content within coding sequences determines the codon (and hence amino-acid) composition rather than being a passive reflection of a preferred codon or amino-acid usage. Hence, this uniform robustness supports the idea that there is little special about particular codons or amino acids. In all three domains of life, it appears that every codon and every amino acid follows a single trend determined by the overall compositional properties of genomes. Even within genomes, the overall composition of individual genes seems to explain up to half the variance in the responses of the individual codons and amino acids. This raises the question: what drives GC content?

A mutational model

The simple empirical relationship between the composition of codons and amino acids and their responses to changing genome composition suggests that these responses might be quantitatively predictable on theoretical grounds. We take as

our starting points Sueoka's hypothesis that genome composition is largely determined by mutation bias, specifically the ratio of AT→GC to GC→AT mutations [3,29,32], and Muto and Osawa's demonstration that different types of sequence, and the three codon positions within exon sequences, vary in their response to genome GC content with the third position changing the fastest, which suggests that the different observed substitution rates might be explained by the differential effects of mutation [53].

We assume for simplicity that the genome is divided into two types of site - constant and variable; that all variable sites respond identically to mutational pressure; and that all third-position sites are variable (that is, that under this initial model there is no selective consequence to silent mutations). Accordingly, GC3 (the third-position GC content) reflects the ratio of AT→GC to GC→AT mutations. When plotted against GC3, the slope of the GC1 and GC2 graphs give the intensity of selection at those two positions relative to GC3 [4], whereas the values at GC3 = 0% and GC3 = 100% give the identity of the bases at the constant positions (Figure 3). In other words, when all the variable sites (as measured by GC3) are as biased as possible towards one state, we assume that any sites at position 1 or 2 that still have the opposite state are maintained by selection and cannot change. Thus, the frequency of each codon can be estimated as:

$$P[XYZ] = P[X_1]P[Y_2]P[Z_3] \quad (1)$$

where $P[X_1]$, $P[Y_2]$, and $P[Z_3]$ denote the probability of getting base X at position 1, base Y at position 2, and base Z at position 3, respectively.

The probability of getting a particular base at a particular position is given by:

$$P[X_n] = P[X|c_n]P[c_n] + P[X|v]P[v_n] \quad (2)$$

where $P[c_n]$ denotes the probability that a site is constant at position n, and $P[v_n]$ denotes the probability that a site is

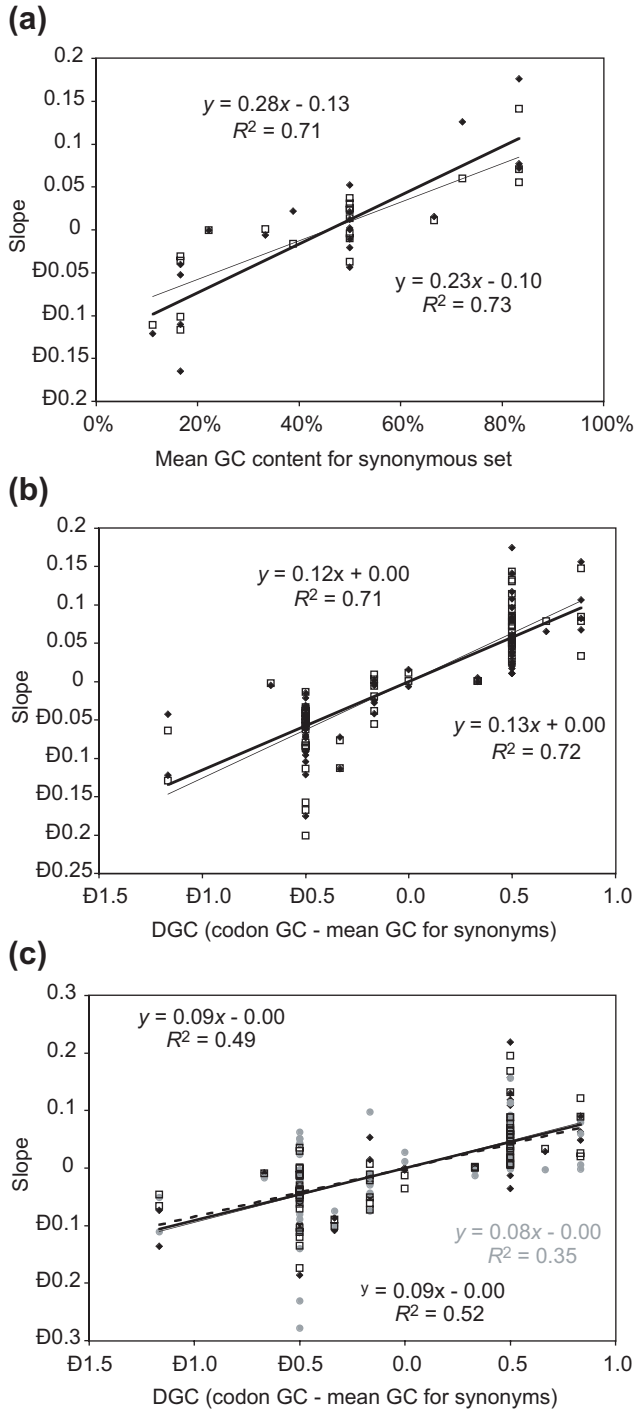


Figure 2

Figure 2
Codon and amino-acid responses are determined by their individual GC content. **(a)** Plot of response to GC content (here, the slope of the regression of absolute frequency in coding sequences on genome GC content) versus composition of the 21 codon sets (20 amino acids and termination) for archaea/bacteria (black symbols, thick lines) and eukaryotes (white symbols, thin lines). **(b)** A similar plot for the 64 codons. Note that, of the three measures of response, the slope is the least highly correlated with codon or amino-acid composition (see Table 2). For amino acids the composition is the mean GC content of their codons (a). For codons (b,c) the composition is the difference (ΔGC) between the codon's GC content and the mean GC content for all codons encoding the corresponding amino acid. **(c)** A response-composition plot of the 64 codons showing response within genomes rather than between them, for a bacterium (*Synechocystis*, black symbols, thick line), an archaean (*Archaeoglobus*, gray symbols, gray line), and a eukaryote *(Drosophila*, white symbols, dashed line). The gray line is almost coincident with the thick line; the points are clustered along the abscissa because the structure of the code restricts the possible GC content of the codon sets.

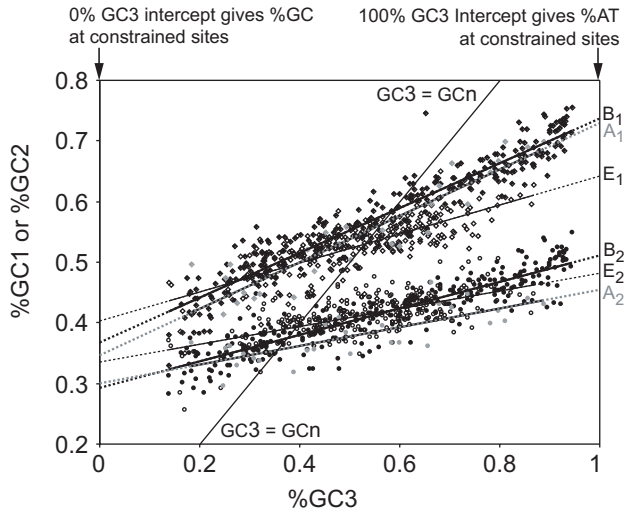


Figure 3
The codon response to genome GC content varies with position. A re-plot of GC3 versus GC1, GC2 from [4], using the additional sequence data now available. Each point represents an organism, classified by domain: archaea, gray; bacteria, black; eukaryotes, white. GC1, diamonds; GC2, squares. Lines are model I least-squares regressions. Where GC3 = 0%, the remaining %GC in position 1 and position 2 is assumed to represent constant sites (that is, those fixed by selection to remain G or C). Similarly, where GC3 = 100%, the remaining %AT in position 1 and position 2 is assumed to represent constant sites where A or T have been fixed.

Table 2

Correlations between composition and response to GC content

	Amino acids		Codons	
Correlation with:	Archaea/ bacteria	Eukaryotes	Archaea/ bacteria	Eukaryotes
Slope	0.842	0.853	0.840	0.849
Intercept	0.878	0.886	0.931	0.910
Correlation coefficient	0.886	0.887	0.934	0.922

Amino acids: correlation of each measure of response with mean codon GC content. Codons: correlations with difference between GC content and mean GC content of synonymous codons. See also Figure 2.

variable at position n (that is, that it is not constant). These two probabilities sum to 1. $P[X|c_n]$ denotes the conditional probability of getting base X at a constant site at position n, and $P[X|v]$ denotes the conditional probability of getting base X at a variable site (assumed to be the same across all three positions).

Comparing the responses of GC1 and GC2 to GC3 (Figure 3, Tables 2 and 3), we can see that:

1. The slopes for GC1 in bacteria and archaea are almost identical. However, the bacterial/archaeal slope differs significantly from the eukaryotic slope. Thus, first-position residues are less labile among the eukaryotes tested. (Note that the number of degrees of freedom is n1 + n2 – 4: the data are constrained by a sum and a regression mean-square).

2. The slopes for GC2 are not significantly different in the three domains.

3. The intercepts for archaea and bacteria do not differ, except for GC2 when GC3 is 100%. The intercepts for the pooled archaea/bacteria sample always differ from those of eukaryotes.

4. Archaea behave far more like bacteria than like eukaryotes. Using the Kolmogorov-Smirnov test in two dimensions as implemented in [64], $N_A = 28$, $N_B = 311$, $N_E = 257$: for GC1 versus GC3, $D_{AB} = 0.19$, $P_{AB} = 0.33$, $D_{AE} = 0.30$, $P_{AE} = 0.03$; for GC2 versus GC3, $D_{AB} = 0.21$, $P_{AB} = 0.02$; $D_{AE} = 0.54$, $P_{AE} = 4 \times 10^{-6}$. Differences between eukaryotes and bacteria, or between eukaryotes and the combined archaeal/bacterial data set, were highly significant ($P < 10^{-6}$ that both represent samples drawn from the same population for each pairwise test).

5. The general patterns are replicated in the three domains. Thus, the relationship between the first-, second- and third-position GC content emerges independently in evolutionarily separate groups. Further subdividing each domain shows that many lineages have explored the same wide range of GC contents, reproducing the same relationships.

The major distinction is between eukaryotes and prokaryotes, which presumably has an ecological or structural rather than a phylogenetic explanation. The slopes may be different because of the great differences in genome size and generation time, or because the partitioning of genetic material into the specialized environment of the nucleus changes patterns of mutation. Eukaryotes also have far more noncoding DNA, which could potentially isolate coding regions from selection for genomic GC content if this were an important force: however, for this analysis we consider only the coding regions, which still differ greatly in composition among different eukaryotes (suggesting that selection has not acted to conserve the GC content of coding sequences). The intercepts may be different because the set of proteins in each domain differs, and so the distribution of nucleotides in critical sites need not be conserved. It is also possible that the selection of genes that have been studied differs between the groups, although the few organisms for which complete genomes are known are not outliers. To reflect the major differences separating the eukaryotes from the other domains, we pooled the archaeal and bacterial data and contrast this

Table 3

Responses of GC1 and GC2 to changes in GC3, by domain

	n		R^2	Slope ± SE	Y at GC3 = 0 ± SE	Y at GC3 = 100 ± SE
Bacteria	311	GC1	0.91	0.370±0.007	0.367±0.004	0.737±0.003
		GC2	0.80	0.219±0.006	0.291±0.004	0.510±0.003
Archaea	28	GC1	0.85	0.38±0.03	0.35±0.02	0.73±0.02
		GC2	0.60	0.16±0.03	0.30±0.01	0.45±0.01
Eukaryotes	257	GC1	0.57	0.24±0.01	0.402±0.008	0.643±0.007
		GC2	0.38	0.15 ±0.01	0.334±0.007	0.482±0.006

Because there is error in both axes, but there should be a definite causal relationship between GC3 and GC1 or GC2, we use model I regression to predict specific values of GC1 or GC2 from a set value of GC3, and thus to calculate the most likely proportion and GC content of constant sites [73].

with the eukaryotic data, fitting slopes and intercepts to the model for the two groups separately.

Does the model accurately reflect codon usage?

The model outlined above requires four parameters (estimated GC1 and GC2 at GC3 = 0% and GC3 = 100%) defined from the data, and uses these to predict the slopes and intercepts of regressions on GC content of the 64 codons and 21 codon sets (20 amino acids and termination). As the model is deterministic, it is not useful for predicting the correlation coefficients. When composition is summed across all three codon positions (for example, CGA and GGT would be counted among the eight codons comprised of two G or C and one A or T, with the A/T at the third position), the model is remarkably accurate for both bacteria (Figure 4a-c) and eukaryotes (Figure 4d), though in each case, GC3 shows the greatest deviations from the model (perhaps indicating that the labile GC3 is 'fine tuned' by the other selective pressures linked to codon usage).

This model even performs moderately well when applied to a random sample of 500 genes from the *Drosophila* genome (Figure 4e): although the unexplained variance is much greater, the points clearly cluster around the three lines predicted by the position-dependent model rather than the single (orange) line that overall composition alone would predict (as in previous, simpler models). In each case, the white-centered lines are the theoretical predictions, and each dark point represents a species. The orange line is the frequency that would be expected from the GC content without taking into account the position-dependent composition biases, that is by $(GC)^n(1\text{-}GC)^{3\text{-}n}$, where GC is the genome GC content and n is the number of G and C in the codon [56]. Taking the position dependence into account thus provides a much better fit to the data.

We compare the predicted and actual slopes for each of the 64 codons in Figure 5. The model explains 77% of the variance in slopes for the regression of codon frequency on GC content in eukaryotes, and 80% of the variance in prokaryotes. Additionally, the model is an unbiased estimator: the slopes are 1, and the intercepts are 0. The results are similar for the intercepts, and for amino-acid usages (Table 4).

Can selection account for the remaining variance?

The four-parameter model discussed above assumes, for the sake of simplicity, that A = T and C = G (Chargaff's rule). For double-stranded DNA molecules, this is necessarily true because of Watson-Crick base pairing. Less well known is the intra-strand Chargaff's rule, which states that the same relationship holds within large, single-stranded DNA molecules (in particular, the two strands of the *Bacillus subtilis* genome) [65,66]. The interpretation of this intra-strand rule is statistical rather than mechanical: if there are no biases in mutation and selection between the two strands, or if genes are distributed evenly between the two strands, a C→G mutation on one strand (for example) cannot be distinguished from a G→C mutation on the other. Thus, the twelve possible nucleotide-substitution rates reduce to just six and, at equilibrium, Chargaff's rule should hold even within the set of coding sequences in a genome [30]. As long as all nucleotide-substitution rates are positive, this equilibrium condition holds for all possible substitution matrices [67].

In actual genomes, however, the intra-strand Chargaff's rule is frequently violated because the leading and lagging strands have different substitution patterns and genes are not evenly distributed [31]. In respect of our model, not only the position of a base but also its identity affects how fast it responds to genome GC content (Table 5). Interestingly, the intra-strand Chargaff's rule is violated in a position-dependent manner. For both prokaryotes and eukaryotes, the third codon position is pyrimidine-rich (C3/G3 = 1.11 and 1.15 respectively; T3/A3 = 1.24 and 1.36), and the first codon position is purine-rich (C1/G1 = 0.58 and 0.62; T1/A1 = 0.63 and 0.65). The second codon position is mixed (C2/G2 = 1.34 and 1.32; T2/A2 = 0.97 and 0.87). Consequently, relaxing the assumption of the intra-strand Chargaff's rule should increase the accuracy of the model.

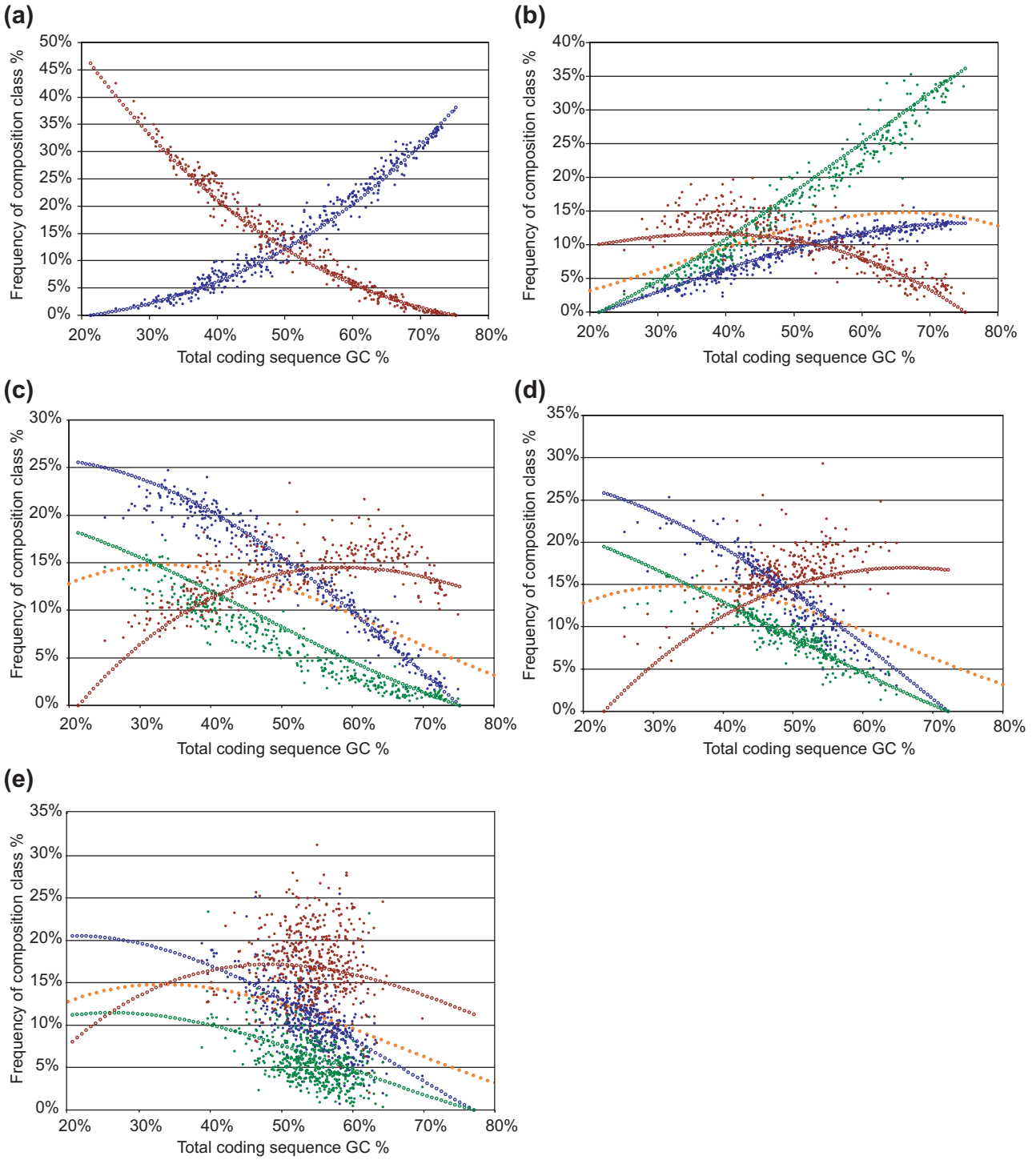


Figure 4
Predicted versus actual responses for sets of codons with identical composition. Each line is the sum of eight codons with the same GC content (by position). Each solid circle is a species. Lines of open circles are the theoretical predictions based on the four-parameter model. **(a)** All-GC (blue) and all-AT (red) codons in prokaryotes. **(b)** Codons with two G or C and one A or T, the minority base being at the first (blue), second (green), or third (red) position. Note that the third-position slope is actually of opposite sign to the first- and second-position slopes. The orange line is what would be expected if there were no position dependence (that is, $P(GC)^2P(AT)$ as in [56]). **(c)** As in (b), but for codons with two A or T and one G or C. In this case, the orange line is $P(AT)^2P(GC)$. **(d)** As in (c), but for eukaryotes. **(e)** As in (d), but now each point is a randomly chosen gene in *Drosophila*.

Table 4

Concordance between predictions and data

		Archaea/bacteria			Eukaryotes		
	Amino acids/codons	Observed	Four-parameter	24-parameter	Observed	Four-parameter	24-parameter
Observed	Slope	-	0.89	0.93	-	0.88	0.92
	Intercept	-	0.91	0.92	-	0.90	0.91
Four-parameter	Slope	0.83	-	0.95	0.80	-	0.96
	Intercept	0.82	-	0.99	0.80	-	0.98
24-parameter	Slope	0.90	0.94	-	0.85	0.95	-
	Intercept	0.86	0.96	-	0.80	0.95	-

Each entry is the correlation coefficient between the predicted and observed values for the 64 codons (above the diagonal) or 21 amino acid sets (below the diagonal) for archaea and bacteria (left) and eukaryotes (right). See Figure 5 for an example (the graph for the first entry on the second column: note that the graph shows R^2, not R).

These differences in composition could reflect coding constraints, if a functional proteome required a particular amino-acid composition. As we have seen, however, the frequency of particular amino acids varies greatly among different organisms, decreasing the likelihood that there is a unique, optimal composition. Additionally, the amino acids respond predictably with changing GC content, in a manner consistent with processes acting only at the level of single nucleotides.

If the bases do not change at the same rates, the assumption that the GC content at each position completely describes the nucleotide composition is unwarranted. The four-parameter model discussed above assumes that, for each of the three codon positions, each of the four bases changes equally rapidly with changing genome GC content. Interestingly, this is not actually the case. For example, G changes far more slowly than any other nucleotide when at the third codon position, but faster than T when at the first or second codon position. Furthermore, A at the second position changes nearly twice as fast as T at the second position (Table 5). This violates the assumption that all variable sites are equal.

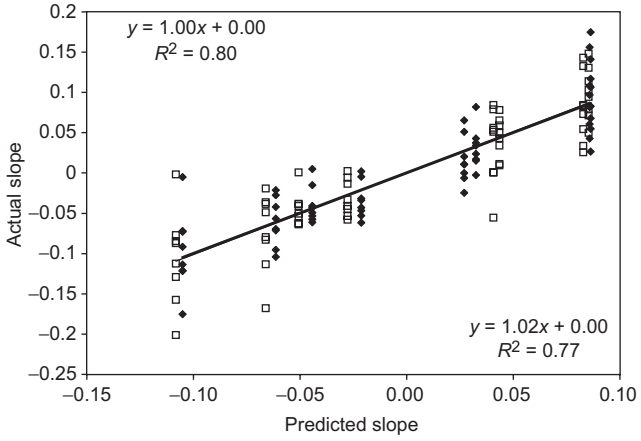


Figure 5
Comparison of predicted versus actual codon responses. Both bacteria/archaea (black) and eukaryotes (white) show a very good fit between the model and the data (in this case, predicted slopes along the *x* axis and actual slopes along the *y* axis). The slope is 1 and passes through the origin in both cases, indicating that the model is an unbiased predictor of codon usage trends. See Table 4 for other comparisons.

Is there any rationale behind these seemingly arbitrary rates? Here, selection rather than mutation may provide an answer. Most mutations are deleterious; furthermore, the greater the effect of a particular change, the less likely it is to be advantageous [68]. We can estimate the average effect of changing each base at each position according to a method used previously to calculate the effects of errors in individual codons [69]. Briefly, for each mutation, the difference in 'polar requirement' [70,71] between amino acids encoded by the original and new codons is squared. The resulting error value is averaged for all applicable mutations, weighting transitions more heavily than transversions because they occur more frequently (in this case, by a factor of 4 as reported in [72] for comparison of human pseudogenes with their functional predecessors). This gives an *a priori* estimate of the impact of a given set of mutations based on the chemical properties of the amino acids and the configuration of the genetic code.

In fact, the mean-square error does an excellent job of accounting for the difference in slopes. The mean-square errors give a logarithmic fit to the rates of change, but because there is no reason to believe this functional relation to be the correct one we used a nonparametric test for correlation (Spearman's rank coefficient [73]). For eukaryotes, r_s = -0.83 (P = 8.3 x 10^{-4}); for prokaryotes, r_s = -0.86 (P = 3.3 x 10^{-4}); n = 12 in both cases. The rank order of the mean-square errors does not change when the modular power and/or the transition bias are varied over the range one to

Table 5

Violations of Chargaff's rule and rate constancy

	|slope A/B|	|slope E|	Error
T1	0.251	0.248	6.591
C1	0.421	0.382	4.640
A1	0.466	0.334	2.474
G1	0.296	0.201	3.930
T2	0.095	0.126	7.738
C2	0.221	0.217	5.256
A2	0.333	0.254	11.984
G2	0.207	0.163	7.487
T3	0.938	0.986	0.065
C3	1.108	1.212	0.065
A3	0.917	1.051	0.076
G3	0.746	0.825	0.082

|slope AB| is the absolute value of the change of a given nucleotide at a given position (relative to total coding sequence GC content) in archaea/bacteria; |slope E| is the corresponding slope in eukaryotes. Error is a measure of the average consequence of a change in a particular base at a particular codon position (for example, T1 is T at the first position, using a methodology based on [69]. See text for explanation.

ten, so the correlations are robust across parameter space. In other words, because the rate of change of the different nucleotides depends on the magnitude of error introduced on average by altering them, we interpret the variable response of GC content at each position to be dependent on base identity as the result of selection against substitution of dissimilar amino acids.

Relaxing the constraints of the model

To take differential selection into account, we relax the assumption of the intra-strand Chargaff's rule as follows. The four-parameter model presented above requires two parameters each for two regression lines, relating the first- and second-position GC content to the third-position GC content. However, if the frequencies of the four nucleotides at each codon position can vary independently with GC content (subject to the constraint that the nucleotide frequencies at each position are constrained by a sum), it is necessary to characterize the regression lines of each base at each position to make predictions about the nucleotide composition of the set of constant bases at each position, and of the most likely states of variable bases for a given GC content.

Hence we constructed a 24-parameter model (4 bases x 3 positions x 2 parameters for each regression line) where, for each position, we plotted the percentage of each of the four bases against total GC content. For a given total GC content, the expected frequency of a particular base at a particular position is estimated directly from its regression line, which is based on two parameters (the slope and the intercept). This takes into account the fact that an organism at a given GC content will predictably violate the intra-strand Chargaff's rule in its coding sequences. This could be considered an extension of Takahata's analysis of rate heterogeneity among the four nucleotides [74], but extended for the reading-frame-dependent selection in coding sequences. The model actually has only 18 degrees of freedom (rather than 24), as the sums are constrained, but predicts a set of codon frequencies that potentially has 63 degrees of freedom.

This 24-parameter model explains somewhat more of the variance in both codon and amino-acid responses (slope and intercept), although the marginal benefit is greater in prokaryotes than in eukaryotes (Table 4). The improvement in R^2 can explain nearly 40% of the variance unexplained by the four-parameter model in some cases (amino-acid slopes in archaea/bacteria), although in other cases the 24-parameter model does not offer an improvement (for example, amino-acid intercepts in eukaryotes). One possible interpretation is that for our data set selection plays a greater role in the genome composition of prokaryotes. This is certainly plausible given the bias in eukaryotic sequences towards large species with small populations. More generally, we may infer that, on the scale of whole genomes, differential mutation and selection between the two strands play relatively little role in determining codon usage.

Conclusions

We have shown that the GC content of individual codons and amino acids is the primary determinant of their response to biases in sequence composition, both among and (to a lesser extent) within genomes. Although the literature contains many examples of correlations between GC content and the frequency of particular codons and amino acids, our model is able to recapture quantitatively the behavior of essentially all codons and amino acids by invoking forces that act only on the level of individual nucleotides. This is likely to be due to a combination of mutation and selection: mutation can act in parallel across an entire genome, changing many sites simultaneously; however, this process is limited by the consequences of error at each position.

The simplest hypothesis, that codon usage depends solely on codon GC content [56], fits the data poorly (compare orange lines with red, green and blue lines in Figure 4). One can, however, explain most of the variance in the response of both codons and amino acids by taking into account the fact that the three codon positions change at different rates, and that the four nucleotides are not evenly distributed among the sites that are functionally constrained. Additionally, accounting for the fact that the four nucleotides change at different rates allows some further improvement, which ranges from minimal to drastic depending on the exact circumstances. This supports the basic principle of neutral

evolution, the idea that most change in nucleotide and protein sequences is driven by mutation and limited by purifying selection that varies for different sites and molecules (reviewed in [75]). Within this context, it supports the idea that most of this neutral change is driven by directional mutation, which thus explains differences in nucleotide composition among species [4].

Although the conclusion that amino acids with GC-rich codon doublets are more frequent in GC-rich genomes, and that those with AU-rich codon doublets are more frequent in AT-rich genomes, is neither new nor surprising [34], our model accurately and quantitatively predicts these responses for essentially all codons and amino acids by invoking forces acting on individual nucleotides. The genetic code constrains which codons and which amino acids can respond to biases in nucleotide composition, in part because mixed codons necessarily respond more slowly to forces acting on particular types of bases than do homogeneous codons. Thus, although GC content only explains the variance in usage of some codons and some amino acids, we can accurately predict which codons and amino acids will show clear responses and, for those that do show clear responses, accurately predict their frequencies in particular genomes (for example, Figure 1 shows an example of a codon for which 85% of the variance in usage is explained by genome GC content, and an amino acid for which 79% of the variance is explained). Thus, especially for species with few close relatives, variable sites may even be more useful for predicting PCR primer sequences than conserved sites, although this will depend on the particular sequence and genome composition.

We have focused on codon usage at the level of whole genomes (or samples of genes where whole genomes are not available), an area that has received relatively little attention. This large-scale view does not consider the selective factors influencing individual genes, and the fact that the model provides much better fit across genomes than within them may reflect local adaptation to factors such as expression level [11]. What remains surprising is that our simple model can explain so much of the variance in codon and amino-acid response to GC content in these different systems. Identifying deviations from the predictions based on nucleotide composition may identify genes that are under unusual selection pressures, whether for a particular amino-acid composition or for a specific pattern or degree of codon bias.

The fact that both amino-acid and codon usage are so closely entwined with genome composition has important practical implications. For phylogenetic analysis, the fact that some amino acids (such as arginine) change rapidly and predictably with GC content slightly undermines the idea that amino-acid sequences are more stable than nucleotide sequences: pairs of species with convergent GC contents might also evolve convergent protein sequences, especially at functionally unconstrained positions. For example, the frequencies of both lysine and arginine are highly (but oppositely) correlated with GC content, and lysine and arginine can easily substitute for one another in proteins. Each of the three domains of life has explored a wide range of genome GC contents, and organisms at the extremes of the range but with different evolutionary histories may share more convergent amino-acid substitutions than currently recognized.

For sequence analysis, the prospects are more promising: given very limited information about a species (the GC content), it may be possible to estimate the codon usage and therefore minimize the degeneracy of PCR primers, even if no closely related species have been characterized. Organisms with extreme genome compositions, or with genome compositions that differ markedly from their close relatives (such as *Mycoplasma pneumoniae* versus other mycoplasmas) should be particularly accessible. This should be especially useful in developmental genetics and in environmental applications where model systems are not available.

The fact that the model holds independently for different lineages of organisms (for example, bacteria and eukaryotes), and, to a lesser extent, for individual genes within species, strongly suggests that the trends are ahistorical. Given rates of change for each nucleotide at each codon position, determined jointly by selection, mutation, and the genetic code structure, we can predict the codon and amino-acid composition of a particular sequence from its overall compositional properties, without reference to related sequences. Interestingly, the history of a sequence seems relatively important in determining its codon and amino-acid usage. This fact is likely to be particularly important in cases where a species diverges greatly in GC content from its closest relatives: knowing its GC content will allow much better prediction of specific gene sequences than simple comparison with conserved sites in related sequences (which may in some cases be similar because of shared genome composition rather than functional constraint).

Finally, our model explains many of the details of individual codon and amino-acid responses over the wide range of genome compositions found in nature. Perhaps surprisingly, individual amino acids with specific structural or functional roles within proteins (such as and arginine) respond to GC content no differently than the rest, and their frequencies can be very sensitive to genome composition despite the effects this might have on the properties of the translated products. This ability of amino-acid frequencies to vary so widely implies that functional proteins may be less constrained by sequence (and therefore easier to evolve) than previously imagined.

Materials and methods

Species and gene codon usage totals were downloaded from CUTG (Codon Usage Tabulated from GenBank) [2,76], which

is based on GenBank Release 117.0. Of 675 species with at least 20 protein-coding sequences tabulated from nuclear DNA, we excluded 53 eukaryotes and 17 bacteria on the grounds that they had alternative genetic codes (for example, *Tetrahymena* and *Mycoplasma*), or had introns accidentally tabulated in the database as part of the coding sequence (for example, *Pongo* and *Homo)*. These were detected as an excess of termination codons greater than 1 per 20 coding sequences (that is, at least 5% more stop codons than genes). We excluded an additional nine eukaryotes for which a few genes had been tabulated repeatedly as independent sequences (for example, *Naja atra),* leaving a sample size of 311 bacteria, 28 archaea, and 257 eukaryotes with at least 20 distinct coding sequences tabulated in the database. The choice of 20 coding sequences was arbitrary, intended to ensure a sufficiently large sample size to estimate properties of entire genomes; raising the stringency to species with 50 or 100 coding sequences (288 and 176 species, respectively) reduced the size of our data set but gave almost identical results (data not shown). We made no attempt to separate the genes by chromosome (for eukaryotes), expression level, or location, except that organellar genes were not considered. Except where otherwise noted, 'total GC' refers to the total GC content of coding sequences, rather than of genomes. These values are sufficiently highly correlated that it makes no difference which is used.

We estimated nucleotide and amino-acid compositions for genomes from the species sum records from CUTG, which sum the codons for all nuclear coding sequences deposited in GenBank for each species. We did not make any effort to exclude short, truncated, duplicated or hypothetical genes, although comparison with a filtered data set based on an earlier release of GenBank revealed no significant differences (data not shown). Thus, genes made contributions proportional to their lengths.

Codon frequencies were calculated both including and excluding termination codons. Data reported here include termination codons. Because termination codons are rare, this does not significantly alter the results, except for allowing inferences about the relative usage of UAA, UAG and UGA as termination signals.

Acknowledgements

We thank Noboru Sueoka, Mike Yarus, Erik Schultes, Jean Lobry, Dawn Brooks, and members of the Yarus and Landweber labs for comments and discussion.

References

1. Sueoka N: **Compositional correlation between deoxyribonucleic acid and protein.** *Cold Spring Harb Symp Quant Biol* 1961, **26:**35-43.
2. **CUTG (Codon Usage Tabulated from GenBank)** [http://www.kazusa.or.jp/codon]
3. Sueoka N: **On the genetic basis of variation and heterogeneity of DNA base composition.** *Proc Natl Acad Sci USA* 1962, **48:**582-592.
4. Sueoka N: **Directional mutation pressure and neutral molecular evolution.** *Proc Natl Acad Sci USA* 1988, **85:**2653-2657.
5. Kimura M: **On the probability of fixation of mutant genes in populations.** *Genetics* 1962, **47:**713-719.
6. Kimura M: **Genetic variability maintained in a finite population due to mutational production of neutral and nearly neutral isoalleles.** *Genet Res* 1968, **11:**247-269.
7. King JL, Jukes TH: **Non-Darwinian evolution.** *Science* 1969, **164:**788-798.
8. Ikemura T: **Correlation between the abundance of *Escherichia coli* transfer RNAs and the occurrence of the respective codons in its protein genes: a proposal for a synonymous codon choice that is optimal for the *E. coli* translational system.** *J Mol Biol* 1981, **151:**389-409.
9. Ikemura T: **Correlation between the abundance of yeast transfer RNAs and the occurrence of the respective codons in protein genes. Differences in synonymous codon choice patterns of yeast and *Escherichia coli* with reference to the abundance of isoaccepting transfer RNAs.** *J Mol Biol* 1982, **158:**573-597.
10. Ikemura T, Ozeki H: **Codon usage and transfer RNA contents: organism-specific codon-choice patterns in reference to the isoacceptor contents.** *Cold Spring Harb Symp Quant Biol* 1983, **47:**1087-1097.
11. Ikemura T: **Codon usage and tRNA content in unicellular and multicellular organisms.** *Mol Biol Evol* 1985, **2:**13-34.
12. Bulmer M: **Coevolution of codon usage and transfer RNA abundance.** *Nature* 1987, **325:**728-730.
13. Ikemura T: **Correlation between codon usage and tRNA content in microorganisms**. In *Transfer RNA in Protein Synthesis*. Edited by Hatfield, DL, Lee, BL. CRC Press: Boca Raton, FL; 1992,87-111.
14. Ikemura T: **Correlation between the abundance of yeast transfer RNAs and the occurrence of respective codons in protein genes.** *J Mol Biol* 1982, **158:**573-597.
15. Gouy M, Gautier C: **Codon usage in bacteria: correlation with gene expressivity.** *Nucleic Acids Res* 1982, **10:**7055-7074.
16. Holm L: **Codon usage and gene expression.** *Nucleic Acids Res* 1986, **14:**3075-3087.
17. Sharp PM, Li WH: **An evolutionary perspective on synonymous codon usage in unicellular organisms.** *J Mol Evol* 1986, **24:**28-38.
18. Sharp PM, Tuohy TM, Mosurski KR: **Codon usage in yeast: cluster analysis clearly differentiates highly and lowly expressed genes.** *Nucleic Acids Res* 1986, **14:**5125-5143.
19. Sharp PM, Li WH: **The codon Adaptation Index - a measure of directional synonymous codon usage bias, and its potential applications.** *Nucleic Acids Res* 1987, **15:**1281-1295.
20. Sharp PM, Devine KM: **Codon usage and gene expression level in *Dictyostelium discoideum*: highly expressed genes do 'prefer' optimal codons.** *Nucleic Acids Res* 1989, **17:**5029-5039.
21. Stenico M, Lloyd AT, Sharp PM: **Codon usage in *Caenorhabditis elegans*: delineation of translational selection and mutational biases.** *Nucleic Acids Res* 1994, **22:**2437-2446.
22. Sharp PM, Cowe E, Higgins DG, Shields DC, Wolfe KH, Wright F: **Codon usage patterns in *Escherichia coli*, *Bacillus subtilis*, *Saccharomyces cerevisiae*, *Schizosaccharomyces pombe*, *Drosophila melanogaster* and *Homo sapiens*; a review of the considerable within-species diversity.** *Nucleic Acids Res* 1988, **16:**8207-8211.
23. Sharp PM, Matassi G: **Codon usage and genome evolution.** *Curr Opin Genet Dev* 1994, **4:**851-860.
24. Bernardi G: **Compositional constraints and genome evolution.** *J Mol Evol* 1986, **24:**1-11.
25. Mouchiroud D, Gautier C: **Codon usage changes and sequence dissimilarity between human and rat.** *J Mol Evol* 1990, **31:**81-91.
26. Karlin S, Mrazek J: **What drives codon choices in human genes?** *J Mol Biol* 1996, **262:**459-472.
27. Antezana MA, Kreitman M: **The nonrandom location of synonymous codons suggests that reading frame-independent forces have patterned codon preferences.** *J Mol Evol* 1999, **49:**36-43.
28. Bernardi G: **Isochores and the evolutionary genomics of vertebrates.** *Gene* 2000, **241:**3-17.
29. Sueoka N: **Directional mutation pressure, selective constraints, and genetic equilibria.** *J Mol Evol* 1992, **34:**95-114.

30. Sueoka N: **Intrastrand parity rules of DNA base composition and usage biases of synonymous codons.** *J Mol Evol* 1995, **40:**318-325.
31. Lobry JR: **Asymmetric substitution patterns in the two DNA strands of bacteria.** *Mol Biol Evol* 1996, **13:**660-665.
32. Sueoka N: **Two aspects of DNA base composition: G+C content and translation-coupled deviation from intra-strand rule of A = T and G = C.** *J Mol Evol* 1999, **49:**49-62.
33. D'Onofrio G, Mouchiroud D, Aissani B, Gautier C, Bernardi G: **Correlations between the compositional properties of human genes, codon usage, and amino acid composition of proteins.** *J Mol Evol* 1991, **32:**504-510.
34. Collins DW, Jukes TH: **Relationship between G + C in silent sites of codons and amino acid composition of human proteins.** *J Mol Evol* 1993, **36:**201-213.
35. Lobry JR, Gautier C: **Hydrophobicity, expressivity and aromaticity are the major trends of amino-acid usage in 999 *Escherichia coli* chromosome-encoded genes.** *Nucleic Acids Res* 1994, **22:**3174-3180.
36. D'Onofrio G, Jabbari K, Musto H, Bernardi G: **The correlation of protein hydropathy with the base composition of coding sequences [published erratum appears in Gene 2000 Jan 11;241(2):341].** *Gene* 1999, **238:**3-14.
37. Adzhubei AA, Adzhubei IA, Krasheninnikov IA, Neidle S: **Non-random usage of 'degenerate' codons is related to protein three-dimensional structure.** *FEBS Lett* 1996, **399:**78-82.
38. Xie T, Ding D, Tao X, Dafu D: **The relationship between synonymous codon usage and protein structure [published erratum appears in FEBS Lett 1998 Oct 16;437(1-2):164].** *FEBS Lett* 1998, **434:**93-96.
39. Gupta SK, Majumdar S, Bhattacharya TK, Ghosh TC: **Studies on the relationships between the synonymous codon usage and protein secondary structural units.** *Biochem Biophys Res Commun* 2000, **269:**692-696.
40. Xia X: **Maximizing transcription efficiency causes codon usage bias.** *Genetics* 1996, **144:**1309-1320.
41. Berg OG, Kurland CG: **Growth rate-optimised tRNA abundance and codon usage.** *J Mol Biol* 1997, **270:**544-550.
42. Xia X: **How optimized is the translational machinery in *Escherichia coli*, *Salmonella typhimurium* and *Saccharomyces cerevisiae*?** *Genetics* 1998, **149:**37-44.
43. Lafay B, Atherton JC, Sharp PM: **Absence of translationally selected synonymous codon usage bias in *Helicobacter pylori*.** *Microbiology* 2000, **146:**851-860.
44. Bains W: **Codon distribution in vertebrate genes may be used to predict gene length.** *J Mol Biol* 1987, **197:**379-388.
45. Eyre-Walker A, Bulmer M: **Reduced synonymous substitution rate at the start of enterobacterial genes.** *Nucleic Acids Res* 1993, **21:**4599-4603.
46. Akashi H: **Synonymous codon usage in *Drosophila melanogaster*: natural selection and translational accuracy.** *Genetics* 1994, **136:**927-935.
47. Eyre-Walker A: **Synonymous codon bias is related to gene length in *Escherichia coli*: selection for translational accuracy?** *Mol Biol Evol* 1996, **13:**864-872.
48. Hasegawa M, Yasunaga T, Miyata T: **Secondary structure of MS2 phage RNA and bias in code word usage.** *Nucleic Acids Res* 1979, **7:**2073-2079.
49. Zama M: **Codon usage and secondary structure of mRNA.** *Nucleic Acids Symp Ser* 1990, **22:**93-94.
50. Gambari R, Nastruzzi C, Barbieri R: **Codon usage and secondary structure of the rabbit alpha-globin mRNA: a hypothesis.** *Biomed Biochim Acta* 1990, **49:**S88-93.
51. Huynen MA, Konings DA, Hogeweg P: **Equal G and C contents in histone genes indicate selection pressures on mRNA secondary structure.** *J Mol Evol* 1992, **34:**280-291.
52. Zama M: **Translational pauses during the synthesis of proteins and mRNA structure.** *Nucleic Acids Symp Ser* 1997, **37:**179-180.
53. Muto A, Osawa S: **The guanine and cytosine content of genomic DNA and bacterial evolution.** *Proc Natl Acad Sci USA* 1987, **84:**166-169.
54. Osawa S, Ohama T, Yamao F, Muto A, Jukes TH, Ozeki H, Umesono K: **Directional mutation pressure and transfer RNA in choice of the third nucleotide of synonymous two-codon sets.** *Proc Natl Acad Sci USA* 1988, **85:**1124-1128.
55. Foster PG, Jermiin LS, Hickey DA: **Nucleotide composition bias affects amino acid content in proteins coded by animal mitochondria.** *J Mol Evol* 1997, **44:**282-288.
56. Lobry JR: **Influence of genomic G+C content on average amino-acid composition of proteins from 59 bacterial species.** *Gene* 1997, **205:**309-316.
57. Gu X, Hewett-Emmett D, Li WH: **Directional mutational pressure affects the amino acid composition and hydrophobicity of proteins in bacteria.** *Genetica* 1998, 102-**103:**383-391.
58. Wilquet V, Van de Casteele M: **The role of the codon first letter in the relationship between genomic GC content and protein amino acid composition.** *Res Microbiol* 1999, **150:**21-32.
59. Oresic M, Shalloway D: **Specific correlations between relative synonymous codon usage and protein secondary structure.** *J Mol Biol* 1998, **281:**31-48.
60. Andersson SG, Kurland CG: **Codon preferences in free-living microorganisms.** *Microbiol Rev* 1990, **54:**198-210.
61. Nakamura Y, Tabata S: **Codon-anticodon assignment and detection of codon usage trends in seven microbial genomes.** *Microb Comp Genomics* 1997, **2:**299-312.
62. Kanaya S, Yamada Y, Kudo Y, Ikemura T: **Studies of codon usage and tRNA genes of 18 unicellular organisms and quantification of *Bacillus subtilis* tRNAs: gene expression level and species-specific diversity of codon usage based on multivariate analysis.** *Gene* 1999, **238:**143-155.
63. Sogin ML, Elwood HJ, Gunderson JH: **Evolutionary diversity of eukaryotic small-subunit rRNA genes.** *Proc Natl Acad Sci USA* 1986, **83:**1383-1387.
64. Press WH, Teukolsky SA, Vettering WT, Flannery BP: *Numerical Recipes in C*. 2nd edn. New York: Cambridge University Press, 1992
65. Karkas JD, Rudner R, Chargaff E: **Separation of *B. subtilis* DNA into complementary strands. II. Template functions and composition as determined by transcription with RNA polymerase.** *Proc Natl Acad Sci USA* 1968, **60:**915-920.
66. Rudner R, Karkas JD, Chargaff E: **Separation of *B. subtilis* DNA into complementary strands. III. Direct analysis.** *Proc Natl Acad Sci USA* 1968, **60:**921-922.
67. Lobry JR: **Properties of a general model of DNA evolution under no-strand-bias conditions [published erratum appears in J Mol Evol 1995 Nov;41(5):680].** *J Mol Evol* 1995, **40:**326-330.
68. Fisher RA: *The Genetical Theory of Natural Selection*. 2nd edn. New York: Dover Publications, 1958.
69. Freeland SJ, Hurst LD: **The genetic code is one in a million.** *J Mol Evol* 1998, **47:**238-248.
70. Woese CR, Dugre DH, Saxinger WC, Dugre SA: **The molecular basis for the genetic code.** *Proc Natl Acad Sci USA* 1966, **55:**966-974.
71. Woese CR, Dugre DH, Dugre SA, Kondo M, Saxinger WC: **On the fundamental nature and evolution of the genetic code.** *Cold Spring Harb Symp Quant Biol* 1966, **31:**723-736.
72. Graur D, Li W: *Fundamentals of Molecular Evolution*. 2nd edn. Sunderland, MA: Sinauer, 2000.
73. Sokal RR, Rohlf FJ: *Biometry: The Principles and Practice of Statistics in Biological Research*. 3rd edn. New York: W.H. Freeman and Company, 1995.
74. Takahata N, Kimura M: **A model of evolutionary base substitutions and its application with special reference to rapid change of pseudogenes.** *Genetics* 1981, **98:**641-657.
75. Ohta T, Gillespie JH: **Development of neutral and nearly neutral theories.** *Theor Popul Biol* 1996, **49:**128-142.
76. Nakamura Y, Gojobori T, Ikemura T: **Codon usage tabulated from international DNA sequence databases: status for the year 2000.** *Nucleic Acids Res* 2000, **28:**292.

4 Conclusions/Research Summary

While the precise evolutionary history of the genetic code remains obscure, the work presented here makes the overall picture somewhat clearer. The most fundamental conclusion is that a pluralistic account of code evolution is required: faced with compelling statistical evidence for more than one mechanism, it is prudent to question whether the mechanisms are mutually exclusive, or if, instead, they may have acted together in shaping modern codon assignments. Research on code evolution has long been dominated by explanations that assume that a single mechanism must have led to all modern codon assignments, erasing any traces of any earlier influences. By shifting the debate from the existence to the relative importance of the many different, plausible mechanisms, it may be possible to make rapid progress towards uncovering the order in which amino acids were added to the code and the influences that led to specific codon assignments.

When I began this project, I assumed that the code was probably the exclusive result of stereochemical constraints. Given an experimental technique, SELEX, and a statistical technique, tests for overabundance of codons in RNA molecules selected to bind specific peptide targets, I thought it likely that the 'true' primordial code would soon be revealed. It soon became clear that the situation was more complex: aptamers can be surprisingly tricky to select, and statistics uninformed by biochemistry can produce misleading conclusions. However, increasingly powerful statistical techniques can extract new information from data that are already in the public domain: with genome projects spending billions of dollars on sequencing, it makes sense to examine what is already available.

Thus, I have tried to make research into code evolution less speculative than has historically been the case by applying quantitative techniques to actual data, reflecting Crick's (1968) desire that speculations about the code be driven by empiricism rather than theory. I was fortunate to enter the field at a time when the first amino acid aptamers had been selected; sequencing efforts had revealed a wide variety of variant genetic codes, and provided a range of complete mitochondrial genomes to work with; fundamental components of the translation apparatus were being sequenced from a bewildering array of species; and computer resources were sufficiently available and inexpensive to allow extensive statistical testing by MCMC (Markov Chain Monte Carlo) and other techniques.

Consequently, it has been possible to test almost every well-defined model of code evolution in a rigorous fashion, and, in many cases, to reconcile the results of different models for which there is strong independent support. The first substantial contribution along these lines was to test whether arginine aptamers showed statistical overrepresentation of Arg codons, or of other specific triplet motifs, at their binding sites. At the time, it seemed likely that information from binding sites for single amino acids would not provide enough resolution to test whether codons were specifically represented; however, a simple test for independence showed that, in fact, arginine binding sites are largely composed of Arg codons, but of no other triplet motifs (Chapter 2.2), as proposed by Prof. Yarus a decade earlier after the discovery that the Group I intron specifically binds arginine using its codons. Testing the robustness of the result required more effort. Base-specific bias at binding sites, nonrandomness of binding sites, and differences in ease of detection of chemical modification/interference at particular nucleotides, could be accounted for by Monte Carlo simulations that explicitly compared the actual sequences to randomized sequences with exactly the same compositions. Another possibility was that the result was due to the specific sequences chosen for characterization. I was able to test this by exhaustively analyzing every combination of sequences from the reported aptamer pools, which showed that the result was remarkably robust to assumptions about binding sites and sequence conservation: furthermore, the association was clearly between

arginine and its codon set, rather than groups of related codons in general, and did not hold for aptamers for the related amino acid citrulline (Chapter 2.3).

The strength of the association between arginine codons and arginine binding sites clearly implied that stereochemistry had played a role in structuring the code, but could it be the only explanation? While I was testing stereochemical models in Princeton, Dr. Freeland was in Cambridge, accumulating evidence that the genetic code minimizes genetic errors far better than would a code chosen at random, that this result held for two very different measures of amino acid similarity, and that historical constraints on the first position base could not explain the pattern. Given evidence strongly in favor of stereochemical *and* adaptive constraints, and that neither of us was about to abandon our favored explanation, the most expedient solution was to assume that both processes, and perhaps others, really did play a role in shaping the modern code structure: the research program was then to determine the relative contribution of each causal influence, rather than to deny the efficacy of all but one of them. This pluralistic approach allowed us to write an influential review in TiBS (Chapter 1.5), which framed the rest of the research program: statistical tests of explicit hypotheses, taking into account all the biochemical detail available.

The evolution of variant genetic codes had received much attention a decade earlier, but progress had been rapid and no comprehensive review of either the phylogeny or biochemical mechanisms of variant genetic codes existed. Without such a catalog of the actual distribution and mechanism of code variation, it was difficult to test ideas about how and why these nonstandard codes evolved. The pattern of changes (Chapter 3.1) turned out to be remarkably interesting, rife with convergent evolution that suggested that specific adaptive or biochemical constraints were at work, limiting the range of variability. Most variant codes are found in mitochondria, and by 2000 a large number of complete mitochondrial genomes had been sequenced. The three models that had been proposed to explain recent code diversification, Osawa and Jukes's Codon Capture, Schultz and Yarus's Codon Ambiguity, and Andersson and Kurland's Genome Reduction, each made specific predictions about which changes would occur most often and/or which types of genomes would be most likely to have variant codes. In order to test these hypotheses quantitatively, I downloaded the complete set of mitochondrial genomes from GenBank, and wrote software to analyze them. As in the evolution of the canonical code, there seemed to be truth to more than one model. Surprisingly, however, the very plausible genome reduction model did not explain the variant codes in mitochondria at all (Chapter 3.2).

In testing the Codon Capture hypothesis, it was necessary to test whether particular codons really did vary systematically with changes in genome composition. The literature on synonymous codon usage is vast and unruly: consequently, I expected that every organism would have its own idiosyncratic codon usage, undermining support for the directional mutation pressure required by Codon Capture. To my surprise, such an underlying principle did exist: codon and amino acid usage is almost quantitatively explained by the base composition when position- and base-dependent purifying selection is taken into account, and this selection can be derived from first principles from the average effect of substituting one amino acid for another under a particular type of mutation. This makes it possible to predict codon and amino acid usage with a considerable degree of accuracy. The surprising result is that amino acid and codon usage vary over a vast range, but this variation (which natural selection could certainly act on if it mattered) can all be explained by neutral models acting at the level of single nucleotides, ignoring protein (and even amino acid and codon) identity (Chapter 3.4). Although this section comprises a minor component of my thesis, I see it as the most substantial contribution, and perhaps the one most likely to yield useful technological developments.

Most research into the molecular mechanisms leading to variant codes had focused on tRNAs, which are relatively easy to isolate and characterize. However, when stop codons change, the release factors that recognize these codons must also be altered in order to prevent premature chain termination. While reviewing biochemical mechanisms of the

evolution of specific components of the translation apparatus for *Cell* (Chapter 1.4), I noticed that the eRF1 (eukaryotic release factor 1) sequence of *Tetrahymena*, which has a variant genetic code, had a change in a critical domain otherwise absolutely conserved even between eukaryotes and archaea. Subsequent sequencing of release factors from other ciliates with and without variant genetic codes showed that the relevant change was not as simple as a single amino acid substitution. In order to find out what changes were important, I developed novel statistical methodology to test association between changes in sequences and changes in traits on a phylogeny with reconstructed ancestral sequences (Chapter 3.3). Remarkably, this test showed that the domains known from the crystal structure to be important in binding in humans were the areas that specifically changed along with the genetic code. Furthermore, several of the amino acid changes that occurred only in ciliates with variant genetic codes were the same as yeast mutants reported to confer suppressor activity. Thus, we were able to use 'Nature's genetic screen' to find functionally important regions of the enzyme *in silico*.

The existence of variant genetic codes raises a fundamental question: why do most organisms use the canonical code, rather than one of the allowed variants? In particular, are these variants better in some way than the canonical code? In order to address this question, and the more general question of what environment the code might be adapted to, I used Dr. Freeland's methodology to test the apparent optimality of the code with a variety of chemical measures and against the known variants. Although the results were somewhat inconclusive, it appears that the code may have been optimized very early, perhaps based on the properties of the free amino acids rather than the properties of the side-chains in the context of modern proteins. I also found that one of the main measures used to support estimates of code optimality, the PAM74-100 matrix, is contaminated with the code's structure and so cannot be used to substantiate these claims, and that a re-measurement of Polar Requirement with modern thin-layer chromatography techniques leads to a highly correlated measure that makes the code look far less optimal (Chapter 2.5). More investigation of the specific types of measurements that support the adaptive results is clearly required.

Finally, the stereochemical result (statistical interaction between codons and binding sites) is open to reinterpretation as new data arrive. A reanalysis of the associations using data from the six amino acids for which aptamers are now available (Chapter 2.1) reconfirms the result for arginine. Overall, however, the statistical evidence for an association with the anticodons is just as convincing as that for an association with the codons. A clear association in one direction or the other would allow us to infer which component(s) of the translation apparatus these primordial sites had evolved into (Chapter 2.3); however, such conclusions are premature. It is possible that even within the specific categories of mechanism, adaptation and stereochemistry, several interacting forces shaped modern codon assignments.

It is clear that the code still conceals many of the secrets of its evolution. However, we now have a statistical and experimental framework that should allow us to determine, perhaps in the near future, the relative roles of selection, chemistry and history in producing both the canonical genetic code and the diversity of modern variants.

5 References

Adzhubei, A. A., I. A. Adzhubei, I. A. Krasheninnikov and S. Neidle (1996). “Non-random usage of 'degenerate' codons is related to protein three- dimensional structure.” FEBS Lett **399**(1-2): 78-82.

Agou, F., S. Quevillon, P. Kerjan and M. Mirande (1998). “Switching the amino acid specificity of an aminoacyl-tRNA synthetase.” Biochemistry **37**(32): 11309-14.

Aita, T., S. Urata and Y. Husimi (2000). “From amino acid landscape to protein landscape: analysis of genetic codes in terms of fitness landscape.” J Mol Evol **50**(4): 313-23.

Akashi, H. (1994). “Synonymous codon usage in Drosophila melanogaster: natural selection and translational accuracy.” Genetics **136**(3): 927-35.

Albert, B., B. Godelle, A. Atlan, R. De Paepe, et al. (1996). “Dynamics of Plant Mitochondrial Genome: Model of a Three-Level Selection Process.” Genetics **144**: 369-382.

Alberti, S. (1997). “The Origin of the Genetic Code and Protein Synthesis.” J Mol Evol **45**: 352-358.

Alberts, B. M. (1986). “The Function of the Hereditary Materials: Biological Catalyses Reflect the Cell's Evolutionary History.” American Zoologist **26**: 781-796.

Alff-Steinberger, C. (1969). “The genetic code and error transmission.” Proc. Natl. Acad. Sci. U. S. A. **64**: 584-591.

Alfonzo, J. D., V. Blanc, A. M. Estevez, M. A. Rubio, et al. (1999). “C to U editing of the anticodon of imported mitochondrial tRNA(Trp) allows decoding of the UGA stop codon in Leishmania tarentolae.” Embo J **18**(24): 7056-62.

Amirnovin, R. (1997). “An Analysis of the Metabolic Theory of the Origin of the Genetic Code.” J Mol Evol **44**: 473-476.

Amirnovin, R. and S. L. Miller (1999). “Response.” J Mol Evol **48**: 253-255.

Andersson, G. E. and C. G. Kurland (1991). “An extreme codon preference strategy: codon reassignment.” Mol Biol Evol **8**(4): 530-44.

Andersson, S. G. and C. G. Kurland (1990). “Codon preferences in free-living microorganisms.” Microbiol Rev **54**(2): 198-210.

Andersson, S. G. and C. G. Kurland (1995). “Genomic evolution drives the evolution of the translation system.” Biochem Cell Biol **73**(11-12): 775-87.

Andersson, S. G. and C. G. Kurland (1998). “Reductive evolution of resident genomes.” Trends Microbiol **6**(7): 263-8.

Antezana, M. A. and M. Kreitman (1999). “The nonrandom location of synonymous codons suggests that reading frame- independent forces have patterned codon preferences.” J Mol Evol **49**(1): 36-43.

Antillon, A. and I. Ortega-Blake (1985). “A group theory analysis of the ambiguities in the genetic code: on the existence of a generalized genetic code.” J Theor Biol **112**(4): 757-69.

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

Arkov, A. L., D. V. Freistroffer, M. Ehrenberg and E. J. Murgola (1998). “Mutations in RNAs of both ribosomal subunits cause defects in translation termination.” Embo J **17**(5): 1507-14.

Arnez, J. G. and D. Moras (1997). “Structural and functional considerations of the aminoacylation reaction.” Trends Biochem Sci **22**(6): 211-6.

Arnez, J. G. and T. A. Steitz (1996). “Crystal structures of three misacylating mutants of Escherichia coli glutaminyl-tRNA synthetase complexed with tRNA(Gln) and ATP.” Biochemistry **35**(47): 14725-33.

Ashraf, S. S., E. Sochacka, R. Cain, R. Guenther, et al. (1999). “Single atom modification (O-->S) of tRNA confers ribosome binding.” RNA **5**(2): 188-94.

Bains, W. (1987). “Codon distribution in vertebrate genes may be used to predict gene length.” J Mol Biol **197**(3): 379-88.

Balasubramanian, R. (1982). “Origin of life: A hypothesis for the origin of adaptor-mediated ordered synthesis of proteins and an explanation for the choice of terminating codons in the genetic code.” Bio Systems **15**: 99-104.

Barrell, B. G., A. T. Bankier and J. Drouin (1979). “A different genetic code in human mitochondria.” Nature **282**(5735): 189-94.

Bashford, J. D., I. Tsohantjis and P. D. Jarvis (1998). “A supersymmetric model for the evolution of the genetic code.” Proc Natl Acad Sci U S A **95**(3): 987-92.

Baumann, U. and J. Oró (1993). “Three stages in the evolution of the genetic code.” Bio Systems **29**: 133-141.

Becker, H. D. and D. Kern (1998). “Thermus thermophilus: a link in evolution of the tRNA-dependent amino acid amidation pathways.” Proc Natl Acad Sci U S A **95**(22): 12832-7.

Becker, H. D., H. Roy, L. Moulinier, M. H. Mazauric, et al. (2000). “Thermus thermophilus contains an eubacterial and an archaebacterial aspartyl-tRNA synthetase.” Biochemistry **39**(12): 3216-30.

Benner, S. A., R. K. Allemann, A. D. Ellington, L. Ge, et al. (1987). “Natural Selection, Protein Engineering, and the Last Riboorganism: Rational Model Building in Biochemistry.” Cold Spring Harbor Symposia on Quantitative Biology **LII**: 53-63.

Benner, S. A., M. A. Cohen and G. H. Gonnet (1994). “Amino acid substitution during functionally constrained divergent evolution of protein sequences.” Protein Eng **7**(11): 1323-32.

Benner, S. A., A. D. Ellington and A. Tauer (1989). “Modern metabolism as a palimpsest of the RNA world.” Proc Natl Acad Sci USA **86**: 7054-7058.

Berg, O. G. and C. G. Kurland (1997). “Growth rate-optimised tRNA abundance and codon usage.” J Mol Biol **270**(4): 544-50.

Bergman, N. H., W. K. Johnston and D. P. Bartel (2000). “Kinetic framework for ligation by an efficient RNA ligase ribozyme.” Biochemistry **39**(11): 3115-23.

Bernardi, G. (1986). “Compositional constraints and genome evolution.” J Mol Evol **24**(1-2): 1-11.

Bernardi, G. (2000). “Isochores and the evolutionary genomics of vertebrates.” Gene **241**(1): 3-17.

Bertram, G., H. A. Bell, D. W. Ritchie, G. Fullerton, et al. (2000). “Terminating eukaryote translation: domain 1 of release factor eRF1 functions in stop codon recognition.” RNA **6**(9): 1236-47.

Björk, G. R. (1998). Modified Nucleosides at Position 35 and 37 of tRNAs and Their Predicted Coding Capacities. Modification and Editing of RNA. H. Grosean and R. Benne. Washington DC, ASM Press.

Blalock, J. E. and E. M. Smith (1984). "Hydropathic Anti-Complementarity of Amino Acids Based on the Genetic Code." Biochem Biophys Res Comm **121**(1): 203-207.

Bonitz, S. G., R. Berlani, G. Coruzzi, M. Li, et al. (1980). "Codon recognition rules in yeast mitochondria." Proc Natl Acad Sci U S A **77**(6): 3167-70.

Boren, T., P. Elias, T. Samuelsson, C. Claesson, et al. (1993). "Undiscriminating codon reading with adenosine in the wobble position." J Mol Biol **230**(3): 739-49.

Brown, G. G. and M. V. Simpson (1982). "Novel features of animal mtDNA evolution as shown by sequences of two rat cytochrome oxidase subunit II genes." Proc Natl Acad Sci U S A **79**(10): 3246-50.

Brown, J. R. and W. F. Doolittle (1999). "Gene descent, duplication, and horizontal transfer in the evolution of glutamyl- and glutaminyl-tRNA synthetases." J Mol Evol **49**(4): 485-95.

Brown, J. R., F. T. Robb, R. Weiss and W. F. Doolittle (1997). "Evidence for the early divergence of tryptophanyl- and tyrosyl-tRNA synthetases." J Mol Evol **45**(1): 9-16.

Brown, J. W. and N. R. Pace (1992). "Ribonuclease P RNA and protein subunits from bacteria." Nucleic Acids Res **20**: 1451-1456.

Buckingham, R. H., G. Grentzmann and L. Kisselev (1997). "Polypeptide chain release factors." Mol Microbiol **24**(3): 449-56.

Budisa, N., C. Minks, S. Alefelder, W. Wenger, et al. (1999). "Toward the experimental codon reassignment in vivo: protein building with an expanded amino acid repertoire." FASEB J **13**(1): 41-51.

Budisa, N., C. Minks, F. J. Medrano, J. Lutz, et al. (1998). "Residue-specific bioincorporation of non-natural, biologically active amino acids into proteins as possible drug carriers: Structure and stability of the *per*-thiaproline mutant of annexin V." Proc Natl Acad Sci **95**: 455-459.

Bulmer, M. (1987). "Coevolution of codon usage and transfer RNA abundance." Nature **325**(6106): 728-30.

Burgstaller, P., M. Kochoyan and M. Famulok (1995). "Structural probing and damage selection of citrulline- and arginine-specific RNA aptamers identify base positions required for binding." Nucleic Acids Res **23**: 4769-4776.

Cairns-Smith, A. G. (1982). Genetic Takeover and the Mineral Origins of Life. Cambridge, Cambridge University Press.

Caron, F. (1990). "Eucaryotic codes." Experientia **46**(11-12): 1106-17.

Castresana, J., G. Feldmaier-Fuchs and S. Paabo (1998). "Codon reassignment and amino acid composition in hemichordate mitochondria." Proc Natl Acad Sci U S A **95**(7): 3703-7.

Cavalier-Smith, T. (1993). "Kingdom protozoa and its 18 phyla." Microbiol Rev **57**(4): 953-94.

Cech, T. R. (1986). "A model for the RNA-catalysed replication of RNA." Proc Natl Acad Sci USA **83**: 4360-4363.

Cech, T. R. (1993). Structure and Mechanism of the Large Catalytic RNAs: Group I and Group II Introns and Ribonuclease P. The RNA World. R. F. Gesteland and J. F. Atkins. New York, Cold Spring Harbor Laboratory Press**:** 239-269.

Cech, T. R. (2000). “Structural biology. The ribosome is a ribozyme.” Science **289**(5481): 878-9.

Cermakian, N. and C. R (1998). Modified Nucleosides Always Were: an Evolutionary Model. Modification and Editing of RNA. H. Grosean and R. Benne. Washington DC, ASM Press.

Chaley, M. B., E. V. Korotkov and D. A. Phoenix (1999). “Relationships among isoacceptor tRNAs seems to support the coevolution theory of the origin of the genetic code.” J Mol Evol **48**(2): 168-77.

Chihade, J. W. and P. Schimmel (1999). “Assembly of a catalytic unit for RNA microhelix aminoacylation using nonspecific RNA binding domains.” Proc Natl Acad Sci U S A **96**(22): 12316-21.

Ciesiolka, J., M. Illangasekare, I. Majerfeld, T. Nickles, et al. (1996). “Affinity Selection-Amplification from Randomized Ribooligonucleotide Pools.” Methods in Enzymology **267**: 315-335.

Clark-Walker, G. D. and G. F. Weiller (1994). “The structure of the small mitochondrial DNA of Kluyveromyces thermotolerans is likely to reflect the ancestral gene order in fungi.” J Mol Evol **38**(6): 593-601.

Collins, D. W. and T. H. Jukes (1993). “Relationship between G + C in silent sites of codons and amino acid composition of human proteins.” J Mol Evol **36**(3): 201-13.

Commans, S. and A. Bock (1999). “Selenocysteine inserting tRNAs: an overview.” FEMS Microbiol Rev **23**(3): 335-51.

Connell, G. J., M. Illangsekare and M. Yarus (1993). “Three Small Ribooligonucleotides with Specific Arginine Sites.” Biochemistry **32**: 5497-5502.

Connell, G. J. and M. Yarus (1994). “RNAs with Dual Specificity and Dual RNAs with Similar Specificity.” Science **264**: 1137-1141.

Corliss, J. O. (1979). The ciliated protozoa. Characterization, classification, and guide to the literature. London, Pergamon Press.

Crick, F. H. (1966). “Codon-anticodon pairing: the wobble hypothesis.” J. Mol. Biol. **19**(2): 548-555.

Crick, F. H. C. (1957). “The structure of nucleic acids and their role in protein synthesis.” Biochem. Soc. Symp. **14**: 25-26.

Crick, F. H. C. (1963). “The Recent Excitement in the Coding Problem.” Progress in Nucleic Acids **1**: 163-217.

Crick, F. H. C. (1967). “An Error in Model Building.” Nature **213**: 798.

Crick, F. H. C. (1967). “Origin of the Genetic Code.” Nature **213**: 119.

Crick, F. H. C. (1968). “The Origin of the Genetic Code.” J. Mol. Biol. **38**: 367-379.

Cullman, G. and J. Labouygues (1983). “Noise Immunity of the Genetic Code.” Bio Systems **16**: 9-29.

Cullman, G. and J. Labouygues (1987). “The logic of the genetic cdoe.” Math Model **8**: 643-646.

Curran, J. F. (1995). “Decoding with the A:I wobble pair is inefficient.” Nucleic Acids Res **23**(4): 683-8.

Curran, J. F. (1998). Modified Nucleosides in Translation. Modification and Editing of RNA. H. Grosean and R. Benne. Washington DC, ASM Press.

Curtis, E. A. and L. F. Landweber (1999). “Evolution of gene scrambling in ciliate micronuclear genes.” Ann N Y Acad Sci **870**: 349-50.

Cusack, S. (1993). “Sequence, structure and evolutionary relationships between class 2 aminoacyl-tRNA synthetases: an update.” Biochimie **75**(12): 1077-81.

D'Onofrio, G., K. Jabbari, H. Musto and G. Bernardi (1999). “The correlation of protein hydropathy with the base composition of coding sequences [published erratum appears in Gene 2000 Jan 11;241(2):341].” Gene **238**(1): 3-14.

D'Onofrio, G., D. Mouchiroud, B. Aissani, C. Gautier, et al. (1991). “Correlations between the compositional properties of human genes, codon usage, and amino acid composition of proteins.” J Mol Evol **32**(6): 504-10.

Dai, X., A. D. Mesmaeker and G. F. Joyce (1995). “Cleavage of an Amide Bond by a Ribozyme.” Science **267**: 237-240.

Davies, J., W. Gilbert and L. Gorini (1964). “Streptomycin, suppression, and the code.” Proc Natl Acad Sci U S A **51**: 883-890.

Davis, B. K. (1999). “Evolution of the genetic code.” Prog Biophys Mol Biol **72**(2): 157-243.

Davydov, O. V. (1995). Problem of the genetic code structure: new data and perspectives. Evolutionary Biology and Related Areas of Physicochemical Biology. B. F. Poglazov, B. I. Kurganov, M. S. Kritsky and K. L. Gladilin. Moscow, Bach Institute of Biochemistry, Russian Academy of Sciences**:** 283-295.

Davydov, O. V. (1996). “Internal Logic of the Genetic Encoding: End-Atom Rules of Doublet Composition.” ISSOL Newsletter **23**(1): 12.

Davydov, O. V. (1998). “Amino acid contribution to the genetic code structure: end-atom chemical rules of doublet composition.” J Theor Biol **193**(4): 679-90.

Dawkins, R. (1976). The Selfish Gene. Oxford, Oxford University Press.

de Duve, C. (1988). “The second genetic code.” Nature **333**: 117-118.

de Duve, C. (1995). Vital Dust. New York, Basic Books.

De Giorgi, C., F. De Luca and C. Saccone (1991). “Mitochondrial DNA in the sea urchin Arbacia lixula: nucleotide sequence differences between two polymorphic molecules indicate asymmetry of mutations.” Gene **103**(2): 249-52.

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. (1989). “Some Aspects of the Organization and Evolution of the Genetic Code.” J Mol Evol **29**: 191-201.

Di Giulio, M. (1991). “On the Relationships between the Genetic Code Coevolution Hypothesis and the Physicochemical Hypothesis.” Z Naturforsch **46c**: 305-312.

Di Giulio, M. (1993). “Origin of Glutaminyl-tRNA Synthetase: An Example of Palimpsest?” J Mol Evol **37**: 5-10.

Di Giulio, M. (1997). “On the Origin of the Genetic Code.” J theor Biol **187**: 573-581.

Di Giulio, M. (1997). “On the RNA World: Evidence in Favor of an Early Ribonucleopeptide World.” L Mol Evol **45**: 571-578.

Di Giulio, M. (1998). “The historical factor: the biosynthetic relationships between amino acids and their physiochemical properties in the origin of the genetic code.” J Mol Evol **46**: 615-621.

Di Giulio, M. (1999). “The coevolution theory of the origin of the genetic code.” J Mol Evol **48**(3): 253-5.

Di Giulio, M. (2000). “Genetic code origin and the strength of natural selection.” J Theor Biol **205**(4): 659-61.

Di Giulio, M. (2000). “The origin of the genetic code.” Trends Biochem Sci **25**(2): 44.

Di Giulio, M. (2001). “The Origin of the Genetic Code cannot be Studied using Measurements based on the PAM Matrix because this Matrix Reflects the Code Itself, Making any such Analyses Tautologous.” J Theor Biol **208**(2): 141-144.

Di Giulio, M., M. R. Capobianco and M. Medugno (1994). “On the Optimization of the Physicochemical DIstances between Amino Acids in the Evolution of the Genetic Code.” J theor Biol **168**: 43-51.

Di Giulio, M. and M. Medugno (2000). “The robust statistical bases of the coevolution theory of genetic code origin.” J Mol Evol **50**(3): 258-63.

Dick, T. P. and W. W. A. Schamel (1995). “Molecular Evolution of Transfer RNA from Two Precursor Hairpins: Implications for the Origin of Protein Synthesis.” J Mol Evol **41**: 1-9.

Dillon, L. S. (1973). “The Origins of the Genetic Code.” Bot Rev **39**: 301-345.

Dontsova, M., L. Frolova, J. Vassilieva, W. Piendl, et al. (2000). “Translation termination factor aRF1 from the archaeon Methanococcus jannaschii is active with eukaryotic ribosomes.” FEBS Lett **472**(2-3): 213-6.

Dufton, M. J. (1983). “The significance of redundancy in the genetic code.” J Theor Biol **102**(4): 521-6.

Dufton, M. J. (1997). “Genetic code synonym quotas and amino acid complexity: cutting the cost of proteins?” J Theor Biol **187**(2): 165-73.

Dunnill, P. (1966). “Triplet Nucleotide—Amino Acid Pairing: A Stereochemical Basis for the Division between Protein and Nonprotein Amino Acids.” Nature **210**: 1267-1268.

Dyson, F. (1985). Origins of life. Cambridge, Cambridge University Press.

Edmonds, C. G., P. F. Crain, R. Gupta, T. Hashizume, et al. (1991). “Posttranscriptional modification of tRNA in thermophilic archaea (Archaebacteria).” J Bacteriol **173**(10): 3138-48.

Egholm, M., O. Buchardt, L. Cristensen, C. Behrens, et al. (1993). “PNA hybridizes to complementary oligonucleotides obeying the Watson-Crick hydrogen-bonding rules.” Nature **365**: 566-568.

Ehara, M., Y. Inagaki, K. I. Watanabe and T. Ohama (2000). “Phylogenetic analysis of diatom coxI genes and implications of a fluctuating GC content on mitochondrial genetic code evolution.” Curr Genet **37**(1): 29-33.

Eigen, M. (1971). “Self-organization of matter and the evolution of biological macromolecules.” Naturwissenschaften **58**: 465-522.

Eigen, M., W. Gardiner, P. Schuster and R. Winkler-Oswatitsch (1981). “The origin of genetic information.” Sci Am **244**: 88-118.

Eigen, M., B. F. Lindemann, M. Tietze, R. Winkler-Oswatitsch, et al. (1989). “How Old Is the Genetic Code? Statistical Geometry of tRNA Provides an Answer.” Science **244**: 673-679.

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

Eigen, M. and R. Winkler-Oswatitsch (1981). “Transfer-RNA, an early gene?” Naturwissenschaften **68**: 282-292.

Ekland, E. H. and D. P. Bartel (1996). “RNA-catalysed RNA polymerization using nucleoside triphosphates.” Nature **382**(6589): 373-6.

Ellington, A. D., M. Khrapov and C. A. Shaw (2000). “The scene of a frozen accident.” RNA **6**(4): 485-98.

Ellington, A. D. and J. W. Szostak (1990). “*In vitro* selection of RNA molecules that bind specific ligands.” Nature **346**: 818-822.

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

Epstein, C. J. (1967). “Non-randomness of Amino-acid Changes in the Evolution of Homologous Proteins.” Nature **215**: 355-359.

Eschenmoser, A. (1999). “Chemical etiology of nucleic acid structure.” Science **284**(5423): 2118-24.

Eyre-Walker, A. (1996). “Synonymous codon bias is related to gene length in Escherichia coli: selection for translational accuracy?” Mol Biol Evol **13**(6): 864-72.

Eyre-Walker, A. and M. Bulmer (1993). “Reduced synonymous substitution rate at the start of enterobacterial genes.” Nucleic Acids Res **21**(19): 4599-603.

Famulok, M. (1994). “Molecular Recognition of Amino Acids by RNA-Aptamers: An L-Citrulline Binding RNA Motif and Its Evolution into an L-Arginine Binder.” J Am Chem Soc **116**: 1698-1706.

Famulok, M. and J. W. Szostak (1992). “Stereospecific Recognition of Tryptophan Agarose by *In Vitro* Selected RNA.” Journal of the American Chemical Society **114**: 3990-3991.

Fan, P., A. K. Suri, R. Fiala, D. Live, et al. (1996). “Molecular recognition in the FMN-RNA aptamer complex.” J Mol Biol **258**: 480-500.

Fauchere, J. and V. Pliska (1983). “Hydrophobic parameters pi of amino acid side chains from the partitioning of N-acetyl-amino-acid amides.” Eur J Med Chem **18**(4): 369-375.

Fendler, J. H., F. Nome and J. Nagyvary (1975). “Compartmentalization of Amino Acids in Surfactant Aggregates.” J Mol Evol **6**: 215-232.

Ferris, J. P., P. C. Joshi, E. H. Edelson and J. G. Lawless (1978). “HCN: A Plausible Source of Purines, Pyrimidines and Amino Acids on the Primitive Earth.” J Mol Evol **11**: 293-311.

Figureau, A. (1987). “Information theory and the genetic code.” Orig Life **17**: 439-449.

Figureau, A. (1989). “Optimization and the genetic code.” Orig Life Evol Biosph **19**: 57-67.

Figureau, A. and M. Pouzet (1984). “Genetic code and optimal resistance to the effect of mutations.” Orig Life Evol Biosph **14**: 579-588.

Fisher, R. A. (1958). The Genetical Theory of Natural Selection. New York, Dover Publications.

Fitch, W. M. (1966). “The relation between frequencies of amino acids and ordered trinucleotides.” J Mol Biol **16**: 1-8.

Fitch, W. M. and K. Upper (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.

Forchhammer, K., K. Boesmiller and A. Bock (1991). “The function of selenocysteine synthase and SELB in the synthesis and incorporation of selenocysteine.” Biochimie **73**(12): 1481-6.

Foster, P. G., L. S. Jermiin and D. A. Hickey (1997). “Nucleotide composition bias affects amino acid content in proteins coded by animal mitochondria.” J Mol Evol **44**(3): 282-8.

Fox, S. W. and H. Dose (1977). Molecular Evolution and the Origin of Life. New York, Academic Press.

Freeland, S. J. and L. D. Hurst (1998). “The genetic code is one in a million.” J Mol Evol **47**(3): 238-248.

Freeland, S. J. and L. D. Hurst (1998). “Load minimization of the code: history does not explain the pattern.” Proc Roy Soc Lond B **265**: 1-9.

Freeland, S. J., R. D. Knight and L. F. Landweber (1999). “Do proteins predate DNA?” Science **286**(5440): 690-2.

Freeland, S. J., R. D. Knight and L. F. Landweber (2000). “Measuring adaptation within the genetic code.” Trends Biochem Sci **25**(2): 44-5.

Freeland, S. J., R. D. Knight, L. F. Landweber and L. D. Hurst (2000). “Early Fixation of an Optimal Genetic Code.” Mol Biol Evol **17**(4): 511-518.

Gambari, R., C. Nastruzzi and R. Barbieri (1990). “Codon usage and secondary structure of the rabbit alpha-globin mRNA: a hypothesis.” Biomed Biochim Acta **49**(2-3): S88-93.

Gamow, G. (1954). “Possible mathematical relation between deoxyribonucleic acid and protein.” Kgl Dansk Videnskab Selskab Biol Medd **22**: 1-13.

Gánti, T. (1975). “Organisation of chemical reactions into dividing and metabolizing units: the chemotons.” BioSystems **7**: 189-195.

Gatti, D. L. and A. Tzagoloff (1991). “Structure and evolution of a group of related aminoacyl-tRNA synthetases.” J Mol Biol **218**(3): 557-68.

Geiger, A., P. Burgstaller, H. von der Eltz, A. Roeder, et al. (1996). “RNA aptamers that bind L-arginine with sub-micromolar dissociation constants and high enantioselectivity.” Nucleic Acids Research **24**(6): 1029-1036.

Giege, R., M. Sissler and C. Florentz (1998). “Universal rules and idiosyncratic features in tRNA identity.” Nucleic Acids Res **26**(22): 5017-35.

Gilbert, W. (1986). “The RNA world.” Nature **319**: 618.

Glasner, M. E., C. C. Yen, E. H. Ekland and D. P. Bartel (2000). “Recognition of nucleoside triphosphates during RNA-catalyzed primer extension.” Biochemistry **39**(50): 15556-62.

Goldberg, A. L. and R. E. Wittes (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**(6): 662-4.

Gould, S. J. and R. Lewontin (1979). “The Spandrels of San Marco and the Panglossian Paradigm: a critique of the adaptationist programme.” Proceedings of the Royal Society of London **205**: 581-598.

Gouy, M. and C. Gautier (1982). “Codon usage in bacteria: correlation with gene expressivity.” Nucleic Acids Res **10**(22): 7055-74.

Grantham, R. (1974). “Amino Acid Difference Formula to Help Explain Protein Evolution.” Science **185**: 862-865.

Graur, D. and W. Li (2000). Fundamentals of Molecular Evolution. Sunderland, Mass., Sinauer.

Grimm, M., C. Brunen-Nieweler, V. Junker, K. Heckmann, et al. (1998). “The hypotrichous ciliate Euplotes octocarinatus has only one type of tRNACys with GCA anticodon encoded on a single macronuclear DNA molecule.” Nucleic Acids Res **26**(20): 4557-65.

Gu, X., D. Hewett-Emmett and W. H. Li (1998). “Directional mutational pressure affects the amino acid composition and hydrophobicity of proteins in bacteria.” Genetica **103**(1-6): 383-91.

Guerrier-Takada, C., K. Gardiner, T. Marsh, N. Pace, et al. (1983). “The RNA Moiety of Ribonuclease P Is the Catalytic Subunit of the Enzyme.” Cell **35**: 849-857.

Gupta, S. K., S. Majumdar, T. K. Bhattacharya and T. C. Ghosh (2000). “Studies on the relationships between the synonymous codon usage and protein secondary structural units.” Biochem Biophys Res Commun **269**(3): 692-6.

Haig, D. and L. D. Hurst (1991). “A Quantitative Measure of Error Minimization in the Genetic Code.” J. Mol. Evol. **33**: 412-417.

Haig, D. and L. D. Hurst (1999). “A quantitative measure of error minimization in the genetic code.” J Mol Evol **49**(5): 708.

Hammerschmidt, B., M. Schlegel, D. H. Lynn, D. D. Leipe, et al. (1996). “Insights into the evolution of nuclear dualism in the ciliates revealed by phylogenetic analysis of rRNA sequences.” J Eukaryot Microbiol **43**: 225-230.

Hanvey, J. C., N. J. Peffer, J. E. Bisi, S. A. Thomson, et al. (1992). “Antisense and antigene properties of peptide nucleic acids.” Science **258**(5087): 1481-5.

Hanyu, N., Y. Kuchino, S. Nishimura and H. Beier (1986). “Dramatic events in ciliate evolution: alteration of UAA and UAG termination codons to glutamine codons due to anticodon mutations in two Tetrahymena tRNAGln.” EMBO J **5**(1307-1311).

Hartman, H. (1995). “Speculations on the Origin of the Genetic Code.” J Mol Evol **40**: 541-544.

Harvey, P. H. and M. D. Pagel (1991). The Comparative Method in Evolutionary Biology. Oxford & NY, Oxford University Press.

Hasegawa, M., T. Yasunaga and T. Miyata (1979). “Secondary structure of MS2 phage RNA and bias in code word usage.” Nucleic Acids Res **7**(7): 2073-9.

Hayashi-Ishimaru, Y., M. Ehara, Y. Inagaki and T. Ohama (1997). “A deviant mitochondrial genetic code in prymnesiophytes (yellow-algae): UGA codon for tryptophan.” Curr Genet **32**: 296-299.

Hayashi-Ishimaru, Y., T. Ohama, Y. Kawatsu, K. Nakamura, et al. (1996). “UAG is a sense codon in several chlorophycean mitochondria.” Curr Genet **30**: 29-33.

Hendry, L. B., E. D. Bransome Jr, M. S. Hutson and L. K. Campbell (1981). “First approximation of a stereochemical rationale for the genetic code based on the topography and physichemical properties of "cavities" constructed from models of DNA.” Proc Natl Acad Sci USA **78**(12): 7440-7444.

Hendry, L. B. and F. H. Whitham (1979). “Stereochemical recognition in nucleic acid-amino acid interactions and its implications in biological coding: a model approach.” Perspect. Biol. Med. **22**: 333-345.

Himeno, H., S. Yoshida, A. Soma and K. Nishikawa (1997). “Only one nucleotide insertion to the long variable arm confers an efficient serine acceptor activity upon Saccharomyces cerevisiae tRNA(Leu) in vitro.” J Mol Biol **268**(4): 704-11.

Hirao, I. and A. D. Ellington (1995). “Re-creating the RNA world.” Current Biology **5**(9): 1017-1022.

Hirsh, D. (1971). “Tryptophan transfer RNA as the UGA suppressor.” J Mol Biol **58**(2): 439-58.

Holm, L. (1986). “Codon usage and gene expression.” Nucleic Acids Res **14**(7): 3075-87.

Hong, K. W., M. Ibba and D. Soll (1998). “Retracing the evolution of amino acid specificity in glutaminyl-tRNA synthetase.” FEBS Lett **434**(1-2): 149-54.

Hopfield, J. J. (1978). “Origin of the genetic code: A testable hypothesis based on tRNA structure, sequence, and kinetic proofreading.” Proc Natl Acad Sci USA **75**(9): 4334-4338.

Horie, N., Z. Yamaizumi, Y. Kuchino, K. Takai, et al. (1999). “Modified nucleosides in the first positions of the anticodons of tRNA(Leu)4 and tRNA(Leu)5 from Escherichia coli.” Biochemistry **38**(1): 207-17.

Hornos, J. E. and Y. M. Hornos (1993). “Algebraic model for the evolution of the genetic code.” Physical Review Letters **71**(26): 4401-4404.

Horowitz, S. and M. A. Gorovsky (1985). “An unusual genetic code in nuclear genes of Tetrahymena.” Proc Natl Acad Sci U S A **82**(8): 2452-5.

Horton, T. L. and L. F. Landweber (2000). “Evolution of four types of RNA editing in myxomycetes.” RNA **6**(10): 1339-46.

Huynen, M. A., D. A. Konings and P. Hogeweg (1992). “Equal G and C contents in histone genes indicate selection pressures on mRNA secondary structure.” J Mol Evol **34**(4): 280-91.

Ibba, M., J. L. Bono, P. A. Rosa and D. Soll (1997). “Archaeal-type lysyl-tRNA synthetase in the Lyme disease spirochete Borrelia burgdorferi.” Proc Natl Acad Sci U S A **94**(26): 14383-14388.

Ibba, M. and D. Soll (1999). “Quality Control Mechanisms During Translation.” Science **286**(5446): 1893-1897.

Ikemura, T. (1981). “Correlation between the abundance of Escherichia coli transfer RNAs and the occurrence of the respective codons in its protein genes: a proposal for a synonymous codon choice that is optimal for the E. coli translational system.” J Mol Biol **151**(3): 389-409.

Ikemura, T. (1982). “Correlation between the abundance of yeast transfer RNAs and the occurrence of the respective codons in protein genes. Differences in synonymous codon choice patterns of yeast and Escherichia coli with reference to the abundance of isoaccepting transfer RNAs.” J Mol Biol **158**(4): 573-97.

Ikemura, T. (1985). “Codon usage and tRNA content in unicellular and multicellular organisms.” Mol Biol Evol **2**(1): 13-34.

Ikemura, T. (1992). Correlation between codon usage and tRNA content in microorganisms. Transfer RNA in protein synthesis. D. L. Hatfield and B. L. Lee. Boca Raton, FL, CRC Press**:** 87-111.

Ikemura, T. (1992). “Correlation between the abundance of yeast transfer RNAs and the occurrence of respective codons in protein genes.” J Mol Biol **158**: 573-597.

Ikemura, T. and H. Ozeki (1983). “Codon usage and transfer RNA contents: organism-specific codon-choice patterns in reference to the isoacceptor contents.” Cold Spring Harb Symp Quant Biol **47**(Pt 2): 1087-97.

Illangasekare, M., G. Sanchez, T. Nickles and M. Yarus (1995). “Aminoacyl-RNA Synthesis Catalyzed by an RNA.” Science **267**: 643-647.

Illangasekare, M. and M. Yarus (1999). “Specific, rapid synthesis of Phe-RNA by RNA.” Proc Natl Acad Sci U S A **96**(10): 5470-5.

Illangasekare, M. and M. Yarus (1999). “A tiny RNA that catalyzes both aminoacyl-RNA and peptidyl-RNA synthesis [In Process Citation].” RNA **5**(11): 1482-9.

Inagaki, Y., Y. Bessho, H. Hori and S. Osawa (1996). “Cloning of the Mycoplasma capricolum gene encoding peptide-chain release factor.” Gene **169**(1): 101-3.

Inagaki, Y., Y. Bessho and S. Osawa (1993). “Lack of peptide-release activity responding to codon UGA in Mycoplasma capricolum.” Nucleic Acids Res **21**(6): 1335-8.

Inagaki, Y., M. Ehara, K. I. Watanabe, Y. Hayashi-Ishimaru, et al. (1998). “Directionally evolving genetic code: the UGA codon from stop to tryptophan in mitochondria.” J Mol Evol **47**(4): 378-84.

Ito, K., M. Uno and Y. Nakamura (2000). “A tripeptide 'anticodon' deciphers stop codons in messenger RNA.” Nature **403**(6770): 680-4.

Janke, A. and S. Paabo (1993). “Editing of a tRNA anticodon in marsupial mitochondria changes its codon recognition.” Nucleic Acids Res **21**(7): 1523-5.

Jay, D. G. and W. Gilbert (1987). “Basic protein enhances the incorporation of DNA into lipid vesicles: model for the formation of primordial cells.” Proc Natl Acad Sci U S A **84**(7): 1978-1980.

Jayaram, B. (1997). “Beyond the Wobble: The Rule of Conjugates.” J Mol Evol **45**: 704-705.

Jiang, F., R. A. Kumar, J. R. A and D. J. Patel (1996). “Structural basis of RNA folding and recognition in an AMP·RNA aptamer complex.” Nature **382**: 183-186.

Jiang, L., A. Majumdar, W. Hu, T. J. Jaishree, et al. (1999). “Saccharide-RNA recognition in a complex formed between neomycin B and an RNA aptamer.” Structure Fold Des **7**(7): 817-27.

Jiang, L. and D. J. Patel (1998). “Solution structure of the tobramycin-RNA aptamer complex.” Nat Struct Biol **5**(9): 769-74.

Jiang, L., A. K. Suri, R. Fiala and D. J. Patel (1997). “Saccharide-RNA recognition in an aminoglycoside antibiotic-RNA aptamer complex.” Chem Biol **4**: 35-50.

Jiménez-Montaño, M. A. (1994). “On the syntactic structure and redundancy distribution of the genetic code.” Bio Systems **32**: 11-23.

Jiménez-Sánchez, A. (1995). “On the Origin and Evolution of the Genetic Code.” J Mol Evol **41**: 712-716.

Joshi, N. V., V. V. Korde and V. Sitaramam (1993). “Logic of the genetic code: Conservation of long-range interactions among amino acids as a prime factor.” J Genet **72**: 47-58.

Joyce, G. F. (1989). “Amplifiction, mutation and selection of catalytic RNA.” Gene **82**: 83-87.

Joyce, G. F. (1989). “RNA evolution and the origins of life.” Nature **338**: 217-224.

Joyce, G. F. and L. E. Orgel (1993). Prospects for Understanding the Origin of the RNA World. The RNA World. R. F. Gesteland and J. F. Atkins. New York, Cold Spring Harbor Laboratory Press**:** 1-25.

Joyce, G. F., A. W. Schwartz, S. L. Miller and L. E. Orgel (1987). “The case for an ancestral genetic system involving simple analogues of the nucleotides.” Proc Natl Acad Sci USA **84**: 4398-4402.

Judson, O. P. and D. Haydon (1999). “The genetic code: what is it good for?” J Mol Evol **49**: 539-550.

Jukes, T. H. (1973). “Arginine as an evolutionary intruder into protein synthesis.” Biochem Biophys Res Comm **53**(3): 709-714.

Jukes, T. H. (1990). “Genetic code 1990. Outlook.” Experientia **46**(11-12): 1149-57.

Jukes, T. H. (1996). “Neutral changes and modifications of the genetic code.” Theor Popul Biol **49**(2): 143-5.

Jukes, T. H. and J. L. King (1971). “Deleterious mutations and neutral substitutions.” Nature **231**(5298): 114-5.

Jukes, T. H. and S. Osawa (1997). “Further Comments on Codon Reassignment.” J Mol Evol **45**: 1-8.

Jukes, T. H., S. Osawa, A. Muto and N. Lehman (1987). “Evolution of Anticodons: Variations in the Genetic Code.” Cold Spring Harbor Symposia on Quantitative Biology **52**: 769-776.

Jungck, J. R. (1978). “The Genetic Code as a Periodic Table.” J Mol Evol **11**: 211-224.

Kanaya, S., Y. Yamada, Y. Kudo and T. Ikemura (1999). “Studies of codon usage and tRNA genes of 18 unicellular organisms and quantification of Bacillus subtilis tRNAs: gene expression level and species-specific diversity of codon usage based on multivariate analysis.” Gene **238**(1): 143-55.

Kano, A., T. Ohama, R. Abe and S. Osawa (1993). “Unassigned or nonsense codons in Micrococcus luteus.” J Mol Biol **230**(1): 51-6.

Karamyshev, A. L., K. Ito and Y. Nakamura (1999). “Polypeptide release factor eRF1 from Tetrahymena thermophila: cDNA cloning, purification and complex formation with yeast eRF3.” FEBS Lett **457**(3): 483-8.

Karkas, J. D., R. Rudner and E. Chargaff (1968). “Separation of B. subtilis DNA into complementary strands. II. Template functions and composition as determined by transcription with RNA polymerase.” Proc Natl Acad Sci U S A **60**(3): 915-920.

Karlin, S. and J. Mrazek (1996). “What drives codon choices in human genes?” J Mol Biol **262**(4): 459-72.

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

Kawashima, S. and M. Kanehisa (2000). “AAindex: amino acid index database.” Nucleic Acids Res **28**(1): 374.

Keefe, A. D. and S. L. Miller (1995). “Are Polyphosphates or Phosphate Esters Prebiotic Reagents?” J Mol Evol **41**: 693-702.

Keeling, P. J. and W. F. Doolittle (1997). “Widespread and Ancient Distribution of a Noncanonical Genetic Code in Diplomonads.” Mol Biol Evol **14**(9): 895-901.

Kimura, M. (1962). “On the probability of fixation of mutant genes in populations.” Genetics **47**: 713-719.

Kimura, M. (1968). “Evolutionary rate at the molecular level.” Nature **217**(129): 624-6.

Kimura, M. (1968). “Genetic variability maintained in a finite population due to mutational production of neutral and nearly neutral isoalleles.” Genet Res **11**(3): 247-69.

King, J. L. and T. H. Jukes (1969). “Non-Darwinian evolution.” Science **164**(881): 788-98.

Klump, H. H. (1993). “The physical basis of the genetic code: the choice between speed and precision.” Arch Biochem Biophys **301**(2): 207-9.

Knight, R. D., S. J. Freeland and L. F. Landweber (1999). “Selection, history and chemistry: the three faces of the genetic code.” Trends Biochem Sci **24**(6): 241-7.

Knight, R. D., S. J. Freeland and L. F. Landweber (2001). “Rewiring the keyboard: evolvability of the genetic code.” Nat Rev Genet **2**: 49-58.

Knight, R. D., S. J. Freeland and L. F. Landweber (2001). “A simple model based on mutation and selection explains trends in codon and amino-acid usage and GC composition within and across genomes.” GenomeBiology **2**(4): http://www.genomebiology.com/2001/2/4/research/0010/.

Knight, R. D. and L. F. Landweber (1998). “Rhyme or reason: RNA-arginine interactions and the genetic code.” Chem Biol **5**(9): R215-20.

Knight, R. D. and L. F. Landweber (2000). “The Early Evolution of the Genetic Code.” Cell **101**: 569-572.

Knight, R. D. and L. F. Landweber (2000). “Guilt by association: the arginine case revisited.” RNA **6**(4): 499-510.

Knight, R. D., L. F. Landweber and M. Yarus (In Press). “How Mitochondria Redefine the Code.” J Mol Evol(Early 2001).

Koshi, J. M. and R. A. Goldstein (1997). “Mutation matrices and physical-chemical properties: correlations and implications.” Proteins **27**(3): 336-44.

Kruger, K., P. J. Grabowski, A. J. Zaug, J. Sands, et al. (1982). “Self-Splicing RNA: Autoexcision and Autocyclization of the Ribosomal RNA Intervening Sequence of Tetrahymena.” Cell **31**: 147-157.

Kuck, U., K. Jekosch and P. Holzamer (2000). “DNA sequence analysis of the complete mitochondrial genome of the green alga Scenedesmus obliquus: evidence for UAG being a leucine and UCA being a non-sense codon.” Gene **253**(1): 13-8.

Kvenvolden, K., J. G. Lawless, K. Pering, E. Peterson, et al. (1970). “Evidence for extraterrestrial amino-acids and hydrocarbons in the Murchison meteorite.” Nature **228**: 923-926.

Kvenvolden, K. A., J. G. Lawless and C. Ponnamperuma (1971). “Nonprotein amino acids in the Murchison meteorite.” Proc Natl Acad Sci USA **68**: 486-490.

Kyte, J. and R. F. Doolittle (1982). “A simple method for displaying the hydropathic character of a protein.” J Mol Biol **157**(1): 105-32.

Lacey, J. C., Jr and D. W. Mullins, Jr (1983). “Experimental Studies Related to the Origin of the Genetic Code and the Process of Protein Synthesis—A Review.” Orig Life **13**: 3-42.

Lacey, J. C., Jr. (1992). “Experimental studies on the origin of the genetic code and the process of protein synthesis: a review update.” Orig Life Evol Biosph **22**: 243-275.

Lacey, J. C., Jr. and K. M. Pruitt (1969). “Origin of the Genetic Code.” Nature **223**: 799-804.

Lacey, J. C., Jr., N. S. M. D. Wickramasinghe, G. W. Cook and G. Anderson (1993). “Couplings of Character and of Chirality in the Origin of the Genetic System.” J Mol Evol **37**: 233-239.

Lafay, B., J. C. Atherton and P. M. Sharp (2000). “Absence of translationally selected synonymous codon usage bias in Helicobacter pylori.” Microbiology **146**(Pt 4): 851-60.

Laforest, M. J., I. Roewer and B. F. Lang (1997). “Mitochondrial tRNAs in the lower fungus Spizellomyces punctatus: tRNA editing and UAG 'stop' codons recognized as leucine.” Nucleic Acids Res **25**(3): 626-32.

Lagerkvist, U. (1978). “"Two out of three": An alternative method for codon reading.” Proc Natl Acad Sci USA **75**(4): 1759-1762.

Lagerkvist, U. (1980). “Codon Misreading: A Restriction Operative in the Evolution of the Genetic Code.” American Scientist **68**: 192-198.

Lagerkvist, U. (1981). “Unorthodox codon reading and the evolution of the genetic code.” Cell **23**: 305-306.

Lake, J. A. (1985). “Evolving ribosome structure: Domains in archaebacteria, eubacteria, eocytes and eucaryotes.” Ann Rev Biochem **54**: 508-530.

Lamond, A. I. and T. J. Gibson (1990). “Catalytic RNA and the origin of genetic systems.” Trends in Genetics **6**(5): 145-149.

Lamour, V., S. Quevillon, S. Diriong, V. C. N'Guyen, et al. (1994). “Evolution of the Glx-tRNA synthetase family: the glutaminyl enzyme as a case of horizontal gene transfer.” Proc Natl Acad Sci U S A **91**(18): 8670-4.

Landweber, L. F. and L. A. Katz (1998). “Evolution: Lost Worlds.” Trends Ecol Evol **13**: 93-94.

Landweber, L. F., P. J. Simon and T. A. Wagner (1998). “Ribozyme Engineering and Early Evolution.” BioScience **48**(2): 94-103.

Larralde, R., M. P. Robertson and S. L. Miller (1995). “Rates of decomposition of ribose and other sugars: Implications for chemical evolution.” Proc. Natl. Acad. Sci. USA **92**: 8158-8160.

Lazcano, A. and S. L. Miller (1996). “The Origin and Early Evolution of Life: Prebiotic Chemistry, the Pre-RNA World, and Time.” Cell **85**: 793-798.

Lee, D. H., K. Severin, Y. Yokobayashi and M. R. Ghadiri (1997). “Emergence of symbiosis in peptide self-replication through a hypercyclic network.” Nature **390**: 591-594.

Lee, N., Y. Bessho, K. Wei, J. W. Szostak, et al. (2000). “Ribozyme-catalyzed tRNA aminoacylation [see comments].” Nat Struct Biol **7**(1): 28-33.

Lehman, N. and T. H. Jukes (1988). “Genetic code development by stop codon takeover.” J Theor Biol **135**(2): 203-14.

Lehmann, U. (1985). “Chromatographic separation as selection process for prebiotic evolution and the origin of the genetic code.” Bio Systems **17**: 193-208.

Leinfelder, W., E. Zehelein, M. A. Mandrand-Berthelot and A. Bock (1988). “Gene for a novel tRNA species that accepts L-serine and cotranslationally inserts selenocysteine.” Nature **331**(6158): 723-5.

Lenhard, B., O. Orellana, M. Ibba and I. Weygand-Durasevic (1999). “tRNA recognition and evolution of determinants in seryl-tRNA synthesis.” Nucleic Acids Res **27**(3): 721-9.

Levitt, M. (1978). “Conformational preferences of amino acids in globular proteins.” Biochemistry **17**(20): 4277-85.

Levy, M. and S. L. Miller (1998). “The stability of the RNA bases: implications for the origin of life.” Proc Natl Acad Sci U S A **95**(14): 7933-8.

Levy, M. and S. L. Miller (1999). “The prebiotic synthesis of modified purines and their potential role in the RNA world.” J Mol Evol **48**(6): 631-7.

Lewin, R. (1986). “RNA catalysis gives fresh perspective on the origin of life.” Science **231**: 545-546.

Li, J., B. Esberg, J. F. Curran and G. R. Bjork (1997). “Three modified nucleosides present in the anticodon stem and loop influence the in vivo aa-tRNA selection in a tRNA-dependent manner.” J Mol Biol **271**(2): 209-21.

Li, T., Y. Li, N. Guo, E. Wang, et al. (1999). “Discrimination of tRNALeu isoacceptors by the insertion mutant of Escherichia coli leucyl-tRNA synthetase.” Biochemistry **38**(28): 9084-8.

Liang, A., C. Brünen-Niewler, T. Muramatsu, Y. Kuchino, et al. (In Press). “The ciliate *Euplotes octocarinatus* expresses two polypeptide release factors of the type eRF1.”.

Liang, A. and K. Heckmann (1993). “Blepharisma uses UAA as a termination codon.” Naturwissenschaften **80**(5): 225-6.

Lin, C. H., W. Wang, R. A. Jones and D. J. Patel (1998). “Formation of an amino-acid-binding pocket through adaptive zippering-up of a large DNA hairpin loop.” Chem Biol **5**(10): 555-72.

Lobry, J. R. (1995). “Properties of a general model of DNA evolution under no-strand-bias conditions [published erratum appears in J Mol Evol 1995 Nov;41(5):680].” J Mol Evol **40**(3): 326-30.

Lobry, J. R. (1996). “Asymmetric substitution patterns in the two DNA strands of bacteria.” Mol Biol Evol **13**(5): 660-5.

Lobry, J. R. (1997). “Influence of genomic G+C content on average amino-acid composition of proteins from 59 bacterial species.” Gene **205**(1-2): 309-16.

Lobry, J. R. and C. Gautier (1994). “Hydrophobicity, expressivity and aromaticity are the major trends of amino-acid usage in 999 Escherichia coli chromosome-encoded genes.” Nucleic Acids Res **22**(15): 3174-80.

Lohse, P. A. and J. W. Szostak (1996). “Ribozyme-catalysed amino-acid transfer reactions.” Nature **381**: 442-444.

Lovett, P. S., N. P. Ambulos, Jr., W. Mulbry, N. Noguchi, et al. (1991). “UGA can be decoded as tryptophan at low efficiency in Bacillus subtilis.” J Bacteriol **173**(5): 1810-2.

Lutz, M. J., J. Horlacher and S. A. Benner (1998). “Recognition of a non-standard base pair by thermostable DNA polymerases.” Bioorg Med Chem Lett **8**(10): 1149-52.

Lynn, D. H. and E. B. Small (1997). “A revised classification of the Phylum Ciliophora Doflein, 1901.” Rev Soc Mex Hist Nat **47**: 65-78.

Mackay, A. L. (1967). “Optimization of the genetic code.” Nature **216**(111): 159-60.

Maddison, D. R. and W. P. Maddison (1998). The Tree of Life: A multi-authored, distributed Internet project containing information about phylogeny and biodiversity. **2000**.

Madore, E., C. Florentz, R. Giege, S. Sekine, et al. (1999). “Effect of modified nucleotides on Escherichia coli tRNAGlu structure and on its aminoacylation by glutamyl-tRNA synthetase. Predominant and distinct roles of the mnm5 and s2 modifications of U34.” Eur J Biochem **266**(3): 1128-35.

Maeshiro, T. and M. Kimura (1998). “The role of robustness and changeability on the origin and evolution of genetic codes.” Proc Natl Acad Sci U S A **95**(9): 5088-93.

Maizels, N. and A. M. Weiner (1987). “Peptide-specific Ribosomes, Genomic Tags, and the Origin of the Genetic Code.” Cold Spring Harbor Symposia on Quantitative Biology **LII**: 743-749.

Maizels, N. and A. M. Weiner (1993). The Genomic Tag Hypothesis: Modern Viruses as Molecular Fossils of Ancient Strategies for Genomic Replication. The RNA World. R.

F. Gesteland and J. F. Atkins. New York, Cold Spring Harbor Laboratory Press**:** 577-602.

Maizels, N. and A. M. Weiner (1994). “Phylogeny from function: Evidence from the molecular fossil record that tRNA originated in replication, not translation.” Proc Natl Acad Sci USA **91**: 6729-6734.

Majerfeld, I. and M. Yarus (1994). “An RNA pocket for an aliphatic hydrophobe.” Nat Struct Biol **1**(5): 287-292.

Majerfeld, I. and M. Yarus (1998). “Isoleucine:RNA sites with essential coding sequences.” RNA **4**: 471-478.

Mannironi, C., C. Scerch, P. Fruscoloni and G. P. Tocchini-Valentini (2000). “Molecular recognition of amino acids by RNA aptamers: the evolution into an L-tyrosine binder of a dopamine-binding RNA motif.” RNA **6**(4): 520-7.

Matsugi, J., K. Murao and H. Ishikura (1998). “Effect of B. subtilis TRNA(Trp) on readthrough rate at an opal UGA codon.” J Biochem (Tokyo) **123**(5): 853-8.

Matsuyama, S., T. Ueda, P. F. Crain, J. A. McCloskey, et al. (1998). “A novel wobble rule found in starfish mitochondria. Presence of 7- methylguanosine at the anticodon wobble position expands decoding capability of tRNA.” J Biol Chem **273**(6): 3363-8.

Mazauric, M. H., H. Roy and D. Kern (1999). “tRNA glycylation system from Thermus thermophilus. tRNAGly identity and functional interrelation with the glycylation systems from other phylae.” Biochemistry **38**(40): 13094-105.

Melcher, G. (1974). “Stereospecificity of the Genetic Code.” J Mol Evol **3**: 121-140.

Mellersh, A. (1993). “A model for the prebiotic synthesis of peptides and the genetic code.” Orig Life Evol Biosph **23**: 261-274.

Meyer, F., H. J. Schmidt, E. Plumper, A. Hasilik, et al. (1991). “UGA is translated as cysteine in pheromone 3 of Euplotes octocarinatus.” Proc Natl Acad Sci U S A **88**(9): 3758-61.

Miller, S. L. (1953). “Production of amino acids under possible primitive earth conditions.” Science **117**: 528-529.

Miller, S. L. (1987). “Which Organic Compounds Could Have Occurred on the Prebiotic Earth?” Cold Spring Harbor Symposia on Quantitative Biology **LII**: 17-27.

Mironova, L. N., O. A. Zelenaia and M. D. Ter-Avanesian (1986). “Nuclear-mitochondrial interactions in yeasts: mitochondrial mutations compensating the respiration deficiency of sup1 and sup2 mutants.” Genetika **22**(2): 200-8.

Miseta, A. (1989). “The role of protein associated amino acid precursor molecules in the organization of genetic codons.” Physiol Chem Phys Med NMR **21**: 237-242.

Morowitz, H. J., J. D. Kostelnik, J. Yang and G. D. Cody (2000). “The origin of intermediary metabolism.” Proc Natl Acad Sci U S A **97**(14): 7704-8.

Motorin, Y. and H. Grosjean (1998). Chemical Structures and Classification of Posttranscriptionally Modified Nucleosides in RNA. Modification and Editing of RNA. H. Grosean and R. Benne. Washington DC, ASM Press.

Mouchiroud, D. and C. Gautier (1990). “Codon usage changes and sequence dissimilarity between human and rat.” J Mol Evol **31**(2): 81-91.

Muramatsu, T., K. Nishikawa, F. Nemoto, Y. Kuchino, et al. (1988). “Codon and amino-acid specificities of a transfer RNA are both converted by a single post-transcriptional modification.” Nature **336**(6195): 179-81.

Murgola, E. J. (1985). “tRNA, suppression, and the code.” Annu Rev Genet **19**: 57-80.

Murgola, E. J. (1995). Translational Suppression: When Two Wrongs DO Make a Right. tRNA: Structure, Biosynthesis and Function. D. Söll and U. RajBhandary. Washington, DC, American Society for Microbiology**:** 491-509.

Muto, A. and S. Osawa (1987). "The guanine and cytosine content of genomic DNA and bacterial evolution." Proc Natl Acad Sci U S A **84**(1): 166-9.

Nagel, G. M. and R. F. Doolittle (1991). "Evolution and relatedness in two aminoacyl-tRNA synthetase families." Proc Natl Acad Sci U S A **88**(18): 8121-5.

Nagel, G. M. and R. F. Doolittle (1995). "Phylogenetic Analysis of the Aminoacyl-tRNA Synthetases." J Mol Evol **40**: 487-498.

Nagyvary, J. and J. H. Fendler (1974). "Origin of the genetic code: A physical-chemical model of primitive codon assignments." Orig Life **5**: 357-362.

Nakamura, Y., T. Gojobori and T. Ikemura (2000). "Codon usage tabulated from international DNA sequence databases: status for the year 2000." Nucleic Acids Res **28**(1): 292.

Nakamura, Y. and S. Tabata (1997). "Codon-anticodon assignment and detection of codon usage trends in seven microbial genomes." Microb Comp Genomics **2**(4): 299-312.

Nakashima, H., K. Nishikawa and T. Ooi (1990). "Distinct character in hydrophobicity of amino acid compositions of mitochondrial proteins." Proteins **8**(2): 173-8.

Neander, K. (1991). "The teleological notion of "function"." Australasian Journal of Philosophy **69.4**: 454-468.

Nelsesteuen, G. L. (1978). "Amino Acid-Directed Nucleic Acid Synthesis." J Mol Evol **11**: 109-120.

Nelson, K. E., M. Levy and S. L. Miller (2000). "Peptide nucleic acids rather than RNA may have been the first genetic molecule." Proc Natl Acad Sci U S A **97**(8): 3868-71.

Nicholas, H. B., Jr and W. H. McClain (1995). "Searching tRNA Sequences for Relatedness to Aminoacyl-tRNA Synthetase Families." J Mol Evol **40**: 482-486.

Nielsen, P. E. (1993). "Peptide nucleic acid (PNA): a model structure for the primordial genetic material?" Orig Life Evol Biosph **23**(5-6): 323-7.

Nielsen, P. E., M. Egholm, R. H. Berg and O. Buchardt (1991). "Sequence-Selective Recognition of DNA by Strand Displacement with a Thymine-Substituted Polyamide." Science **254**: 1497-1500.

Nissen, P., J. Hansen, N. Ban, P. B. Moore, et al. (2000). "The structural basis of ribosome activity in peptide bond synthesis." Science **289**(5481): 920-30.

Noller, H. F. (1993). On the Origin of the Ribosome: Coevolution of Subdomains of tRNA and rRNA. The RNA World. R. F. Gesteland and J. F. Atkins. New York, Cold Spring Harbor Laboratory Press**:** 137-156.

Noller, H. F., V. Hoffarth and L. Zimniak (1992). "Unusual Resistance of Peptidyl Transferase to Protein Extraction Procedures." Science **256**: 1416-1419.

Oba, T., Y. Andachi, A. Muto and S. Osawa (1991). "CGG: an unassigned or nonsense codon in Mycoplasma capricolum." Proc Natl Acad Sci U S A **88**(3): 921-5.

Ogawa, A. K., Y. Q. Wu, D. L. McMinn, J. Q. Liu, et al. (2000). "Efforts toward the expansion of the genetic alphabet: Information storage and replication with unnatural hydrophobic base pairs." J Am Chem Soc **122**: 3274-3287.

Ohta, T. and J. H. Gillespie (1996). "Development of Neutral and Nearly Neutral Theories." Theor Popul Biol **49**(2): 128-42.

Olins, D. E., A. L. Olins and P. H. von Hippel (1967). “Model Nucleoprotein Complexes: Studies on the Interaction of Cationic Homopolypeptides with DNA.” J Mol Biol **24**: 157-176.

Oobatake, M. and T. Ooi (1977). “An analysis of non-bonded energy of proteins.” J Theor Biol **67**(3): 567-84.

Oresic, M. and D. Shalloway (1998). “Specific correlations between relative synonymous codon usage and protein secondary structure.” J Mol Biol **281**(1): 31-48.

Orgel, L. E. (1968). “Evolution of the Genetic Apparatus.” J Mol Biol **38**: 381-393.

Orgel, L. E. (1986). “RNA Catalysis and the Origins of Life.” J theor Biol **123**: 127-149.

Orgel, L. E. (1990). “Adding to the genetic alphabet.” Nature **343**.

Orgel, L. E. (2000). “Self-organizing biochemical cycles.” Proc Natl Acad Sci U S A **97**(23): 12503-7.

Oró, J. and A. P. Kimball (1962). “Synthesis of Purines under Possible Primitive Earth Conditions II. Purine Intermediates from Hydrogen Cyanide.” Arch Biochem Biophys **96**: 293-323.

Oró, J. and P. Kimball (1961). “Synthesis of Purines under Possible Primitive Earth Conditions. I. Adenine from Hydrogen Cyanide.” Arch Biochem Biophys **94**: 217-227.

Osawa, S. (1995). Evolution of the Genetic Code. Oxford, Oxford University Press.

Osawa, S. and T. H. Jukes (1988). “Evolution of the genetic code as affected by anticodon content.” Trends Genet. **4**(7): 191-198.

Osawa, S. and T. H. Jukes (1989). “Codon Reassignment (Codon Capture) in Evolution.” J Mol Evol **28**: 271-278.

Osawa, S., T. H. Jukes, A. Muto, F. Yamao, et al. (1987). “Role of directional mutation pressure in the evolution of the eubacterial genetic code.” Cold Spring Harb Symp Quant Biol **52**: 777-89.

Osawa, S., T. H. Jukes, K. Watanabe and A. Muto (1992). “Recent evidence for evolution of the genetic code.” Microbiol Rev **56**(1): 229-64.

Osawa, S., T. Ohama, T. H. Jukes and K. Watanabe (1989). “Evolution of the mitochondrial genetic code. I. Origin of AGR serine and stop codons in metazoan mitochondria.” J Mol Evol **29**(3): 202-7.

Osawa, S., T. Ohama, F. Yamao, A. Muto, et al. (1988). “Directional mutation pressure and transfer RNA in choice of the third nucleotide of synonymous two-codon sets.” Proc Natl Acad Sci U S A **85**(4): 1124-8.

Ota, T. and M. Kimura (1971). “Amino acid composition of proteins as a product of molecular evolution.” Science **174**(5): 150-3.

Pallanck, K., M. Pak and L. H. Schulman (1995). tRNA Discrimination in Aminoacylation. tRNA: Structure, Biosynthesis, and Function. D. Söll and U. RajBhandary. Washington, DC, American Society for Microbiology**:** 371-394.

Pelc, S. R. (1965). “Correlation between coding triplets and amino acids.” Nature **207**: 597-599.

Pelc, S. R. and M. G. E. Welton (1966). “Stereochemical relationship between coding triplets and amino-acids.” Nature **209**: 868-872.

Perreau, V. M., G. Keith, W. M. Holmes, A. Przykorska, et al. (1999). “The Candida albicans CUG-decoding ser-tRNA has an atypical anticodon stem-loop structure.” J Mol Biol **293**(5): 1039-53.

Perret, V., A. Garcia, H. Grosjean, J. P. Ebel, et al. (1990). “Relaxation of a transfer RNA specificity by removal of modified nucleotides.” Nature **344**(6268): 787-9.

Philippe, H. and P. Forterre (1999). “The rooting of the universal tree of life is not reliable.” J Mol Evol **49**(4): 509-23.

Piccirilli, J. A., T. Krauch, S. E. Moroney and S. A. Benner (1990). “Enzymatic incorporation of a new base pair into DNA and RNA extends the genetic alphabet.” Nature **343**: 33-37.

Podder, S. K. and H. S. Basu (1984). “Specificity of protein-nucleic acid interaction and the biochemcial evolution.” Orig Life **14**: 477-484.

Porschke, D. (1985). “Differential Effect of Amino Acid Residues on the Stability of Double Helices Formed from Polyribonucleotides and Its Possible Relation to the Evolution of the Genetic Code.” J Mol Evol **21**: 192-198.

Preer, J. R., L. B. Preer, B. M. Rudman and A. J. Barnett (1985). “Deviation from the universal code shown by the gene for surface protein 51A in Paramecium.” Nature **314**(6007): 188-90.

Press, W. H., S. A. Teukolsky, W. T. Vettering and B. P. Flannery (1992). Numerical Recipes in C. New York, Cambridge University Press.

Radzicka, A., G. B. Young and R. Wolfenden (1993). “Lack of water transport by amino acid side chains or peptides entering a nonpolar environment.” Biochemistry **32**(27): 6807-9.

Ralph, R. K. (1968). “A suggestion on the origin of the genetic code.” Biochem Biophys Res Comm **33**(2): 213-218.

Raszka, M. and M. Mandel (1972). “Is There A Physical Chemical Basis for the Present Genetic Code?” J Mol Evol **2**: 38-43.

Rendell, M. S., J. P. Harlos and R. Rein (1971). “Specificity in the Genetic Code: The Role of Nucleotide Base-Amino Acid Interaction.” Biopolymers **10**: 2083-2094.

Reuben, J. and F. E. Polk (1980). “Nucleotide-Amino Acid Interactions and Their Relation to the Genetic Code.” J Mol Evol **15**: 103-112.

Ribas de Pouplana, L., M. Frugier, C. L. Quinn and P. Schimmel (1996). “Evidence that two present-day components needed for the genetic code appeared after nucleated cells separated from eubacteria.” Proc Natl Acad Sci U S A **93**(1): 166-70.

Ribas de Pouplana, L., R. J. Turner, B. A. Steer and P. Schimmel (1998). “Genetic code origins: tRNAs older than their synthetases?” Proc Natl Acad Sci U S A **95**(19): 11295-11300.

Ring, D., Y. Wolman, N. Friedmann and S. L. Miller (1972). “Prebiotic Synthesis of Hydrophobic and Protein Amino Acids.” Proc Natl Acad Sci USA **69**(3): 765-768.

Robertson, D. L. and G. F. Joyce (1990). “Selection in vitro of an RNA enzyme that specifically cleaves single- stranded DNA.” Nature **344**: 467-468.

Robertson, S. A., K. Harada, A. D. Frankel and D. E. Wemmer (In Press). “Structure Determination and Binding Kinetics of a DNA Aptamer-Arginamide Complex.” Biochemistry.

Robson, B. and E. Suzuki (1976). “Conformational properties of amino acid residues in globular proteins.” J Mol Biol **107**(3): 327-56.

Rodin, S., A. Rodin and S. Ohno (1996). “The presence of codon-anticodon pairs in the acceptor stem of tRNAs.” Proc Natl Acad Sci USA **93**: 4537-4542.

Rogers, K. C. and D. Soll (1995). “Divergence of glutamate and glutamine aminoacylation pathways: providing the evolutionary rationale for mischarging.” J Mol Evol **40**(5): 476-81.

Rogers, M. J., J. Simmons, R. T. Walker, W. G. Weisburg, et al. (1985). “Construction of the mycoplasma evolutionary tree from 5S rRNA sequence data.” Proc Natl Acad Sci U S A **82**(4): 1160-4.

Ronneberg, T. A., L. F. Landweber and S. J. Freeland (2000). “Testing a biosynthetic theory of the genetic code: fact or artifact?” Proc Natl Acad Sci U S A **97**(25): 13690-5.

Root-Bernstein, R. S. (1982). “Amino Acid Pairing.” J theor Biol **94**: 885-894.

Root-Bernstein, R. S. (1982). “On the Origin of the Genetic Code.” J theor Biol **94**: 895-904.

Rose, G. D. and R. Wolfenden (1993). “Hydrogen bonding, hydrophobicity, packing, and protein folding.” Annu Rev Biophys Biomol Struct **22**: 381-415.

Rudner, R., J. D. Karkas and E. Chargaff (1968). “Separation of B. subtilis DNA into complementary strands. III. Direct analysis.” Proc Natl Acad Sci U S A **60**(3): 921-922.

Saks, M. E. and J. R. Sampson (1995). “Evolution of tRNA recognition systems and tRNA gene sequences.” J Mol Evol **40**(5): 509-518.

Saks, M. E., J. R. Sampson and J. Abelson (1998). “Evolution of a Transfer RNA Gene Through a Point Mutation in the Anticodon.” Science **279**: 1665-1670.

Santos, M. A., C. Cheesman, V. Costa, P. Moradas-Ferreira, et al. (1999). “Selective advantages created by codon ambiguity allowed for the evolution of an alternative genetic code in Candida spp.” Mol Microbiol **31**(3): 937-47.

Santos, M. A. S., V. M. Perreau and M. F. Tuite (1996). “Transfer RNA structural change is a key element in the reassignment of the CUG codon in *Candida albicans*.” EMBO **15**(18): 5060-5068.

Santos, M. A. S. and M. F. Tuite (1995). “The CUG codon is decoded *in vivo* as serine and not leucine in *Candida albicans*.” Nucleic Acids Research **23**(9): 1481-1486.

Saxinger, C. and C. Ponnamperuma (1971). “Experimental Investigation on the Origin of the Genetic Code.” J Mol Evol **1**: 63-73.

Saxinger, C. and C. Ponnamperuma (1974). “Interactions between amino acids and nucleotides in the prebiotic milieu.” Orig Life **5**: 189-200.

Schimmel, P. (1995). “An Operational RNA Code for Amino Acids and Variations in Critical Nucleotide Sequences in Evolution.” J Mol Evol **40**: 531-536.

Schimmel, P., R. Giege, D. Moras and S. Yokoyama (1993). “An operational genetic code for amino acids and possible relationship to genetic code.” Proc Natl Acad Sci USA **90**: 8763-8768.

Schimmel, P. R. and D. Söll (1979). “Aminoacyl-tRNA synthetases: general features and recognition of transfer RNAs.” Annu Rev Biochem **48**: 601-648.

Schneider, S. U. and E. J. de Groot (1991). “Sequences of two rbcS cDNA clones of Batophora oerstedii: structural and evolutionary considerations.” Curr Genet **20**(1-2): 173-5.

Schön, A., C. G. Kannangara, S. Gough and D. Söll (1988). “Protein biosynthesis in organelles requires misaminoacylation of tRNA.” Nature **331**: 187-190.

Schrödinger, E. (1945). What Is Life? Cambridge, Cambridge University Press.

Schulman, L. H. and H. Pelka (1989). “The Anticodon Contains a Major Element of the Identity of Arginine Transfer RNAs.” Science **246**: 1595-1597.

Schultz, D. W. and M. Yarus (1994). “Transfer RNA Mutation and the Malleability of the Genetic Code.” J Mol Biol **235**: 1377-1380.

Schultz, D. W. and M. Yarus (1994). “tRNA structure and ribosomal function. I. tRNA nucleotide 27-43 mutations enhance first position wobble.” J Mol Biol **235**(5): 1381-94.

Schultz, D. W. and M. Yarus (1994). “tRNA structure and ribosomal function. II. Interaction between anticodon helix and other tRNA mutations.” J Mol Biol **235**(5): 1395-405.

Schultz, D. W. and M. Yarus (1996). “On Malleability in the Genetic Code.” J Mol Evol **42**: 597-601.

Schwartz, A. W. (1997). “Speculation on the RNA Precursor Problem.” J theor Biol **187**: 523-527.

Schwartz, A. W. and R. M. de Graaf (1993). “The Prebiotic Synthesis of Carbohydrates: A Reassessment.” J Mol Evol **36**: 101-106.

Senger, B., S. Auxilien, U. Englisch, F. Cramer, et al. (1997). “The modified wobble base inosine in yeast tRNAIle is a positive determinant for aminoacylation by isoleucyl-tRNA synthetase.” Biochemistry **36**(27): 8269-75.

Severin, K., D. L. Lee, A. J. Kennan and M. R. Ghadiri (1997). “A synthetic peptide ligase.” Nature **389**: 706-709.

Shannon, C. E. and W. Weaver (1949). The Mathematical Theory of Communication. Urbana, Ill., Univ. of Illinois Press.

Sharp, P. M., E. Cowe, D. G. Higgins, D. C. Shields, et al. (1988). “Codon usage patterns in Escherichia coli, Bacillus subtilis, Saccharomyces cerevisiae, Schizosaccharomyces pombe, Drosophila melanogaster and Homo sapiens; a review of the considerable within- species diversity.” Nucleic Acids Res **16**(17): 8207-11.

Sharp, P. M. and K. M. Devine (1989). “Codon usage and gene expression level in Dictyostelium discoideum: highly expressed genes do 'prefer' optimal codons.” Nucleic Acids Res **17**(13): 5029-39.

Sharp, P. M. and W. H. Li (1986). “An evolutionary perspective on synonymous codon usage in unicellular organisms.” J Mol Evol **24**(1-2): 28-38.

Sharp, P. M. and W. H. Li (1987). “The codon Adaptation Index--a measure of directional synonymous codon usage bias, and its potential applications.” Nucleic Acids Res **15**(3): 1281-95.

Sharp, P. M. and G. Matassi (1994). “Codon usage and genome evolution.” Curr Opin Genet Dev **4**(6): 851-60.

Sharp, P. M., T. M. Tuohy and K. R. Mosurski (1986). “Codon usage in yeast: cluster analysis clearly differentiates highly and lowly expressed genes.” Nucleic Acids Res **14**(13): 5125-43.

Shepherd, J. C. (1981). “Periodic correlations in DNA sequences and evidence suggesting their evolutionary origin in a comma-less genetic code.” J Mol Evol **17**(2): 94-102.

Shi, P., A. M. Weiner and N. Maizels (1998). “A top-half tDNA minihelix is a good substrate for the eubacterial CCA-adding enzyme.” RNA **4**: 276-284.

Shiba, K., H. Motegi, M. Yoshida and T. Noda (1998). “Human asparaginyl-tRNA synthetase: molecular cloning and the inference of the evolutionary history of Asx-tRNA synthetase family.” Nucleic Acids Res **26**(22): 5045-51.

Shimizu, M. (1982). “Molecular Basis for the Genetic Code.” J Mol Evol **18**: 297-303.

Shimizu, M. (1995). “Specific Aminoacylation of C4N Hairpin RNAs with the Cognate Aminoacyl-Adenylates in the Presence of a Dipeptide: Origin of the Genetic Code.” J Biochem **117**: 23-26.

Siatecka, M., M. Rozek, J. Barciszewski and M. Mirande (1998). “Modular evolution of the Glx-tRNA synthetase family--rooting of the evolutionary tree between the bacteria and archaea/eukarya branches.” Eur J Biochem **256**(1): 80-7.

Sitaramam, V. (1989). “Genetic code preferentially conserves long-range interactions among the amino acids.” FEBS Lett **247**(1): 46-50.

Sjöström, M. and S. Wold (1985). “A Multivariate Study of the Relationship Between the Genetic Code and the Physical-Chemical Properties of Amino Acids.” J Mol Evol **22**: 272-277.

Small, E. B. and D. H. Lynn (1981). “A new macrosystem for the phylum Ciliophora doflein, 1901.” Biosystems **14**(3-4): 387-401.

Small, I., H. Wintz, K. Akashi and H. Mireau (1998). “Two birds with one stone: genes that encode products targeted to two or more compartments.” Plant Mol Biol **38**(1-2): 265-77.

Sogin, M. L., H. J. Elwood and J. H. Gunderson (1986). “Evolutionary diversity of eukaryotic small-subunit rRNA genes.” Proc Natl Acad Sci U S A **83**(5): 1383-7.

Sogin, M. L., A. Ingold, M. Karlok, H. Nielsen, et al. (1986). “Phylogenetic evidence for the acquisition of ribosomal RNA introns subsequent to the divergence of some of the major *Tetrahymena* groups.” EMBO J **5**: 3625-3630.

Sokal, R. R. and F. J. Rohlf (1995). Biometry: The Principles and Practice of Statistics in Biological Research. New York, W. H. Freeman and Company.

Song, H., P. Mugnier, A. K. Das, H. M. Webb, et al. (2000). “The crystal structure of human eukaryotic release factor eRF1-- mechanism of stop codon recognition and peptidyl-tRNA hydrolysis.” Cell **100**(3): 311-21.

Sonneborn, T. M. (1965). Degeneracy of the Genetic Code: Extent, Nature, and Genetic Implications. Evolving Genes and Proteins. V. Bryson and H. J. Vogel. New York, Academic Press**:** 377-297.

Soto, M. A. and C. J. Toha (1985). “A hardware interpretation of the evolution of the genetic code.” BioSystems **18**: 209-215.

Sowerby, S. J., C. A. Cohn, W. M. Heckl and N. G. Holm (2001). “Differential adsorption of nucleic acid bases: Relevance to the origin of life.” Proc Natl Acad Sci U S A **98**(3): 820-2.

Sowerby, S. J. and W. M. Heckl (1998). “The role of self-assembled monolayers of the purine and pyrimidine bases in the emergence of life.” Orig Life Evol Biosph **28**(3): 283-310.

Sowerby, S. J., P. A. Stockwell, W. M. Heckl and G. B. Petersen (2000). “Self-programmable, self-assembling two-dimensional genetic matter.” Orig Life Evol Biosph **30**(1): 81-99.

Speyer, J. F., C. B. Lengyel, A. J. Wahba, R. S. Gardner, et al. (1963). “Synthetic polynucleotides and the amino acid code.” Cold Spring Harbor Symp Quant Biol **28**: 559-567.

Stechmann, A., M. Schlegel and D. H. Lynn (1998). “Phylogenetic relationships between Prostome and Colpodean ciliates tested by small subunit rRNA sequences.” Mol Phylogenet Evol **9**: 48-54.

Stenico, M., A. T. Lloyd and P. M. Sharp (1994). “Codon usage in Caenorhabditis elegans: delineation of translational selection and mutational biases.” Nucleic Acids Res **22**(13): 2437-46.

Sueoka, N. (1961). “Compositional Correlation between Deoxyribonucleic Acid and Protein.” Cold Spring Harb Symp Quant Biol **26**: 35-43.

Sueoka, N. (1962). “On the genetic basis of variation and heterogeneity of DNA base composition.” Proc Natl Acad Sci USA **48**(4): 582-592.

Sueoka, N. (1988). “Directional mutation pressure and neutral molecular evolution.” Proc Natl Acad Sci U S A **85**(8): 2653-7.

Sueoka, N. (1992). “Directional mutation pressure, selective constraints, and genetic equilibria.” J Mol Evol **34**(2): 95-114.

Sueoka, N. (1995). “Intrastrand parity rules of DNA base composition and usage biases of synonymous codons.” J Mol Evol **40**(3): 318-25.

Sueoka, N. (1999). “Two aspects of DNA base composition: G+C content and translation-coupled deviation from intra-strand rule of A = T and G = C.” J Mol Evol **49**(1): 49-62.

Sugita, T. and T. Nakase (1999). “Non-universal usage of the leucine CUG codon and the molecular phylogeny of the genus Candida.” Syst Appl Microbiol **22**(1): 79-86.

Suzuki, T., T. Ueda and K. Watanabe (1997). “The 'polysemous' codon--a codon with multiple amino acid assignment caused by dual specificity of tRNA identity.” EMBO J **16**(5): 1122-34.

Swanson, R. (1984). “A unifying concept for the amino acid code.” Bull Math Biol **46**: 187-203.

Szathmary, E. (1991). “Four letters in the genetic alphabet: a frozen evolutionary optimum?” Proc R Soc Lond B Biol Sci **245**(1313): 91-9.

Szathmary, E. (1992). “What is the optimum size for the genetic alphabet?” Proc Natl Acad Sci U S A **89**(7): 2614-8.

Szathmáry, E. (1991). “Codon Swapping as a Possible Evolutionary Mechanism.” J Mol Evol **32**: 178-182.

Szathmáry, E. (1993). “Coding coenzyme handles: A hypothesis for the origin of the genetic code.” Proc Natl Acad Sci USA **90**: 9916-9920.

Szathmáry, E. (1999). “The origin of the genetic code: amino acids as cofactors in an RNA world.” Trends Genet **15**(6): 223-9.

Szathmáry, E. and J. Maynard Smith (1995). “The major evolutionary transitions.” Nature **374**: 227-232.

Szathmáry, E. and J. Maynard Smith (1997). “From Replicators to Reproducers: the First Major Transitions Leading to Life.” J theor Biol **187**: 555-571.

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

Szostak, J. W. and A. D. Ellington (1993). In Vitro Selection of Functional RNA Sequences. The RNA World. R. F. Gesteland and J. F. Atkins. New York, Cold Spring Harbor Laboratory Press**:** 511-533.

Takahata, N. and M. Kimura (1981). “A model of evolutionary base substitutions and its application with special reference to rapid change of pseudogenes.” Genetics **98**(3): 641-57.

Takai, K., S. Okumura, K. Hosono, S. Yokoyama, et al. (1999). “A single uridine modification at the wobble position of an artificial tRNA enhances wobbling in an Escherichia coli cell-free translation system.” FEBS Lett **447**(1): 1-4.

Takai, K., H. Takaku and S. Yokoyama (1996). “Codon-reading specificity of an unmodified form of Escherichia coli tRNA1Ser in cell-free protein synthesis.” Nucleic Acids Res **24**(15): 2894-9.

Takai, K., H. Takaku and S. Yokoyama (1999). “In vitro codon-reading specificities of unmodified tRNA molecules with different anticodons on the sequence background of Escherichia coli tRNASer.” Biochem Biophys Res Commun **257**(3): 662-7.

Tao, J. and A. D. Frankel (1992). “Specific binding of arginine to TAR RNA.” Proc. Natl. Acad. Sci. **89**: 2723-2726.

Tao, J. and A. D. Frankel (1996). “Arginine-Binding RNAs Resembling TAR Identified by in Vitro Selection.” Biochemistry **35**: 2229-2238.

Tate, W. P., J. B. Mansell, S. A. Mannering, J. H. Irvine, et al. (1999). “UGA: a dual signal for 'stop' and for recoding in protein synthesis.” Biochemistry (Mosc) **64**(12): 1342-53.

Tate, W. P., E. S. Poole, M. E. Dalphin, L. L. Major, et al. (1996). “The translational stop signal: Codon with a context, or extended factor recognition element?” Biochimie **78**: 945-952.

Taylor, F. J. R. and D. Coates (1989). “The code within the codons.” Bio Systems **22**: 177-187.

Telford, M. J., E. A. Herniou, R. B. Russell and D. T. Littlewood (2000). “Changes in mitochondrial genetic codes as phylogenetic characters: two examples from the flatworms [In Process Citation].” Proc Natl Acad Sci U S A **97**(21): 11359-64.

Thompson, D. (1917). On Growth and Form. Cambridge, Cambridge University Press.

Tolstrup, N., J. Toftgard, J. Engelbrecht and S. Brunak (1994). “Neural network model of the genetic code is strongly correlated to the GES scale of amino acid transfer free energies.” J Mol Biol **243**(5): 816-20.

Tomii, K. and M. Kanehisa (1996). “Analysis of amino acid indices and mutation matrices for sequence comparison and structure prediction of proteins.” Protein Eng **9**(1): 27-36.

Tomita, K., T. Ueda, S. Ishiwa, P. F. Crain, et al. (1999). “Codon reading patterns in Drosophila melanogaster mitochondria based on their tRNA sequences: a unique wobble rule in animal mitochondria.” Nucleic Acids Res **27**(21): 4291-7.

Tomita, K., T. Ueda and K. Watanabe (1997). “5-formylcytidine (f5C) found at the wobble position of the anticodon of squid mitochondrial tRNA(Met)CAU.” Nucleic Acids Symp Ser **37**: 197-8.

Tomita, K., T. Ueda and K. Watanabe (1998). “7-Methylguanosine at the anticodon wobble position of squid mitochondrial tRNA(Ser)GCU: molecular basis for assignment of AGA/AGG codons as serine in invertebrate mitochondria.” Biochim Biophys Acta **1399**(1): 78-82.

Tomita, K., T. Ueda and K. Watanabe (1999). “The presence of pseudouridine in the anticodon alters the genetic code: a possible mechanism for assignment of the AAA lysine codon as asparagine in echinoderm mitochondria.” Nucleic Acids Res **27**(7): 1683-9.

Topal, M. D. and J. R. Fresco (1976). “Base pairing and fidelity in codon-anticodon interaction.” Nature **263**(5575): 289-93.

Topal, M. D. and J. R. Fresco (1976). “Complementary base pairing and the origin of substitution mutations.” Nature **263**(5575): 285-9.

Tourancheau, A. B., N. Tsao, L. A. Klobutcher, R. E. Pearlman, et al. (1995). “Genetic code deviations in the ciliates: evidence for multiple and independent events.” EMBO **14**(13): 3262-3267.

Trifonov, E. and T. Bettecken (1997). “Sequence fossils, triplet expansion, and reconstruction of earliest codons.” GENE **205**(1-2): 1-6.

Trifonov, E. N. (2000). “Consensus temporal order of amino acids and evolution of the triplet code.” Gene **261**(1): 139-151.

Trinquier, G. and Y. H. Sanejouand (1998). “Which effective property of amino acids is best preserved by the genetic code?” Protein Eng **11**(3): 153-69.

Tuerk, C. and L. Gold (1990). “Systematic Evolution of Ligands by Exponential Enrichment: RNA Ligands to Bacteriophage T4 DNA Polymerase.” Science **249**: 505-510.

Tumbula, D. L., H. D. Becker, W. Z. Chang and D. Soll (2000). “Domain-specific recruitment of amide amino acids for protein synthesis.” Nature **407**(6800): 106-10.

Unrau, P. J. and D. P. Bartel (1998). “RNA-catalysed nucleotide synthesis.” Nature **395**(6699): 260-3.

Visser, C. M. and R. M. Kellog (1978). “Biotin. Its Place in Evolution.” J Mol Evol **11**: 171-187.

Voet, D. and J. G. Voet (1995). Biochemistry. New York, John Wiley & Sons.

Volkenstein, M. V. (1965). “Coding of polar and non-polar amino acids.” Nature **207**: 294-295.

Volkenstein, M. V. (1966). “The genetic coding of protein structure.” Biochim Biophys Acta **119**: 421-424.

von Dohren, H., R. Dieckmann and M. Pavela-Vrancic (1999). “The nonribosomal code.” Chem Biol **6**(10): R273-9.

Wächtershäuser, G. (1988). “Before Enzymes and Templates: Theory of Surface Metabolism.” Microbiological Reviews **52**(4): 452-484.

Wächtershäuser, G. (1990). “Evolution of the first metabolic cycles.” Proc Natl Acad Sci USA **87**: 200-204.

Wagner, G. P. and L. Altenberg (1996). “Complex adaptations and the evolution of evolvability.” Evolution **50**(3): 967-976.

Watanabe, K. and S. Osawa (1995). tRNA Sequences and Variations in the Genetic Code. tRNA: Structure, Biosynthesis and Function. D. Söll and U. RajBhandary. Washington DC, ASM Press.

Watanabe, Y., H. Tsurui, T. Ueda, R. Furushima, et al. (1994). “Primary and higher order structures of nematode (Ascaris suum) mitochondrial tRNAs lacking either the T or D stem.” J Biol Chem **269**(36): 22902-6.

Watanabe, Y., H. Tsurui, T. Ueda, R. Furusihima-Shimogawara, et al. (1997). “Primary sequence of mitochondrial tRNA(Arg) of a nematode Ascaris suum: occurrence of unmodified adenosine at the first position of the anticodon.” Biochim Biophys Acta **1350**(2): 119-22.

Weber, A. L. (1987). “The triose model: glyceraldehyde as a source of energy and monomers for prebiotic condensation reactions.” Orig Life Evol Biosph **17**(2): 107-19.

Weber, A. L. (1989). “Thermal synthesis and hydrolysis of polyglyceric acid.” Orig Life Evol Biosph **19**: 7-19.

Weber, A. L. and J. C. Lacey Jr (1978). “Genetic Code Correlations: Amino Acids and Their Anticodon Nucleotides.” J Mol Evol **11**: 199-210.

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

Welch, M., I. Majerfeld and M. Yarus (1997). “23S rRNA Similarity from Selection for Peptidyl Transferase Mimicry.” Biochemistry **36**: 6614-6623.

Wertz, D. H. and H. A. Scheraga (1978). “Influence of water on protein structure. An analysis of the preferences of amino acid residues for the inside or outside and for specific conformations in a protein molecule.” Macromolecules **11**(1): 9-15.

Wetzel, R. (1995). “Evolution of the Aminoacyl-tRNA Synthetases and the Origin of the Genetic Code.” J Mol Evol **40**: 545-550.

White, H. B., III (1976). “Coenzymes as Fossils of an Earlier Metabolic State.” J Mol Evol **7**: 101-104.

Wiegand, T. W., R. C. Janssen and B. E. Eaton (1997). “Selection of RNA amide synthases.” Chem Biol **4**: 675-683.

Wilquet, V. and M. Van de Casteele (1999). “The role of the codon first letter in the relationship between genomic GC content and protein amino acid composition.” Res Microbiol **150**(1): 21-32.

Wilson, R. J. and D. H. Williamson (1997). “Extrachromosomal DNA in the Apicomplexa.” Microbiol Mol Biol Rev **61**(1): 1-16.

Wittung, P., P. E. Nielsen, O. Buchardt, M. Egholm, et al. (1994). “DNA-like double helix formed by peptide nucleic acid.” Nature **368**: 561-563.

Woese, C. R. (1965). “On the evolution of the genetic code.” Proc Natl Acad Sci USA **54**: 1546-1552.

Woese, C. R. (1965). “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. (1969). “Models for the Evolution of Codon Assignments.” J Mol Biol **43**: 235-240.

Woese, C. R. (1973). “Evolution of the genetic code.” Naturwissenschaften **60**(10): 447-59.

Woese, C. R., D. H. Dugre, S. A. Dugre, M. Kondo, et al. (1966). “On the fundamental nature and evolution of the genetic code.” Cold Spring Harb. Symp. Quant. Biol. **31**: 723-736.

Woese, C. R., D. H. Dugre, W. C. Saxinger and S. A. Dugre (1966). “The molecular basis for the genetic code.” Proc Natl Acad Sci USA **55**: 966-974.

Woese, C. R., G. E. Fox, L. Zablen, T. Uchida, et al. (1975). “Conservation of primary structure in 16S ribosomal RNA.” Nature **254**(5495): 83-6.

Wolfenden, R., L. Andersson, P. M. Cullis and C. C. Southgate (1981). “Affinities of amino acid side chains for solvent water.” Biochemistry **20**(4): 849-55.

Wolman, Y., W. J. Haverland and S. L. Miller (1972). “Nonprotein Amino Acids from Spark Discharges and Their Comparison with the Murchison Meteorite Amino Acids.” Proc Natl Acad Sci USA **69**: 809-811.

Wolstenholme, D. R., J. L. Macfarlane, R. Okimoto, D. O. Clary, et al. (1987). “Bizarre tRNAs inferred from DNA sequences of mitochondrial genomes of nematode worms.” Proc Natl Acad Sci U S A **84**(5): 1324-8.

Wong, J. T. (1980). “Role of minimization of chemical distances between amino acids in the evolution of the genetic code.” Proc Natl Acad Sci USA **77**(2): 1083-1086.

Wong, J. T.-F. (1975). “A Co-Evolution Theory of the Genetic Code.” Proc Natl Acad Sci USA **72**(5): 1909-1912.

Wong, J. T.-F. (1976). “The evolution of a universal genetic code.” Proc Natl Acad Sci USA **73**(7): 2336-2340.

Wong, J. T.-F. (1981). “Coevolution of genetic code and amino acid biosynthesis.” TIBS **6**: 33-36.

Wong, J. T.-F. (1983). “Membership mutation of the genetic code: loss of fitness by tryptophan.” Proc Natl Acad Sci USA **80**: 6303-6306.

Wong, J. T.-F. and P. M. Bronskill (1979). “Inadequacy of Prebiotic Synthesis as Origin of Proteinaceous Amino Acids.” J Mol Evol **13**: 115-125.

Wright, A. G., B. A. Dehority and D. H. Lynn (1997). “Phylogeny of the rumen ciliates *Entodinium, Epidinium* and *Polyplastron* (Litostomatea: Entodiniomorphida) inferred from small subunit ribosomal RNA sequences.” J Eukaryot Microbiol **44**: 61-67.

Wright, E. V. (1939). Gadsby: a story of over 50,000 words without using the letter "E". Los Angeles, Wetzel Publishing Co.

Wu, Y., A. K. Ogawa, M. Berger, D. L. McMinn, et al. (2000). “Efforts toward expansion of the genetic alphabet: Optimization of interbase hydrophobic interactions.” J Am Chem Soc **122**: 7621-7632.

Xia, X. (1996). “Maximizing transcription efficiency causes codon usage bias.” Genetics **144**(3): 1309-20.

Xia, X. (1998). “How optimized is the translational machinery in Escherichia coli, Salmonella typhimurium and Saccharomyces cerevisiae?” Genetics **149**(1): 37-44.

Xia, X. and W. H. Li (1998). “What amino acid properties affect protein evolution?” J Mol Evol **47**(5): 557-64.

Xie, T., D. Ding, X. Tao and D. Dafu (1998). “The relationship between synonymous codon usage and protein structure [published erratum appears in FEBS Lett 1998 Oct 16;437(1-2):164].” FEBS Lett **434**(1-2): 93-6.

Yamao, F., A. Muto, Y. Kawauchi, M. Iwami, et al. (1985). “UGA is read as tryptophan in Mycoplasma capricolum.” Proc Natl Acad Sci U S A **82**(8): 2306-9.

Yang, Y., M. Kochoyan, P. Burgstaller, E. Westhof, et al. (1996). “Structural Basis of Ligand Discrimination by Two Related RNA Aptamers Resolved by NMR Spectroscopy.” Science **272**: 1343-1346.

Yarus, M. (1988). “A specific amino acid binding site composed of RNA.” Science **240**: 1751-1758.

Yarus, M. (1989). “Specificity of Arginine Binding by the *Tetrahymena* Intron.” Biochemistry **28**: 980-988.

Yarus, M. (1991). “An RNA-Amino Acid Complex and the Origin of the Genetic Code.” New Biologist **3**(2): 183-189.

Yarus, M. (1993). An RNA-Amino Acid Affinity. The RNA World. R. F. Gesteland and J. F. Atkins. New York, Cold Spring Harbor Laboratory Press**:** 205-217.

Yarus, M. (1998). “Amino Acids as RNA Ligands: a Direct-RNA-Template Theory for the Code's Origin.” J Mol Evol **47**: 109-117.

Yarus, M. (2000). “RNA-ligand chemistry: a testable source for the genetic code.” RNA **6**(4): 475-84.

Yarus, M. and E. L. Christian (1989). “Genetic Code Origins.” Nature **342**: 349-350.

Yarus, M. and I. Majerfield (1992). “Co-optimization of Ribozyme Substrate Stacking and L-Arginine Binding.” J. Mol. Biol. **225**: 945-949.

Yarus, M. and D. W. Schultz (1997). “Response: Further Comments on Codon Reassignment.” J Mol Evol **45**: 1-8.

Yasuhira, S. and L. Simpson (1997). “Phylogenetic affinity of mitochondria of Euglena gracilis and kinetoplastids using cytochrome oxidase I and hsp60.” J Mol Evol **44**(3): 341-7.

Ycas, M. (1969). The Biological Code. Amsterdam, North-Holland publishing Company.

Yokoyama, S. and S. Nishimura (1995). Modified Nucleosides and Codon Recognition. tRNA: Structure, Biosynthesis and Function. D. Söll and U. RajBhandary. Washington DC, ASM Press.

Zama, M. (1990). “Codon usage and secondary structure of mRNA.” Nucleic Acids Symp Ser **22**: 93-4.

Zama, M. (1997). “Translational pauses during the synthesis of proteins and mRNA structure.” Nucleic Acids Symp Ser **37**: 179-80.

Zhang, B. and T. R. Cech (1997). “Peptide bond formation by in vitro selected ribozymes.” Nature **390**: 96-100.

Zimmermann, G. R., T. P. Shields, R. D. Jenison, C. L. Wick, et al. (1998). “A semiconserved residue inhibits complex formation by stabilizing interactions in the free state of a theophylline-binding RNA.” Biochemistry **37**(25): 9186-92.

Zinnen, S. and M. Yarus (1995). “An RNA pocket for the planar aromatic side chains of phenylalanine and tryptophane.” Nucleic Acids Symposium Series **33**: 148-151.

Zinoni, F., A. Birkmann, W. Leinfelder and A. Böck (1987). “Cotranslational insertion of selenocysteine into a formate dehydrogenase from *Escherichia coli* directed by a UGA codon.” Proc Natl Acad Sci USA **84**: 3156-3160.

Zinoni, F., J. Heider and A. Bock (1990). “Features of the formate dehydrogenase mRNA necessary for decoding of the UGA codon as selenocysteine.” Proc Natl Acad Sci U S A **87**(12): 4660-4.

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