A new classification scheme of the genetic code

Thomas Wilhelm* and Svetlana Nikolajewa

Institute of Molecular Biotechnology
Beutenbergstr.11
07745 Jena, Germany

Tel. +49 3641 656208
Fax +49 3641 656191
Email: wilhelm@imb-jena.de

* corresponding author

Abstract

Since the early days of the discovery of the genetic code non-random patterns have been searched for in the code in the hope of providing information about its origin and early evolution. Here we present a new classification scheme of the genetic code that is based on a binary representation of the purines and pyrimidines. This scheme reveals known patterns more clearly than the common one, for instance the classification of strong, mixed, and weak codons as well as the ordering of codon families. Furthermore, new patterns have been found that have not been described before: nearly all quantitative amino acid properties, such as Woese's polarity or the specific volume, show a perfect correlation to Lagerkvist's codon-anticodon binding strength.

Our new scheme leads to new ideas about the evolution of the genetic code. It is hypothesized that it started with a binary doublet code and developed via a quaternary doublet code into the contemporary triplet code. Furthermore, arguments are presented against suggestions that a "simplet" code, where only the mid-base was informational, was at the origin of the genetic code.

Key Words: genetic code, origin of life, evolution, doublet code, pattern, amino acid properties

Introduction

Crick (1968) introduced the notion that the genetic code is simply the result of pure chance or a "frozen accident" and that it therefore does not need any further evolutionary explanation. Later, this view was questioned. Although certain knowledge of the origin and early stages of life is not likely to be obtained, there are some hints of possible evolutionary scenarios of the genetic code. One direction of research (the "top-down approach" (Szathmary 1999)) analyzes patterns in the contemporary code (Knight and Landweber 1998, Szathmary 1999) and tries to infer appropriate chemical and selective forces. The bottom-up approach, on the other hand, is rooted in biochemistry and aims at constructing plausible scenarios for the origin of coding (Topal and Fresco 1976, Maizels and Weiner 1987, Szathmary 1993).

It has been appreciated for a long time that the genetic code assigns similar amino acids to similar codons (Sonneborn 1965, Woese 1965, Zuckerkandl and Pauling 1965, Crick 1968). Two different rationales have been presented: first, mutation (Sonneborn 1965, Zuckerkandl and Pauling 1965) and translation (Woese 1967, Haig and Hurst 1991, Freeland and Hurst 1998) error minimization (or both (Ardell and Sella 2002)), and second, similar amino acids tend to directly interact with similar RNA sequences (Woese et al. 1966, Yarus 1998, 2000). Landweber and coworkers found further evidence to support both hypotheses. Extending previous work (Haig and Hurst 1991, Freeland and Hurst 1998), by quantifying amino acid similarity, these authors were able to show that "the canonical code is at or very close to a global optimum for error minimization" (Freeland et al. 2000). Based on the earlier work of Yarus (cf. Yarus 1998, 2000), by doing a statistical analysis of RNA aptamers (nucleic-acid molecules selected to bind specific ligands) they concluded that there is "the strongest support for an intrinsic affinity between any amino acid and its codons" (Knight and Landweber 1998). It has also been proposed that instead of the actual codons, some derivatives of them, such as the anticodons (Dunnill 1966, Jungck 1978) or codon-anticodon duplexes (Alberti 1997) were the original amino acid binding motifs. It could also be that the original amino acid recognition

took place at the tRNA acceptor stem (Hopfield 1978) or that the specificity of aminoacylation is determined by the interaction of the tRNA synthetase with its tRNA (Weiner and Maizels 1987). Szathmary (1999) proposed that amino acid-RNA allocation took place even before the appearance of tRNA. He also gave a possible evolutionary scenario for the development of an anticodon hairpin to a longer structure with an operational code at the acceptor stem.

Several patterns of the genetic code have been identified which can be illustrated within the classical scheme.

The common scheme of the genetic code

The common scheme of the genetic code (Alberts et al. 2002)
contains 4^3=64 codons, a three-dimensional matrix where each dimension represents one of the three positions in the triplet code (Fig.1). Viewed this way, some patterns emerge: The first codon position seems to be correlated with amino acid biosynthetic pathways (Wong 1975, Taylor and Coates 1989), and to their evolution as evaluated by synthetic "primordial soup" experiments (Eigen 1977, Schwemmler 1994).The second position is correlated with the hydropathic properties of the amino acids (Crick 1968, Wolfenden et al. 1979, Taylor and Coates 1989), and the degeneracy of the third position could be related to the molecular weight or size of the amino acids (Hasegawa and Miyata 1980, Taylor and Coates 1989).

Lagerkvist (1978, 1981) divided the common illustration scheme (Fig.1) into a left part (containing the first and second column, i.e. U and C in the second position of the codon, respectively) and a right part (the third and fourth column, A and G in the second position). He observed that codon families (the amino acid of a codon family is uniquely determined by the first two nucleotides of a codon, cf. shaded regions in Fig.1) have a much higher probability to appear in the left part. Furthermore, he found that "strong" codons (the first two nucleotides in the codon are G and/or C) always represent codon families, while "weak" codons (A and/or U as the first two nucleotides) never do so. "Mixed" codons in the right part of the scheme never represent codon families, whereas mixed codons in the left part always stand for a codon family. Lagerkvist (1978) speculated "that interactions between mixed codons and their anticodons are stronger in the left half of the codon square".

However, most amino acid properties show no clear pattern in the common scheme of the genetic code. Instead Jungck (1978) used 15 different quantitative measures of amino acid properties such as polarity or molecular volume to demonstrate that these properties are generally more closely correlated with anticodon than with codon dinucleoside monophosphate properties. This supports the hypothesis that the relationship between amino acids and their anticodon dinucleosides was the basis for the origin of the genetic code.

In this article we follow the "top-down approach" towards understanding the organization of the genetic code. We are thereby led to propose a new classification scheme for the code that helps us to identify new patterns which in turn suggest new speculations about its origin.

Results

A new classification scheme of the genetic code

Fig.2 shows our new scheme for presenting the genetic code. It is based on a binary classification of nucleic acid bases. The two components of all nucleic acids, purines and pyrimidines, are denoted by 1 and 0, respectively. The 8 rows in Fig.2 represent the 2^3=8 possible combinations of three binary digits. Since there are two purines (A,G) and two pyrimidines (U,C) for each row, there again exist 8 possibilities.

Our first observation is that 4 (and not 8) columns are sufficient to place all 20 amino acids, as well as the termination codons. Each row contains exactly 4 different amino acids (including the termination codon). In the standard code, exceptions are the second row with two leucines and in the fourth row the AU* start codon. Note that here are also the deviations from the standard code. Interestingly, the yeast mitochondrial code shows no exception: each row contains exactly four different entries in four different columns. In this spirit the yeast mitochondrial code is the most regular one. The notice that in our scheme four columns are sufficient reflects the well-known fact that if the third position is important (in exactly half of our table this is not the case), then it is only decisive if there is either a purine (1) or a pyrimidine (0) (Fitch and Upper 1987), i.e. the third position is analyzed in a binary manner (Taylor and Coates 1989). This has been explained by Crick's wobble hypothesis (Crick 1966) whereupon the first two nucleotides of the codon pair with their anticodon bases according to Watson-Crick rules, but the third base pairs according to the wobble rules where G can also pair with U, for instance. The third codon position is exclusively analyzed in a binary manner in the mitochondrial codes of yeast, vertebrates, invertebrates, coelenterates and flatworms, as well as in the codes of mold, protozoan and mycoplasma/spiroplasma; for the other codes there are a few exceptions (cf. Elzanowski and Ostell 2000). Note that these few exceptions always have a purine at the third position of the codon (e.g. AUA (Ile) and AUG (Met) in the standard code).

Our scheme yields some support for the "adaptive genetic code" hypothesis (Freeland 2002) which states that the code has evolved to minimize the deleterious effects of mutation and translation error (Haig and Hurst 1991, Freeland and Hurst 1998). The purine-pyrimidine binary coding scheme, given in Fig.2, gives a much higher regularity than a binary coding according to the base pairs (A,U – 1; G,C – 0). This corresponds to the known fact that transition mutations (e.g. purine A vs. purine G) occur more frequently than transversion mutations (e.g. purine A vs. pyrimidine U).

A second observation concerns the order of the columns. In the first column the first two positions are G and C. These always pair with their anticodon base via 3 hydrogen bonds, i.e. the first two bases together always guarantee 6 hydrogen bonds. For that reason Lagerkvist (1978) called them strong codons. In the second and third column, the first two bases guarantee exactly 5 bonds (mixed codons) and in the fourth column just 4 bonds (weak codons). This pattern corresponds very well to the importance of the third base in the triplet codon: if the first bases are G and/or C (first column), the third base is never important, and in the second and third column, the third base is important in exactly half of the cases (if there is a purine in the second position – lower half of the table). In the fourth column the third base is always necessary for the determination of the correct amino acid. In Fig.2, the order of codon families is illustrated by the shaded regions. It seems that for the first column, the first two bases alone guarantee sufficient stability in the codon-anticodon pairing to ensure the correct choice of the amino acid. In the case of mixed codons (second and third column) a codon family is guaranteed if there is a pyrimidine in the second position. Going beyond Lagerkvists counting of hydrogen bonds, others provided some quantitative information about nucleotide binding strengths (Ornstein and Fresco 1983).

A third observation refers to two perfect symmetries in our scheme. The first is the codon-anticodon symmetry: the thick horizontal line in Fig.2 marks the symmetry axis. For instance, codon CCC (Pro, first column, first row) has the anticodon GGG (Gly, first column, last row), or codon ACG (Thr, third column, fourth row) has the anticodon UGC (Cys, third column, fifth row). The second perfect symmetry is the point symmetry corresponding to Halitsky's family – nonfamily symmetry operation ("E-M bifurcation", Halitsky 2003), indicated by the point in the center of Fig.2. Halitsky observed that all the 32 "family codons" CC*, CU*, UC* GC*, GU*, AC*, CG*, GG* can be mapped into the 32 "nonfamily codons" UU*, AU*, CA*, UG*, UA*, GA*, AG*, AA* by exchanging the two amino bases A and C

with one another, and the two keto bases U and G with one another. For instance, the family codon GUA (Val) is mapped into the nonfamily codon UGC (Cys). Thus, this point symmetry is behind the family – nonfamily symmetry in our scheme (shaded vs. unshaded regions).

A fourth observation concerns the deviations of non-standard genetic codes. As can be seen in Fig.2, nearly all deviations occur in codons with a purine at the third position. The only exception is the yeast mitochondrial code where CU* does not code for Leu, but rather for Thr.

Our fifth observation refers to hitherto unknown regularities of amino acid properties. Jungck (1978) collected 15 different measures of amino acid properties, as well as three measures for dinucleoside monophosphates. For all of these 18 measures we arranged a table with 8 rows and 4 columns corresponding to the scheme in Fig.2 (for AU(G/A) we took the Met values (e.g. vertebrate mitochondrial code), for UA(G/A) the Tyr values (mitochondrial flatworm code)). Then we analyzed all row and column sums. The row sums show a strong monotonicity just for the three dinucleoside monophosphate measures and for the hydropohobicity measure of Levitt. However, amazingly, the column sums of nearly all measures are perfectly correlated to the corresponding codon-anticodon binding strength in the sense of Lagerkvist (1978, 1981), in the following simply denoted as codon strength. This is demonstrated in Table 1. For this table we averaged the column sums of the second and third column, giving one “mixed codons” column. As can be seen in Table 1 there are just two exceptions. In the polarity measure of Zimmerman, the deviation is only very weak and in contradiction to all other measures, here the values for the amino acids vary by orders of magnitude. A problem only arises for the three hydrophobicity measures: The two monotonic measures “Levitt” and “BullBreese” are anticorrelated, and the “Jones” measure is not monotonic. The anticorrelation was already found by Jungck (1978), but he did not comment on this.

The fact that the order of the second and third column is not fixed is also underlined by an individual consideration of the two mixed codon columns, instead of the averaging done in Table 1. In about half of the cases the order of the second and third column should be exchanged to guarantee the strong monotonicity of the amino acid measures as function of the column number.

The observed pattern of strong correlation between amino acid properties and codon strength (considers just the first two nucleotides) implies that both first positions together, and not the first or second position alone must have been important for the amino acid – codon assignment in the evolution of the genetic code.

Evolution of the genetic code

What do the observed patterns tell us about the evolution of the genetic code? The so-called biosynthetic theory assumes that the genetic code evolved from a simpler form that encoded fewer amino acids (Crick 1968). A special version of this theory has been given by Wong (1975) who proposes that the genetic code coevolved with the invention of biosynthetic pathways for new amino acids. Although it has been shown that his analyzes rest on wrong assumptions (Ronneberg et al. 2000), it is generally accepted that one can discriminate evolutionary old and new amino acids (Alberts et al. 2002). Of course it could be that the binding allocation between nucleic acid molecules (RNAs or even PNAs (Knight and Landweber 2000b)) and amino acids did not start until all 20 amino acids were available; but it seems simpler to assume that as soon as there were amino acids and nucleic acids available (produced abioticly), both began to bind to each other. It now seems clear that “the code probably underwent a process of expansion from relatively few amino acids to the modern complement of 20” (Knight and Landweber 2000b).

Does our scheme yield some hints as to the evolution of the code? We already noted that the third nucleotide is nearly always (two exceptions in the standard code) analyzed just in

a binary manner. Taking this for granted, we can reduce our originally 8x8 scheme to a 8x4 scheme (shown in Fig.2). Looking at this scheme, we observe a high redundancy for each second row. Therefore, it is tempting to speculate that there was a period during code evolution where the third position was not needed at all. Assuming this, we can cancel each second row and are left with a pure doublet code that encodes 4x4=16 amino acids (or 15 plus a termination codon). Perhaps then, a doublet code preceded the triplet code, as already had been speculated (Jukes 1973, Hayes 1998).

Conceivably, codon expansion from doublet to triplet could have arisen before this, or possibly not until all 16 amino acids were encoded. If one assumes the latter, then it is interesting to postulate for each doublet the corresponding old amino acid. Met (Wong 1975), Trp, Gln, Asn (Knight and Landweber 2000b), and Tyr (Alberts et al. 2002) seem to be newer amino acids. As mentioned above, Szathmary (1999) proposed an evolutionary mechanism of tRNA formation. In principle, this mechanism could also work starting with doublets instead of triplets. It should be possible to gain experimental evidence for a doublet code by studying amino acid – nucleic acid doublet binding in the same way as has been done for triplets. Knight and Landweber (2000a) showed that Arg triplet codons alone significantly associate with arginine binding sites. Perhaps the doublets show a higher specificity.

However, by proposing a doublet code one faces the frameshifting problem. It seems to be unthinkable that a sudden transition from a two-letter to a three-letter frame ever occurred. Instead, one can imagine a gradual evolution with an ancient three-letter reading frame where just the first two letters have been analyzed by an ancient translation machinery. However, one then wonders about such inefficient use of coding space. Perhaps the ancient translation machinery could simply for stereochemical reasons not analyze a two-letter frame. In this context it is also interesting to note that even our contemporary code is somehow ‘inefficient’: already a quaternary doublet code can encode 16 amino acids (or 15 plus a termination codon). For just four (or fife) further amino acids a third letter is necessary. Of course, this inefficiency has the advantage of robustness enhancing redundancy.

Szathmary (1992, 2003) proposed a model which yields the result that two different base pairs represent an optimal compromise between the overall copying fidelity and an overall reproduction rate (metabolic efficiency). He assumed that the genetic code was developed before evolution invented proofreading. For higher copying fidelity (due to proofreading, etc.), the model predicts that three different base pairs are better than just two. It is tempting to speculate that in the earliest phases of biological evolution with the lowest copying fidelity just one base pair could have worked as well (The copying fidelity is always highest for just one base pair. Nevertheless, Szathmary’s simple model gives no one base pair optimum, but a more detailed model for the metabolic efficiency could do so.). So, perhaps, nucleic acid – amino acid mapping started with a binary code. This is in accordance with earlier speculations that the first genetic material contained only a single base-pairing unit (Crick 1968, Orgel 1968). An important argument in this context is the chemical instability of cytosine, so that it may be difficult to establish a genetic system with G-C base pairing (Levy and Miller 1998). Wächtershäuser (1988) proposed an all-purine precursor of nucleic acids. However, for the sake of self-replication it is more obvious to assume a two-letter code that can give rise to complementary base pairing. Jimenez-Sanchez (1995) argued for an early (binary) A-U coding. Recently, a ribozyme composed of only two different nucleotides has been found by in vitro evolution that contained the pyrimidine uracil and the purine 2,6-diaminopurine (Reader and Joyce 2002). Note that uracil is the biosynthetic precursor of the pyrimidines cytosine and thymine (the corresponding precursor of the purines adenine and guanine is hypoxanthine).

Of course, a binary encoding also would be the most aesthetic version from a purely mathematical point of view. A binary triplet code would represent just one column in our scheme (Fig.2). Given the high redundancy between the rows, it is unlikely that this ever happened. However, an even simpler coding, a binary doublet code, seems conceivable. It is

tempting to speculate which four amino acids, one per two consecutive rows, were the first encoded ones. In the first two rows (two pyrimidines, i.e. 00) Ser seems to be the oldest amino acid, and in the third and fourth row (10) Ala (Wong 1975).
On the other hand the 01 rows obviously contain no really old amino acid while the 11 rows contain more than one: Gly, Asp, Glu (Wong 1975).

One could speculate that the termination marker was important from the very beginning and resulted in coding by the 01 binary doublet. It has been noted that the five amino acids coded by G** (Ala, Val, Gly, Asp, Glu) are all at or near the head of the amino acid synthesis pathways (Taylor and Coates 1989) and also the most abundantly formed ones in abiotic synthesis experiments (Miller 1953, 1987). Furthermore, it has been shown recently by extensive statistical analyzes that the frequencies of all fife G** amino acids are significantly greater in evolutionary conserved residues and it has been concluded that "these amino acids may have been the first introduced into the genetic code" (Brooks and Fresco 2002, 2003, Brooks et al. 2002). This is also consistent with physicochemical arguments proposing that the first sense codons had the form G** (Eigen and Schuster 1978). However, Gly is biochemically built from Ser, so Ser can be assumed as prior. It could be that in the beginning of nucleic acid – amino acid assignment Asp and Glu competed for the 11 doublet. Of course, code transfer from one amino acid to another one might also have occurred (Wong 1975).

Another scenario consistent with a binary doublet code has been given by Fitch's "ambiguity reduction" hypothesis (Fitch and Upper 1987). It states that early in evolution there was an ambiguity in the charging of amino acids to anticodon acceptors: in a first step just *pyrimidine* codons (*0*), coding for hydrophobic amino acids, and *purine* codons (*1*), coding for hydrophilic amino acids, has been distinguished (binary singulet code). In a second step the more refined binary doublet code (00*, 01*, 10*, 11*) evolved.

The idea that the doublet code was just the second state in the evolution of the genetic code and that this evolution started with just the mid-base as coding, has been worked out by others, who termed it "simplet" code (McClendon 1986, Schwemmler 1994). However, in this hypothesis both old amino acids Ser (UC*) and Ala (GC*), as well as Asp (GA*) and Glu (GA*), cannot be discriminated. We therefore suggest that the first two positions were equally important from the very beginning. Although our suggestion also does not allow discrimination between the related amino acids Asp and Glu, it nevertheless allows discrimination between the functionally divergent amino acids Ser and Ala. A further argument for the evolutionary importance of the first two nucleotides is the strong correlation observed between codon strength and the amino acid properties.

Conclusion

Taylor and Coates (1989) stated that "Many parts of the patterns (of the genetic code) have been seen by others but … it is the synthesis that adds up to the most interesting …new insights." In this spirit, we note that in the work presented here different patterns appear more clearly than in the common scheme of the genetic code. An example is Lagerkvists (1978) observation that all strong codons represent codon families while weak codons do not. Mixed codons represent codon families in half of the cases. Our presentation of the code also highlights new patterns, which were not seen before. As summarized in Table 1, nearly all measures of the amino acid properties strongly correlate with the codon strengths. Furthermore, there is a perfect codon – anticodon symmetry as well as point-symmetry corresponding to the family – nonfamily symmetry operation (Halitsky 2003) in the here presented scheme.

With regard to evolution, we hypothesize that codon assignments started from a binary doublet code (e.g., hypoxanthin and uracil) and developed later to a quaternary doublet code (A, G, C, U); thereafter, expansion to a triplet code took place. Although the third position is needed for correct amino acid recognition, still until now it is nearly always analyzed in a

binary manner. The conclusion that code evolution must have started with doublets and not with a single letter is also underlined by the correlation observed here between properties of amino acids and the codon strengths.

Acknowledgments. We thank two anonymous reviewers for many valuable comments and refering us to relevant literature, and A. Beyer, F. Grosse and M.-L. Merten for critical reading of the manuscript. This work was supported by Grant 0312704E of the Bundesministerium für Bildung und Forschung.

References

Alberti S (1997) The origin of the genetic code and protein synthesis. J. Mol. Evol. 45:352-358

Alberts B, Johnson A, Lewis J, Raff M, Roberts K, Walter P (2002) Molecular Biology of the Cell. Garland Science, NY

Ardell DH, Sella G (2002) No accident: genetic codes freeze in error-correcting patterns of the standard genetic code. Phil. Trans. R. Soc. Lond. B 357:1625-1642

Brooks DJ, Fresco JR (2002) Increased frequency of cysteine, tyrosine, and phenylalanine residues since the last universal ancestor. Mol. Cell. Prot. 1.2:125-131

Brooks DJ, Fresco JR, Lesk AM, Singh M (2002) Evolution of amino acid frequencies in proteins over deep time: inferred order of introduction of amino acids into the genetic code. Mol. Biol. Evol. 19(10):1645-1655

Brooks DJ, Fresco JR (2003) Greater GNN pattern bias in sequence elements encoding conserved residues of ancient proteins may be an indicator of amino acid composition of early proteins. Gene 303:177-185

Crick FHC (1966) Codon-anticodon pairing: the wobble hypothesis. J. Mol. Biol. 19:548-555

Crick FHC (1968) The origin of the genetic code. J. Mol. Biol. 38:367-379

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

Eigen M (1977) The hypercycle. A principle of natural self-organization. Part A: Emergence of the hypercycle. Naturwissenschaften 64:541-565

Eigen M, Schuster P (1978) The hypercycle: a principle of natural self-organization. Naturwissenschaften 65:341-368

Elzanowski A, Ostell J (2000) Genetic codes. http://www3.ncbi.nlm.nih.gov/htbin-post/Taxonomy/wprintgc?mode=t#SG1

Fitch WM, Upper K (1987) The phylogeny of tRNA sequences provides evidence for ambiguity reduction in the origin of the genetic code. Cold Spring Harbor Symposia on Quantitative Biology 52:759-767

Freeland SJ (2002) The Darwinian genetic code: An adaptation for adapting? Genetic Progamming and Evolvable Machines 3:113-127

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

Freeland SJ, Knight RD, Landweber LF, Hurst LD (2000) Early fixation of an optimal genetic code. Mol. Biol. Evol. 17:511-518

Haig D, Hurst LD (1991) A quantitative measure of error minimization in the genetic code. J. Mol. Evol. 33:412-417

Halitsky D (2003) Extending the (hexa-)rhombic dodecahedral model of the genetic code: the code's 6-fold degeneracies and the orthogonal projections of the 5-cube as 3-cube. Contributed paper (983-92-151), American Mathematical Society; and personal communication

Hasegawa M, Miyata T (1980) On the antisymmetry of the amino acid code table. Orig. Life 10:265-270

Hayes B (1998) The invention of the genetic code. Amer. Scientist 86:8-14

Hopfield JJ (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

Jimenez-Sanchez, A (1995) On the origin and evolution of the genetic code. J. Mol. Evol. 41:712-716

Jukes, TH (1973) Possibilities for the evolution of the genetic code from a preceding form. Nature 246:22-26

Jungck JR (1978) The genetic code as a periodic table. J. Mol. Evol. 11:211-224

Knight RD, Landweber LF (1998) Rhyme or reason: RNA-arginine interactions and the genetic code. Chem. & Biol. 5:R215-R220

Knight RD, Landweber LF (2000a) Guilt by association: the arginine case revisited. RNA 6:499-510

Knight RD, Landweber LF (2000b) The early evolution of the genetic code. Cell 101:569-572

Knight RD, Freeland SJ, Landweber LF (2001) Rewiring the keyboard: evolvability of the genetic code. Nat. Rev. Gen. 2:49-58

Lagerkvist U (1978) "Two out of three": An alternative method for codon reading. Proc. Natl. Acad. Sci. USA 75:1759-1762

Lagerkvist U (1981) Unorthodox codon reading and the evolution of the genetic code. Cell 23:305-306

Levy M, Miller SL (1998) The stability of the RNA bases: implications for the origin of life. Proc. Natl. Acad. Sci. USA 95:7933-7938

Maizels N, Weiner AM (1987) Peptide-specific ribosomes, genomic tags, and the origin of the genetic code. Cold Spring Harbor Symposia on Quantitative Biology 52:743-749

McClendon JH (1986) The relationship between the origins of the biosynthetic paths to the amino acids and their coding. Origins Life 16:269-270

Miller SL (1953) Production of amino acids under possible primitive earth conditions. Science 117:528-529

Miller SL (1987) Which organic compounds could have occurred on the prebiotic earth? Cold Spring Harbor Symposia on Quantitative Biology 52:17-27

Orgel LE (1968) Evolution of the genetic apparatus. J. Mol. Biol. 38:381-393

Ornstein RL, Fresco JR (1983) Correlation of Tm, sequence, and ΔH of complementary RNA helices and comparison with DNA helices. Biopolymers 22:2001-2016

Osawa S, Jukes TH, Watanabe K, Muto A (1992) Recent evidence for the evolution of the genetic code. Microbiol. Rev. 56(1):229-264

Reader JS, Joyce GF (2002) A ribozyme composed of only two different nucleotides. Nature 420:841-844

Ronneberg TA, Landweber LF, Freeland SJ (2000) Testing a biosynthetic theory of the genetic code: fact or artifact? Proc. Natl. Acad. Sci. USA 97:13690-13695

Schwemmler W (1994) Reconstruction of cell evolution: A periodic system of cells. CRC Press, Boca Raton, FL

Sonneborn TM (1965) Degeneracy of the genetic code: extent, nature, and genetic implications. In: Bryson V, Vogel HJ (eds) Evolving Genes and Proteins. Academic Press, NY, pp. 297-377

Szathmary E (1992) What is the optimum size for the genetic alphabet? Proc. Natl. Acad. Sci. USA 89:2614-2618

Szathmary E (1993) Coding coenzyme handles: A hypothesis for the origin of the genetic code. Proc. Natl. Acad. Sci. USA 90:9916-9920

Szathmary E (1999) The origin of the genetic code. TIG 15:223-229

Szathmary E (2003) Why are there four letters in the genetic alphabet? Nat. Rev. Gen. 4:995-1001

Thanbichler M, Böck A (2002) The function of SECIS RNA in translational control of gene expression in *Escherichia coli*. EMBO J. 21:6925-6934

Taylor FJR, Coates D (1989) The code within the codons. BioSystems 22:177-187

Topal MD, Fresco JR (1976) Base pairing and fidelity in codon-anticodon interaction. Nature 263:289-293

Wächtershäuser G (1988) An all-purine precursor of nucleic acids. Proc. Natl. Acad. Sci. USA 85:1134-1135

Weiner AM, Maizels N (1987) tRNA-like structures tag the 3' ends of genomic RNA molecules for replication: Implications for the origin of protein synthesis. Proc. Natl. Acad. Sci. USA 84:7383-7390

Woese CR (1965) On the evolution of the genetic code. Proc. Natl. Acad. Sci. USA 54:1546-1552

Woese CR (1967) The genetic code: The molecular basis for Genetic Expression. Harper & Row, NY

Woese CR, Dugre DH, Saxinger WC, Dugre SA (1966) The molecular basis for the genetic code. Proc. Natl. Acad. Sci. USA 55:966-974

Wolfenden RV, Cullis PM, Southgate CCF (1979) Water, protein folding, and the genetic code. Science 206:575-577

Wong JT-F (1975) A co-evolution theory of the genetic code. Proc. Natl. Acad. Sci. USA 72:1909-1912

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

Zuckerkandl E, Pauling L (1965) Evolutionary divergence and convergence in proteins. In:

Bryson V, Vogel HJ (eds) Evolving Genes and Proteins. Academic Press, NY, pp. 97-167

Measure	Strong codons	Mixed codons	Weak codons
	Dinucleoside monophosphates		
Hydrophilicity (Weber & Lacey 1978)	1.686	1.434	1.235
Hydrophilicity (Barzilay et al. 1973)	2.72	2.26	2.26
Hydrophobicity (Garel et al. 1973)	2.556	3.413	3.982
	Amino acids		
Molec. Weight (Handbook value)	907	1065.6	1217.5
Molec. Volume (Grantham 1974)	381	637.5	906
Refractivity (Jones 1975)	83.86	140.03	186.51
Alpha pK1 (Zimmermann et al. 1968)	16.96	17.11	17.43
Bulkiness (Zimmermann et al. 1968)	93.22	124.345	143.54
Specific volume (McMeekin et al. 1964)	5.26	5.37	5.8
Polarity (Zimmerman et al. 1968)	107.16	109.58	58.14
Polarity (Woese et al. 1967)	61.2	59.15	51
Polarity (Grantham 1974)	71.2	67	56.3
Hydrophobicity (Jones 1975)	9.18	8.385	16.93
Hydrophobicity (Levitt 1976)	-2.2	1.6	8.8
Hydrophobicity (Bull & Breese 1974)	3880	-165	-6790
Hydrophilicity (Weber & Lacey 1978)	7.02	6.585	5.59
Partition coefficient (Garel et al. 1973)	1.88	5.58	7.6
Sequence Frequency (Jungck 1971)	4280	3522	2966

Table 1 Correlation of codon strength and amino acid properties.
Averaged values (per column, in our scheme of Fig. 2) of quantified dinucleoside monophosphate properties (codon and anticodon values give the same average, because of the codon-anticodon symmetry) and amino acid properties for strong, mixed and weak codons. Each row represents one of the measures published by Jungck (1978; This paper contains (in its Table 1) all detailed references as well as a short note to the determination procedure.).

Figure Legends

Figure 1
The common presentation of the standard (`universal´) genetic code. All deviations from this code (Elzanowski and Ostell 2000) are thought to be the result of later mutations (Osawa et al. 1992, Knight and Landweber 2000b, Knight et al. 2001). Shaded regions show codon families.

Figure 2
A new classification scheme of the standard genetic code based on a binary representation of purines (1) and pyrimidines (0). The third base is given in parenthesis. When there are differences between the standard code and any other code, the number of deviations from the standard code is indicated. This comparison is based on 16 non-standard codes (Elzanowski and Ostell 2000). For instance, in the UG(G/A) field, 0/9 indicates that UGG encodes for Trp in all codes, but UGA is not the termination codon in 9 of the 16 non-standard codes: in 8 different mitochondrial codes UGA encodes Trp, and in the euplotid nuclear code it represents Cys. It is interesting that at least in some bacteria the 21st amino acid, selenocysteine, can also be encoded by UGA (Osawa et al. 1992, Thanbichler and Böck 2002). Another example is the CU(G/A) field. In the yeast mitochondrion CUG and CUA encode Thr, in the alternative yeast nuclear code CUG represents Ser.
Shaded regions show codon families. The point in the center indicates the perfect point symmetry in this scheme, according to Halitsky's family – nonfamily symmetry operation (Halitsky 2003). The thick horizontal line marks the symmetry axis for codon-anticodon symmetry.

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

Figure 1

Code	Strong codons **6** hydrogen bonds	Mixed codons **5** hydrogen bonds (first G or C)	Mixed codons **5** hydrogen bonds (first U or A)	Weak codons **4** hydrogen bonds
000	**Pro CC** ***(C/U)***	**Leu CU** ***(C/U)*** 1/1	**Ser UC** ***(C/U)***	**Phe UU** ***(C/U)***
001	**Pro CC** ***(G/A)***	**Leu CU** ***(G/A)*** 2/1	**Ser UC** ***(G/A)*** 0/1	**Leu UU** ***(G/A)*** 0/1
100	**Ala GC** ***(C/U)***	**Val GU** ***(C/U)***	**Thr AC** ***(C/U)***	**Ile AU** ***(C/U)***
101	**Ala GC** ***(G/A)***	**Val GU** ***(G/A)***	**Thr AC** ***(G/A)***	**Met/Ile AU** ***(G/A)*** 0/5
010	**Arg CG** ***(C/U)***	**His CA** ***(C/U)***	**Cys UG** ***(C/U)***	**Tyr UA** ***(C/U)***
011	**Arg CG** *(G/A)*	**Gln CA** ***(G/A)***	**Trp/Stop UG** ***(G/A)*** 0/9	**Stop UA** ***(G/A)*** 4/2
110	**Gly GG** ***(C/U)***	**Asp GA** ***(C/U)***	**Ser AG** ***(C/U)***	**Asn AA** ***(C/U)***
111	**Gly GG** ***(G/A)***	**Glu GA** ***(G/A)***	**Arg AG** ***(G/A)*** 6/6	**Lys AA** ***(G/A)*** 0/3

Figure 2